Schriftenreihe IWAR

264

TECHNISCHE
UNIVERSITÄT
DARMSTADT

IWAR

*Luz Daniela Alejo Alvarez*

# On a deeper understanding of data-driven approaches in the current framework of wastewater treatment: looking inside the black-box

Herausgeber:
Verein zur Förderung des Instituts IWAR der TU Darmstadt e. V.

# On a deeper understanding of data-driven approaches in the current framework of wastewater treatment: looking inside the black-box

Vom Fachbereich Bau- und Umweltingenieurwissenschaften
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades eines
Doktors (Dr.-Ing.) genehmigte Dissertation

von

Luz Daniela Alejo Alvarez, P.Eng.
aus Concepción, Chile

Darmstadt 2020
D 17

Luz Daniela Alejo Alvarez

On a deeper understanding of data-driven approaches in the current framework of wastewater treatment: looking inside the black-box

Darmstadt, Technische Universität Darmstadt
Jahr der Veröffentlichung der Dissertation auf TUprints: 2021

URN: nbn:de:tuda-tuprints-134630
URI: https://tuprints.ulb.tu-darmstadt.de/id/eprint/13463

Hrsg.: Verein zur Förderung des Instituts IWAR der TU Darmstadt e. V.
Darmstadt: Eigenverlag, 2020
(Schriftenreihe IWAR 264)

ISSN 0721-5282
ISBN 978-3-940897-65-7

Referentin: Prof. Dr. Susanne Lackner
Korreferent: Prof. Dr. John Atkinson-Abutridy

Tag der schriftlichen Einreichung: 27.05.2020
Tag der mündlichen Prüfung: 30.07.2020

Herstellung: Lasertype GmbH
Holzhofallee 19
64295 Darmstadt

Vertrieb: Institut IWAR
TU Darmstadt
Franziska-Braun-Straße 7
64287 Darmstadt
Telefon: 06151 / 16 20301
Telefax: 06151 / 16 20305

*“When Nature finishes producing its own species, man begins with the help of Nature, to create an infinity of species”*
*Leonardo da Vinci*

*To my family…*

## Acknowledgements

I'm inspired by books, the color, the fonts, the weight and of course, the knowledge and content inside. Whenever I select one, I assume that the person/people who wrote it are passionate and dedicated scientists, writers or dreamers. Accordingly, I respect and enjoy reading them.

I have a double life, one is filled with books, late nights of work and dreaming with numbers and ideas, it is a rewarding world. My second one is by far superior, it is filled with inspiring people, whom without their help, I would have not gotten here…

Endlessly will I be grateful with my supervisors Susanne and John, for fruitful discussions, boosting advice, and most important, their friendship. They allowed me to learn from them and supported me these past three years. They are not just accomplished researchers and scientists but incredible human beings. Susanne, thanks for trusting in my work and taking a risk with what I do, for the freedom to get my crazy ideas in motion.

I want to thank my family, for their unconditional support and encouragement. Thank you mom and dad for giving me the opportunity to learn music, explore different sports, learn languages, read my homework while I was a kid and for encouraging me on doing everything with passion, to keep my values intact and stay humble. My parents never let me think that a goal was too big for me, encouraging me to be whomever I wanted to be, flagging their efforts as they guided me. I would like to thank my sister Lucía (Clu), for taking care of me, making sure that I would feed my self properly, and bringing herself to Germany with all her culinary magic that transports me to the flavors of my childhood. I would like to thank Carlos, my fiancée, for his love, containment, support, bearing with me these past stressful months and being there through all the good, bad and worst, you have been an outstanding partner.

I consider myself very lucky, because I had the fortune to meet kind and caring people during my studies in Chile and Germany. Jose Luis, Marianne and Fernanda, thank you for your friendship during these past 10 years. A special thanks to you Jose Luis, for sharing your loving family with me (almost since I arrived from Bolivia to Chile), I thank them too for their love, amazing food and company in Guarilihue. Gracias!

To my gang here in Germany, Hajo, Thomas, Philipp, Luisa, Vava, Michael and Nastia. You gave full colors to my PhD. Thomas, was the best office-mate a workaholic can get, his love for science is inspiring! Vane, *hermana*, you are my Amos, and working with you is pure joy. Hajo, thank you for your music enlighten and remembering me often that there is life beyond work. To all the members of the Lackner Lab; Vera, Shelesh, Laura, Jochen for their companionship, I feel very proud to work along such great researchers.

Bogdan and Cristina, for teaching me the value of true friendship. Thank you for the countless good moments, trips and Pălincă shots, for your friendship and love.

Last but not least, I want to thank the 12 year old girl that was eager to become a scientist. A lot of hard work and passion allowed me to start the road in the service of Science. Pipelines fueled with dreams were engine and are engine for me, to explore and learn every day.

These final thank notes go to my grandmother, my utmost inspiration. She is until today the strongest woman I know.

## Abstract

Machine learning (ML) is one of the most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence (AI) and data science. The effect of ML is broadly felt across a range of industries concerned with data intensive issues, such as consumer services, banking, astronomy and empirical sciences, among others. In the field of wastewater treatment, the origin of vast data generation came along with automation of wastewater treatment plants (WWTP). Additionally, an increase of the computing and storage capacity, allowed large amounts of information to be generated in the water sector coming from different sources to be stored. The information from WWTP, that is generated and recorded involves complex and heterogeneous data sources; on-line from sensors, on/off control data from pumps and equipment and off-line measurements from laboratories. Sensors are able to record measurements every few seconds, thus, generating thousands of data points daily. The data generated in laboratories in wastewater treatment is crucial to evaluate the quality of the water in any biological wastewater treatment process (bWWTP) and often to validate the sensors information. However, due to the costs and time involved, the frequency of sampling for laboratory measurements is often dramatically reduced compared to sensors. Thus, the resulting database (from sensors and laboratories), involve varying frequencies of sampling and thus a highly heterogeneous dataset.

Current research on data-driven methods in wastewater treatment has focused mainly on predictive tasks, to forecast the effluent composition and performance of different bWWTP, the latter also widely studied by activated sludge models (ASM). Although the outcome could be similar with both approaches, the application and the input information to the models is very different. Data-driven approaches require enough data to perform an analysis task, they are data driven. However, the nature of ASM models is phenomenological, which aims to describe the biochemical interaction between the microbial community in the wastewater system and main pollutants in the wastewater; organic matter, nitrogen, phosphorus and other dissolved nutrients. Both approaches provide useful and important information from the process performance, however it is utmost important to distinguish and clarify the differences and goals of ASM-type models and ML-based tasks in the current framework of wastewater treatment. The main reasons that moved the wastewater treatment community to apply these methods in predictive tasks are two-fold; i) is the availability of data gathered from monitoring different bWWTP and ii) the already mentioned complexity of biological processes. The high adaptability of ML methods to dynamic systems has conducted the research community to a wide application of these methods. However, a key issue emerges from the literature. The current studies related to data-driven methods in wastewater treatment do not explicitly describe the pre-processing techniques applied, the amount of the data used for analysis, the frequency considered for the data selection and the rationale behind the selection of the dataset size. The majority of the studies use similar input parameters to those used in ASM-type models, ignoring the potential use of other parameters which are monitored in any bWWTP and not necessarily implemented in the mechanistic models; oxidation reduction potential (ORP), conductivity, turbidity, etc. Thus, yet, potentialities of data-driven methods are being ignored and on the other side, relevant information is omitted in most of the studies published.

As previously stated, the diversity of data sources in wastewater treatment is clear. However, the combination of these data sources for extraction of knowledge is not yet studied in bWWTP. Hence, the main goal of this doctoral dissertation is to increase the general understanding of the state of the art ML methods in wastewater treatment focusing on; i) heterogeneous datasets analysis, ii) the suitability of data-driven methods for these datasets and iii) novel approaches to extract novel knowledge from these datasets. This work demonstrates the importance of data selection in heterogeneous datasets to extract reliable information. The outcome of different data-driven methods change dramatically with different amount of data considered in analysis. This was evidenced when a municipal WWTP was studied. To solve this problem, a methodology to extract a significant subset out of a total raw heterogeneous dataset was developed; optimizing the size of the dataset. The definition of a score-function, allowed the optimization of a subset which was comprised by a set of representative parameters or features (and observations) and then applied to build highly accurate models. Although, feature engineering is a well-

developed field in data-science, not yet explored in wastewater treatment. New engineered features allowed to build highly accurate models for the prediction of complex bWWTP where data limitation was an issue. As well, an alternative methodology is proposed in this work to combine even more heterogeneous data sources to efficiently extract novel knowledge from complex bWWTP and that can be applied to similar complex bWWTP.

Although the contributions of this doctoral dissertation are important, yet the main limitation of this work is the extension of the analysis to similar processes i.e. to evaluate if the knowledge gained from the processes studied are particular to these systems or similar patterns eco in comparable processes, for example, do the patterns in all municipal WWTP are similar?

After showing the impact of the amount of data in different data-driven tasks. Existing data quality metrics for specific data sources in wastewater treatment (except for sensor data) need to be addressed, since are currently disconnected from the specific contextual characteristics. The need to revise data quality metrics for different sources of data in wastewater treatment is necessary, mainly when dealing with heterogeneous datasets. These issues however, are out of the focus of this work.

## Kurzfassung

Das maschinelle Lernen (ML) ist eines der am schnellsten wachsenden technischen Gebiete, das an der Schnittstelle von Informatik und Statistik liegt und den Kern der künstlichen Intelligenz (KI) und der Datenwissenschaft bildet. Die Anwendung von ML ist in einer Reihe von Branchen, die sich mit datenintensiven Themen befassen, wie z.B. Verbraucherservice, Bankenwesen, Astronomie und empirische Wissenschaften usw., weit verbreitet. Im Bereich der Abwasserbehandlung ging der Ursprung der umfangreichen Datengenerierung mit der Automatisierung von Kläranlagen (KA) einher. Zusätzlich ermöglichte eine Erhöhung der Rechen- und Speicherkapazität die Speicherung großer Mengen an Informationen aus verschiedenen Quellen auch im Wassersektor. Die von Kläranlagen erzeugten Informationen umfassen komplexe und heterogene Datenquellen; dazu zählen Daten von Onlinesensoren, on/off Steuerungsdaten von Pumpen und Geräten und Offline-Messungen in Laboratorien. Sensoren sind in der Lage im Sekundentakt Messwerte aufzuzeichnen und so täglich tausende von Datenpunkten zu generieren. Die Labordaten sind entscheidend für die Bewertung der Wasserqualität in den biologischen Stufen einer KA und oft auch für die Validierung der Sensorinformationen. Aufgrund der Kosten und des Zeitaufwands ist die Häufigkeit der Probenahmen für Labormessungen jedoch oft drastisch reduziert. Die daraus resultierende Datenbank (aus Sensor- und Labordaten) beinhaltet daher unterschiedliche Probenahmehäufigkeiten und enthält somit einen sehr heterogenen Datensatz.

Die aktuelle Forschung zu datengestützten Methoden in der Abwasserbehandlung hat sich hauptsächlich auf vorausschauende Aufgaben konzentriert, um die Abwasserzusammensetzung und die Leistung von KAs vorherzusagen, wobei letztere bisher weitgehend mit Belebtschlamm-Modellen (*activated sludge models*, ASM) untersucht werden. Obwohl das Ergebnis bei beiden Ansätzen ähnlich sein könnte, sind die Anwendung und die Eingabeinformationen zu den Modellen sehr unterschiedlich. Datengetriebene Ansätze benötigen genügend Daten, um eine Analyseaufgabe durchzuführen, sie sind datengetrieben. Die Natur der ASM-Modelle ist jedoch mechanistisch, d.h. sie zielen darauf ab, die biochemischen Wechselwirkungen zwischen der mikrobiellen Gemeinschaft im Abwassersystem und den Hauptschadstoffen im Abwasser - organische Substanz, Stickstoff, Phosphor und andere gelöste Stoffe - zu beschreiben. Beide Ansätze liefern nützliche und wichtige Informationen aus der Prozessleistung, es ist jedoch äußerst wichtig, die Unterschiede und Ziele von ASM-Modellen und ML-basierten Ansätzen im aktuellen Rahmen der Abwasserbehandlung zu unterscheiden und zu klären. Die Hauptgründe, warum diese Methoden bei Vorhersageaufgaben Anwendung finden sind zweifach: i) die Verfügbarkeit von Daten, die aus der Überwachung verschiedener Kläranlagen gewonnen wurden, und ii) die bereits erwähnte Komplexität der biologischen Prozesse. Die hohe Anpassungsfähigkeit von ML-Methoden an dynamische Systeme hat zu einer breiten Anwendung dieser Methoden geführt. Aus der Literatur geht jedoch eine Schlüsselfrage hervor. Die aktuellen Studien, die sich auf datengesteuerte Methoden in der Abwasserbehandlung beziehen, beschreiben nicht explizit die angewandten Datenaufbereitungsschritte, die Menge der für die Analyse verwendeten Daten, die für die Datenauswahl in Betracht gezogene Häufigkeit des Datenaufkommens und die Begründung für die Auswahl der Datensatzgröße. Die Mehrheit der Studien verwendet ähnliche Eingabeparameter wie in ASM-Modellen, wobei die potenzielle Verwendung anderer Parameter ignoriert wird, die in jeder Kläranlage überwacht und nicht unbedingt in den mechanistischen Modellen implementiert werden, z.B. Redox-Potential, Leitfähigkeit, Trübung usw. Somit werden die Möglichkeiten datengesteuerter Methoden ignoriert und andererseits werden relevante Informationen in den meisten der veröffentlichten Studien ausgelassen.

Wie bereits erwähnt, ist die Vielfalt der Datenquellen in der Abwasserbehandlung offensichtlich. Die Kombination dieser Datenquellen für die Wissensextraktion wird in der Kläranlage jedoch noch nicht untersucht. Daher ist das Hauptziel dieser Dissertation die Verbesserung des allgemeinen Verständnisses des Standes der Technik von ML-Methoden in der Abwasserbehandlung mit folgenden Schwerpunkten: i) Analyse heterogener Datensätze, ii) Eignung von datengetriebenen Methoden für diese Datensätze und iii) neue Ansätze zur Extraktion neuen Wissens aus diesen Datensätzen. Diese Arbeit zeigt die Bedeutung der Datenauswahl in heterogenen Datensätzen, um zuverlässige Informationen zu extrahieren. Die Ergebnisse verschiedener datengetriebener Methoden ändern sich dramatisch, wenn unterschiedliche Datenmengen bei der Analyse berücksichtigt werden. Dies wurde bei der Untersuchung einer kommunalen Kläranlage nachgewiesen. Um dieses Problem zu lösen, wurde eine Methode zur Extraktion einer signifikanten Teilmenge aus einem gesamten heterogenen Rohdatensatz entwickelt, wobei die Größe des Datensatzes optimiert wurde. Die Definition einer Score-Funktion ermöglichte die Optimierung einer Teilmenge, die aus einer Reihe repräsentativer Parameter oder Merkmale (und Beobachtungen) bestand und dann zur Erstellung hochgenauer Modelle angewandt wurde. Obwohl das Feature-Engineering ein gut entwickeltes Gebiet der Datenwissenschaft ist, ist es in der Abwasserbehandlung noch nicht erforscht. Dank neu entwickelter Merkmale konnten hochpräzise Modelle für die Vorhersage komplexer Kläranlagen erstellt werden, bei denen Datenbeschränkungen ein Problem darstellten. Außerdem wird in dieser Arbeit eine alternative Methodik vorgeschlagen, um noch heterogenere Datenquellen zu kombinieren, um auf effiziente Weise neues Wissen aus komplexen Kläranlagendaten zu extrahieren, das auf ähnliche komplexe Fälle angewendet werden kann.

Die Beiträge dieser Doktorarbeit stellen einen wichtigen Beitrag dar, jedoch mit der Einschränkung, dass in dieser Arbeit noch keine Anwendungen der Analysen auf ähnliche Systeme erfolgte. Zukünftig sollte daher beurteilt werden, ob die Erkenntnisse, die aus den untersuchten Prozessen gewonnen wurden, für diese Systeme oder für ähnliche Muster in vergleichbaren Prozessen typisch sind, z.B.: Sind bestimmte Muster in allen kommunalen Kläranlagen ähnlich?

## Resumen

El aprendizaje automático (o popularmente *Machine Learning, ML*) es uno de los campos técnicos de más rápido crecimiento, situado en la intersección de la informática y la estadística, y en el centro de la inteligencia artificial (IA) y la ciencia de los datos (*Data Science*). El efecto de ML se deja sentir ampliamente en toda una serie de industrias que se ocupan de ramos relacionados con la utilización intensiva de datos, como los servicios de consumo, bancarios, astronomía y ciencias empíricas, entre otras. En el campo del tratamiento de aguas residuales, el origen de la vasta generación de datos llegó junto con la automatización de las plantas de tratamiento de aguas residuales. Además, el aumento de la capacidad de almacenamiento, permitió generar grandes cantidades de información en el sector hídrico procedente de diferentes fuentes. La información de las plantas de tratamiento de aguas residuales que se genera y registra, implica fuentes de datos complejas y heterogéneas; datos procedentes de sensores generados en línea, datos de control de equipos; *on/off* y mediciones *off-line*, datos generados en laboratorios de análisis de muestras.

Los sensores son capaces de registrar mediciones cada pocos segundos, generando así miles de puntos de datos diariamente. Los datos generados en laboratorios son cruciales para evaluar la calidad del agua en cualquier proceso de tratamiento biológico de aguas residuales (*bWWTP; biological Wastewater Treatment Plants*) y a menudo para validar la información de los sensores. Sin embargo, debido a los costos y el tiempo que implica, la frecuencia de muestreo para las mediciones de laboratorio suele reducirse drásticamente en comparación con los sensores. Consequentemente, la base de datos resultante (de los sensores y los laboratorios), implica frecuencias de muestreo variables y por lo tanto un conjunto de datos muy heterogéneos.

Las investigaciones actuales sobre métodos basados en datos o *Machine Learning* en el tratamiento de aguas residuales se han centrado principalmente en tareas de predicción, para pronosticar la composición de los efluentes y el rendimiento de los diferentes procesos biológicos de tratamiento de aguas residuales, estas últimas también se han estudiado ampliamente mediante modelos de lodos activados (ampliamente conocidos como *activated sludge models; ASM*). Si bien el resultado es similar con ambos enfoques, la aplicación y la información de entrada a los modelos es muy diferente. Los enfoques basados en *Machine Learning* requieren datos suficientes para realizar una tarea de análisis, son impulsados por datos. Sin embargo, la naturaleza de los modelos mecanísticos (ASM) es fenomenológica, que tiene por objeto describir la interacción bioquímica entre la comunidad microbiana del sistema de aguas residuales y los principales contaminantes de las aguas residuales; materia orgánica, nitrógeno, fósforo y otros nutrientes disueltos. Ambos enfoques proporcionan información útil e importante sobre el funcionamiento del proceso, pero es sumamente importante distinguir y aclarar las diferencias y objetivos de los modelos del tipo ASM y las tareas basadas en *Machine Learning* en el marco actual del tratamiento de aguas residuales.

Las principales razones que motivaron a la comunidad de tratamiento de aguas residuales a aplicar métodos basaods en datos en las tareas de predicción son dos: i) es la disponibilidad de bases de datos del monitoreio de diferentes procesos biológicos relacionados al tratamiento de aguas residuales y ii) la ya mencionada complejidad de los procesos biológicos. La gran adaptabilidad de los métodos basados en datos a los sistemas dinámicos ha llevado a la comunidad de investigadores a una amplia aplicación de estos métodos. Sin embargo, de la bibliografía surge una cuestión clave. Los estudios actuales relacionados con los métodos basados en datos en el tratamiento de aguas residuales no describen explícitamente las técnicas de pretratamiento aplicadas, la cantidad de datos utilizados para el análisis, la frecuencia considerada para la selección de los datos y el fundamento de la selección del tamaño del conjunto de datos, entre otros.

Después de una revisión rigurosa de los estudios en la literatura, la mayoría de los estudios relacionados con modelos basados en *Machine Learning* utilizan parámetros de entrada similares a los utilizados en los modelos mecanísticos, ignorando el uso potencial de otros parámetros como; potencial de reducción de oxidación, conductividad, turbidez, etc. Como consecuencia, se ignoran las potencialidades de los métodos basados en datos y, por otra parte, se omite información pertinente en la mayoría de los estudios publicados, limitando el alcance de estos modelos y estudios.

La diversidad de fuentes de datos en el tratamiento de aguas residuales es evidente, previamente establecido.

Sin embargo, la combinación de estas fuentes de datos y su uso para la extracción de conocimiento no se ha estudiado aún en el campo de tratamiento biológico de aguas residuales. Por lo tanto, el objetivo principal de esta tesis doctoral es aumentar la comprensión general del estado de arte los métodos de *Machine Learning* en el tratamiento de aguas residuales, centrándose en: i) el análisis de bases datos heterogéneos en este campo y, ii) la adaptabilidad de los métodos basados *Machine Learning* para bases de datos heterogéneos y iii) metodologías innovadoras para la extracción de información relevante de diferentes procesos biológicos relacionados con tratamiento de aguas residuales.

Este trabajo demuestra la importancia de la metodología aplicada para la extracción de información en bases de datos heterogéneos. El resultado de los diferentes métodos basados en datos cambia drásticamente con la cantidad de datos considerados en el análisis. Lo último fue demostrado en este trabajo en el estudio de una planta de tratamiento de aguas residuales municipal. Para resolver este problema, se desarrolló una metodología para extraer un subconjunto significativo de un conjunto de datos heterogéneos; optimizando el tamaño del conjunto de datos.

La definición de una función de *scoring*, permitió la optimizayción un dataset heterogéneo; número de parámetros y observaciones. Esta función fue aplicada en la construcción de modelos de gran precisión, demostrando la importancia de pasos posteriores en la obtención de un modelo adecuado.

Aunque el campo de *feature engineering* es un campo desarrollado el área de ciencia de los computadoras, es aún un campo no explorado en el tratamiento de aguas residuales. En este trabajo doctoral, *feature engineering* permitió construir modelos altamente precisos para la predicción de sistemas complejos donde existía limitación de datos. Además, en este trabajo se propone una metodología innovadora que permite combinar conjuntos de datos procedientes de diferentes fuentes para la extracción eficiente de conocimiento nuevo de procesos biológicos relacionados con tratamiento de aguas reiduales, estos métodos pueden ser aplicados a sistemas biológicos similares.

Al demostrar el impacto de la cantidad en diferentes tareas basadas en datos en el área de tratamiento de aguas residuales. Es necesario abordar la métrica de calidad de los datos existentes para fuentes de datos específicas en el tratamiento de aguas residuales (excepto los datos de los sensores), ya que actualmente están desconectados de las características contextuales específicas. Es necesario revisar las métricas de calidad de los datos para diferentes fuentes de datos en el tratamiento de aguas residuales, principalmente cuando se trata de conjuntos de datos heterogéneos. Sin embargo, estos puntos no son objeto de este trabajo.

## Contents

## List of Abbreviations

| Abbreviation | Description | Units |
|---|---|---|
| AC eff. | acid capacity effluent | mmol $l^{-1}$ |
| AC inf. | acid capacity influent | mmol $l^{-1}$ |
| AC inf. (AS) | acid capacity influent activated sludge (AS) | mmol $l^{-1}$ |
| ADM1 | anaerobic digestion model 1 | - |
| AF | aeration fraction | - |
| Air flow (L1) | air flow AS stage 1 | $Nm^3$ $h^{-1}$ |
| Air flow (L2) | air flow AS stage 2 | $Nm^3$ $h^{-1}$ |
| Air flow (L3) | air flow AS stage 3 | $Nm^3$ $h^{-1}$ |
| Air temp. | air temperature | $^oC$ |
| Anammox | Anaerobic ammonium oxidation | - |
| ANFIS | adaptive neuro fuzzy interference systems | - |
| ANN | artificial neural networks | - |
| AOB | aerobic oxidizing bacteria | - |
| ASM1 | Activated sludge model 1 | - |
| ASM2 | Activated sludge model 2 | - |
| ASM2d | Activated sludge model 2d | - |
| ASM3 | Activated sludge model 3 | - |
| ATP | adenosine triphosphate | - |
| BOD | Biological oxygen demand | - |
| BOD5 eff. (24h) | five day BOD concentration 24h mixed sample | mg $l^{-1}$ |
| BOD5 eff. (2h) | five day BOD effluent concentration 2h mixed sample | mg $l^{-1}$ |
| BOD5 inf. | five day BOD influent concentration | mg $l^{-1}$ |
| BOD5 inf. (AS) | five day BOD influent concentration to AS | mg $l^{-1}$ |
| BWTP | Biological wastewater treatment processes | - |
| CART | classification and regression trees | - |
| CFD | computational fluid dynamics | - |
| CNN | convolutional neural networks | - |
| Coag. Dos. (L1) | coagulant (FeCl3) dosage concentration (L1) | l $d^{-1}$ |
| Coag. Dos. (L2) | coagulant (FeCl3) dosage concentration (L2) | l $d^{-1}$ |
| Coag. Dos. (L3) | coagulant (FeCl3) dosage concentration (L3) | l $d^{-1}$ |
| COD | Chemical oxygen demand | - |
| COD eff. (24h) | COD effluent concentration (24h mixed sample) | mg $l^{-1}$ |
| COD eff. (2h) | COD effluent concentration (2h mixed sample) | mg $l^{-1}$ |
| COD eff. (online) | COD effluent concentration online sensor | mg $l^{-1}$ |
| COD inf. | COD influent concentration | mg $l^{-1}$ |
| COD inf. (AS) | COD influent concentration to AS | mg $l^{-1}$ |
| Cond. Inf. | conductivity in the influent | mS $cm^{-1}$ |
| CV | Control volume | - |
| DBSCAN | Density-Based Spatial Clustering and Application | - |
| DNN | Deep neural networks | - |
| DO | dissolved oxygen | - |

| | | |
|---|---|---|
| DO M1 (L1) | dissolved oxygen (DO) concentration measurment 1 (M1) in L1 | mg $l^{-1}$ |
| DO M1 (L2) | DO concentration M1 in L2 | mg $l^{-1}$ |
| DO M1 (L3) | DO concentration M1 in L3 | mg $l^{-1}$ |
| DO M2 (L1) | DO concentration M1 in L3 | mg $l^{-1}$ |
| DO M2 (L2) | DO concentration measurement No.2 (M2) in L2 | mg $l^{-1}$ |
| DO M2 (L3) | DO concentration M2 in L3 | mg $l^{-1}$ |
| EBPR | enhanced biological phosphorus removal processes | - |
| GAOs | glycogen accumulating organisms | - |
| HB | heterotrophic bacteria | - |
| IAWPRC | International Association on Water Pollution Research and Control | - |
| ICA | instrumentation control and automation technology | - |
| Influent flow (L1) | Influent flow to L1 | $m^3$ $d^{-1}$ |
| Influent flow (L2) | Influent flow to L2 | $m^3$ $d^{-1}$ |
| Influent flow (L3) | Influent flow to L3 | $m^3$ $d^{-1}$ |
| IWA | International water association | - |
| load BOD5 eff. | BOD load effluent | kg $d^{-1}$ |
| load BOD5 inf. | BOD load influent | kg $d^{-1}$ |
| load Coag. Dos. (L1) | load coagulant dosage in L1 | kg $d^{-1}$ |
| load Coag. Dos. (L2) | load coagulant dosage in L2 | kg $d^{-1}$ |
| load Coag. Dos. (L3) | load coagulant dosage in L3 | kg $d^{-1}$ |
| load COD eff. | COD load effluent | kg $d^{-1}$ |
| load COD inf. | COD load influent | kg $d^{-1}$ |
| load $NH_4$-N eff. | $NH_4$-N load effluent | kg $d^{-1}$ |
| load $NH_4$-N inf. | $NH_4$-N load influent | kg $d^{-1}$ |
| load $NO_2$-N eff. | $NO_2$-N load effluent | kg $d^{-1}$ |
| load $NO_2$-N inf. | $NO_2$-N load influent | kg $d^{-1}$ |
| load $NO_3$-N eff. | $NO_3$-N load effluent | kg $d^{-1}$ |
| load $NO_3$-N inf. | $NO_3$-N load influent | kg $d^{-1}$ |
| load $PO_4$-P eff. (online) | $PO_4$-P load effluent online sensors | kg $d^{-1}$ |
| load PP eff. (online and lab) | particulate phosphorus load effluent from online and lab | kg $d^{-1}$ |
| load PP eff. (online) | particulate phosphorus load effluent online | kg $d^{-1}$ |
| load SS eff. | suspended solids (SS) load effluent | kg $d^{-1}$ |
| load TKN eff. | total Kjekdahl nitrogen load effluent | kg $d^{-1}$ |
| load TKN inf. | total Kjekdahl nitrogen load influent | kg $d^{-1}$ |
| load TN eff. | total nitrogen load effluent | kg $d^{-1}$ |
| load TN inf. | total nitrogen load influent | kg $d^{-1}$ |

| | | |
|---|---|---|
| load TN inorg. eff. | total inorganic nitrogen load effluent | kg $d^{-1}$ |
| load TN inorg. inf. | total inorganic nitrogen load influent | kg $d^{-1}$ |
| load TP eff. | total phosphorus load effluent | kg $d^{-1}$ |
| load TP eff. (online) | total phosphorus load effluent from online sensor | kg $d^{-1}$ |
| load TP inf. | total phosphorus load influent | kg $d^{-1}$ |
| LOI eff. (L1) (AS) | loss on ignition (LOI) concentration effluent AS in L1 | % |
| LOI eff. (L2) (AS) | LOI concentration effluent AS in L2 | % |
| LOI eff. (L3) (AS) | LOI concentration effluent AS in L3 | % |
| LTU | linear threshold unit | - |
| MABR | membrane aerated biofilm reactors | - |
| MAR | Missing At Random | - |
| MBBR | membrane bed biofilm reactors | - |
| MCAR | Missing Completely At Random | - |
| ML | machine learning | - |
| MNAR | Missing not at random | - |
| MPC | model predictive control | - |
| MSE | Mean squared error | - |
| NADH | nicotinamide- adenine dinucleotide | - |
| N-DN | nitrification and denitrification | - |
| $NH_4$-N eff. (24h) | $NH_4$-N concentration in effluent (24h mixed sample) | mg $l^{-1}$ |
| $NH_4$-N eff. (2h) | $NH_4$-N concentration in effluent (2h mixed sample) | mg $l^{-1}$ |
| $NH_4$-N eff. (online) | $NH_4$-N concentration in effluent from online sensor | mg $l^{-1}$ |
| $NH_4$-N inf. | $NH_4$-N concentration in influent (lab) | mg $l^{-1}$ |
| $NH_4$-N inf. (AS) | $NH_4$-N concentration in influent of AS | mg $l^{-1}$ |
| $NH_4$-N inf. (AS) (online) | $NH_4$-N concentration in influent of AS from online sensor | mg $l^{-1}$ |
| NN | neural networks | - |
| $NO_2$-N eff. (24h) | $NO_2$-N concentration in effluent (24h mixed sample) | mg $l^{-1}$ |
| $NO_2$-N eff. (2h) | $NO_2$-N concentration in effluent (2h mixed sample) | mg $l^{-1}$ |
| $NO_2$-N inf. | $NO_2$-N concentration in influent | mg $l^{-1}$ |
| $NO_2$-N inf. (AS) | $NO_2$-N concentration in influent of AS | mg $l^{-1}$ |
| $NO_3$-N eff. (24h) | $NO_3$-N concentration in effluent (24h mixed sample) | mg $l^{-1}$ |
| $NO_3$-N eff. (2h) | $NO_3$-N concentration in effluent (2h mixed sample) | mg $l^{-1}$ |
| $NO_3$-N inf. | $NO_3$-N concentration in influent | mg $l^{-1}$ |
| $NO_3$-N inf. (AS) | $NO_3$-N concentration in influent of AS | mg $l^{-1}$ |
| NOB | nitrite oxidizing bacteria | - |
| NOx-N eff. (online) | $NO_3$-N +$NO_2$-N concentration in effluent online sensor | mg $l^{-1}$ |
| ORP | Oxidation Reduction Potential | mV |
| PCA | Principal component analysis | - |

| | | |
|---|---|---|
| pH (L2) | pH in stage 2 (L2) | - |
| pH eff. | pH effluent | - |
| pH inf. | pH influent | - |
| pH inf. (AS) | pH influent of AS | - |
| PHA | poly-hydroxy-alkanoates | - |
| PLC | programmable logic controllers | - |
| PN-A | Partial nitritation anammox | - |
| $PO_4$-P eff. (online) | $PO_4$-P concentration in effluent from online sensor | mg $l^{-1}$ |
| $PO_4$-P inf. (AS) (online) | $PO_4$-P concentration in influent of AS from online sensor | mg $l^{-1}$ |
| RAS flow (L1) | Return activated sludge flow in L1 | $m^3$ $d^{-1}$ |
| RAS flow (L2) | Return activated sludge flow in L2 | $m^3$ $d^{-1}$ |
| RBC | rotating biological contactor | - |
| RBOM | readily biodegradable organic matter | - |
| Recirc. Flow (L1) | Recirculation flow in L1 | $m^3$ $d^{-1}$ |
| Recirc. Flow (L2) | Recirculation flow in L2 | $m^3$ $d^{-1}$ |
| Recirc. Flow (L3) | Recirculation flow in L3 | $m^3$ $d^{-1}$ |
| RFE | recursive feature elimination | - |
| RTU | remote technical units | - |
| SAD | simultaneous anaerobic ammonium oxidation and heterotrophic denitrification process | |
| SBOM | slowly biodegradable organic matter | - |
| SBR | sequencing batch reactor | - |
| SCADA | Supervisory Control And Data Acquisition | - |
| SHARON | Single reactor High activity Ammonia Removal over Nitrite | - |
| SS eff. (online) | Suspended solids concentration effluent from online sensor | mg $l^{-1}$ |
| SSV eff. (L1) (AS) | Settled sludge volume effluent of AS in L1 | ml $l^{-1}$ |
| SSV eff. (L2) (AS) | Settled sludge volume effluent of AS in L2 | ml $l^{-1}$ |
| SSV eff. (L3) (AS) | Settled sludge volume effluent of AS in L3 | ml $l^{-1}$ |
| SVI eff. (L1) (AS) | Sludge volume index effluent in L1 | ml $g^{-1}$ |
| SVI eff. (L2) (AS) | Sludge volume index effluent in L2 | ml $g^{-1}$ |
| SVI eff. (L3) (AS) | Sludge volume index effluent in L3 | ml $g^{-1}$ |
| SVR | Support vector regression | - |
| Temp. (L2) | temperature L2 | $^oC$ |
| Temp. eff. | temperature effluent | $^oC$ |
| Temp. inf. | temperature influent | $^oC$ |
| TKN | total Kjeldahl nitrogen | - |
| TKN eff. (24h) | TKN concentration effluent (24h mixed sample) | mg $l^{-1}$ |
| TKN eff. (2h) | TKN concentration effluent (2h mixed sample) | mg $l^{-1}$ |
| TKN inf. | TKN concentration influent | mg $l^{-1}$ |

| | | |
|---|---|---|
| TKN inf. (AS) | TKN concentration influent in AS | mg l$^{-1}$ |
| TN eff. (24h) | total nitrogen effluent concentration (24h mixed sample) | mg l$^{-1}$ |
| TN eff. (2h) | total nitrogen effluent concentration (2h mixed sample) | mg l$^{-1}$ |
| TN inf. | total nitrogen influent concentration | mg l$^{-1}$ |
| TN inf. (AS) | total nitrogen influent concentration of AS | mg l$^{-1}$ |
| TN inorg. eff. (24h) | total inorganic nitrogen concentration in effluent (24h mixed sample) | mg l$^{-1}$ |
| TN inorg. eff. (2h) | total inorganic nitrogen concentration in effluent (2h mixed sample) | mg l$^{-1}$ |
| TN inorg. Inf. | total inorganic nitrogen concentration in influent | mg l$^{-1}$ |
| TN inorg. inf. (AS) | total inorganic nitrogen concentration in influent of AS | mg l$^{-1}$ |
| TOC eff. (online) | total organic carbon concentration in the effluent from online sensor | mg l$^{-1}$ |
| Total Coag. Dos. | total coagulant dosage (from three stages: L1-L3) | l d$^{-1}$ |
| Total eff. | total effluent flow | m$^3$ d$^{-1}$ |
| Total eff. (DW) | total effluent flow in dryweather | m$^3$ d$^{-1}$ |
| Total eff. (min 21D) | total effluent flow minimum of 21 days | m$^3$ d$^{-1}$ |
| Total inf. | total influent flow | m$^3$ d$^{-1}$ |
| Total load Coag. Dos. | total load of coagulant dosage | kg d$^{-1}$ |
| Total load Coag. Dos. (mol) | total load coagulant dosage (in mol) | mol d$^{-1}$ |
| TP eff. (24h) | total phosphorus effluent concentration (24h mixed sample) | mg l$^{-1}$ |
| TP eff. (2h) | total phosphorus effluent concentration (2h mixed sample) | mg l$^{-1}$ |
| TP eff. (online) | total phosphorus effluent concentration from online sensor | mg l$^{-1}$ |
| TP inf. | total phosphorus concentration influent | mg l$^{-1}$ |
| TP inf. (AS) | total phosphorus concentration influent AS | mg l$^{-1}$ |
| TS (Exess sludge flow) | total solids concentration in the excess sludge flow | mg l$^{-1}$ |
| TS eff. (L1) (AS) | total solids concentration in the effluent of AS in L1 | mg l$^{-1}$ |
| TS eff. (L2) (AS) | total solids concentration in the effluent of AS in L2 | mg l$^{-1}$ |
| TS eff. (L3) (AS) | total solids concentration in the effluent of AS in L3 | mg l$^{-1}$ |
| TSS | Total Suspended Solids | - |
| weather key | weather identification key | - |
| wwt | wastewater treatment | - |
| WWTP | wastewater treatment plants | - |

## List of Figures

## List of Tables

# 1 Introduction

Machine learning (ML) is one of the most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence (AI) and data science (Jordan and Mitchell, 2015). The effect of ML is broadly felt across a range of industries concerned with data intensive issues, such as consumer services, astronomy and empirical sciences, among others. In the field of wastewater treatment, the origin of vast data generation came along with automation of wastewater treatment plants (WWTP). Additionally, an increase of the computing and storage capacity, allowed large amounts of information to be generated in the water sector coming from different sources to be stored. The information from WWTP, that is generated and recorded commonly in SCADA (SCADA: an acronym for Supervisory Control and Data Acquisition) systems, involves complex and heterogeneous data sources; on-line from sensors, on/off control data from pumps and equipment and off-line measurements from laboratories. These large databases allow operators and engineers to monitor individual equipment and processes performance as well (and most important), the water quality to comply with environmental regulations. However, further extraction of knowledge from these databases is challenging without the aid of advanced statistical tools and most important, a clear methodology to extract actionable and reliable knowledge. Data driven methods based on ML provide a wide variety of powerful tools for analysis. In this work, these methods have been applied to analyze high throughput experimental data sources in wastewater treatment in novel ways.

## 1.1 Objectives

The value that advanced data-driven methods bring to the field of wastewater treatment comes from expanding the range of data sources analyzed and improving the quality of the analysis for reliable results. When widening the variety of data sources, the analysis requires more effort than for traditional data sources; robotics, computer vision, among others. In these fields, the rate to which the data is generated is often uniform and orders of magnitude higher than for the water sector (except for sensors). For example, the *FaceForensics++* dataset –a known dataset used for evaluation of new methods for face recognition –, contains over 1.8 million images that could be used for training and validation (Rössler et al., 2019) In wastewater treatment processes, the density of data is considerably different. Sensors are able to record measurements every few seconds, thus, generating thousands of data points daily. The data generated in laboratories in wastewater treatment is crucial to evaluate the quality of the water in any biological wastewater treatment process and often to validate the sensors information. However, due to the costs and time involved, the frequency of sampling for laboratory measurements is often dramatically reduced compared to sensors. Thus, the resulting database (from sensors and laboratories), involve varying frequencies of sampling and thus a highly heterogeneous dataset. Additionally, the suitability and implementation of the methods for these data sources differs to what is observed in the traditional fields mentioned. After a thorough revision of the literature related to data-driven applications in wastewater treatment, the suitability and methodology for the application of these methods in wastewater treatment is questioned. Up to date, the literature has not addressed the nature of data sources in wastewater treatment, and accordingly, the suitability of the methods for these different sources of data to make the most out of the data mining process. The consequences of handling highly heterogeneous datasets lead to more profound issues of data analysis such as the problem of missing values and data limitation for analysis, which are not explored in the current framework of wastewater treatment. Therefore, an important contribution from this work is to develop a methodology based on the different data sources found in wastewater treatment and how to combine them for knowledge extraction, starting from describing the data generated in wastewater treatment to end-to-end applications of supervised and or unsupervised ML methods, knowledge extraction and analysis.

Based on the outcomes of earlier research, six major research questions (RQ) are built at the start of this work:

**RQ1:** It is essential to distinguish and clarify the differences and goals of activated sludge models (ASM-type) models and ML-based tasks in the current framework of wastewater treatment. What is the state of the art towards the application of data-driven methods in the water sector compared to mechanistic approaches? Which are the limitations of both approaches in the water sector?
**RQ2:** The data sources in wastewater treatment have different natures; online from sensors, on/off data from equipment and off-line data from laboratories. For one particular period of operation, the amount of information gathered from these sources is significantly different resulting in a heterogeneous dataset; the amount of data points differs from parameter to parameter. How sensitive are the data-driven methods to the amount of data and parameters considered for analysis? How the results from a data-driven task will change with different sizes of data?
**RQ3:** Following RQ2. Which subset from the total raw data collected would be the most significant for further data-driven tasks? How this subset can be selected? Is it possible to optimize both the parameters considered for prediction (input to the model) and the amount of information (size of the dataset)?
**RQ4:** Following RQ3. When dealing with heterogeneous datasets. How these datasets can be combined with even small datasets; in biological wastewater treatment processes a good example are biomass batch activity tests. How these different sources of data can be combined to create a significant dataset and which tools can be applied to extract knowledge from it?
**RQ5:** Following RQ3. Given heterogeneous datasets from a bWWTP with limited amount of data, and additionally, a process that is not yet well studied from a mechanistic modeling perspective. How can ML methods be applied to extract knowledge and model this complex bWWTP?

## 1.2 Outline

To provide a structured answer to the aforementioned research questions, this work is composed of multiple chapters. Chapter 2 presents an update and thorough revision of the current trends regarding ASM modeling and data-driven methods based on ML, focusing on the differences of both approaches for the study of wastewater treatment processes. Chapter 3 is focused on a detailed description of data generation in wastewater treatment, sources of data, pre-processing of the different databases in wastewater treatment and dealing with problems such as handling missing values and finally, state of the art of unsupervised and supervised ML methods which will be applied along the following chapters of this work. Chapter 4 centers on the integration of different sources of data and the extraction of knowledge a municipal WWTP. In this chapter, the problem of how data analyses are affected by the amount and structure of data considered in the analysis is addressed. In Chapter 5, feature selection methods are optimized towards the selection of the best configuration of features and number of observations applied in the effluent prediction of two full-scale partial nitritation-Anammox sequencing batch reactors. Special focus on feature selection and feature engineering processes are discussed. Chapter 6 illustrates a methodology for knowledge extraction of combined heterogeneous datasets. This methodology was applied to lab and full scale partial nitritation-anammox systems. Chapter 7 shows the application of ensemble learning to extract knowledge and build predictive models in advanced wastewater treatment, where few to non-existing mathematical methods are available. Finally, Chapter 8 summarizes the main outcomes and contributions of this thesis, highlighting future perspectives and challenges in the current framework of data-science in wastewater treatment. Figure 1.1 summarizes the outline of this work along the chapters.

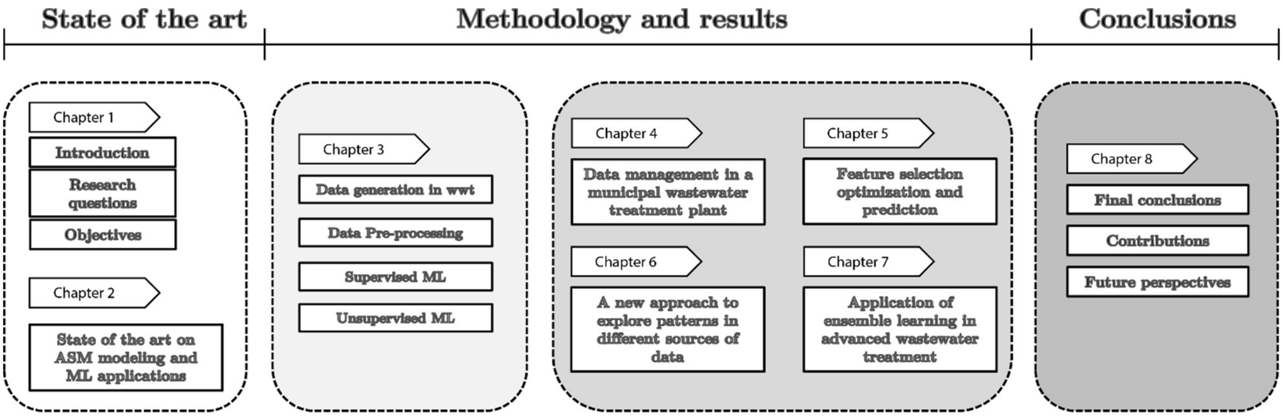

State of the art
Methodology and results
Conclusions
Chapter 1
Introduction
Research questions
Objectives
Chapter 2
State of the art on ASM modeling and ML applications
Chapter 3
Data generation in wwt
Data Pre-processing
Supervised ML
Unsupervised ML
Chapter 4
Data management in a municipal wastewater treatment plant
Chapter 5
Feature selection optimization and prediction
Chapter 6
A new approach to explore patterns in different sources of data
Chapter 7
Application of ensemble learning in advanced wastewater treatment
Chapter 8
Final conclusions
Contributions
Future perspectives

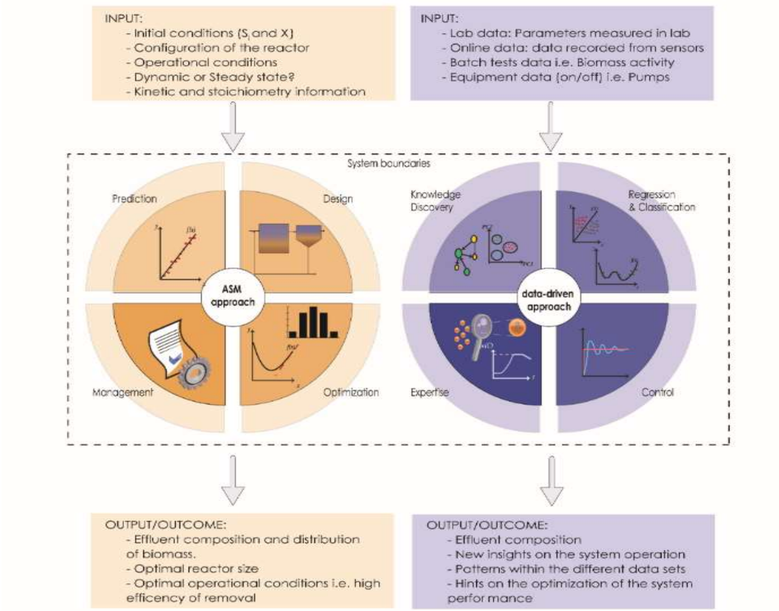
Figure 1. 1 Outline and distribution in this work

# 2 Literature Review

## 2.1 Introduction

Biological wastewater treatment processes (bWWTP) are complex systems where several chains of interconnected biological reactions occur simultaneously. With the aim of biologists, chemists and engineers to understand the nature of these processes and optimize them, complex mathematical models were developed over the years. The activated sludge model 1 (ASM1) was the first internationally coordinated effort for standardizing activated sludge modeling. ASM1 considers carbon oxidation and nitrogen removal from wastewater through biological nitrification and denitrification (N-DN). This mathematical model was developed by a group in the International Water Association (IWA) (Henze et al., 2000). The ASM1 has demonstrated to be essential in understanding biological nutrient removal in many WWTP and has been applied to lab, pilot and full-scale systems. Some years after ASM1 was released, enhanced biological phosphorus removal processes (EBPR) were modeled through ASM2. These initial models were well received by the community. Experiences of engineers, researchers and users with these models were key to propose improvements to ASM1 and ASM2, as a result ASM3 and ASM2d were developed. The constant reinforcement of these models is still proposing modifications and improvements today, which go along with technology advances that allow a more detailed study of the complex systems in bWWTP.

On the other hand, after the first applications of artificial intelligence or AI in different fields of science, the interest and potentiality in the field of wastewater treatment has increased exponentially in the past two decades (Corominas et al., 2018). The application of supervised ML techniques to forecast the effluent composition of biological wastewater treatment processes and the development of *intelligent* controllers are by far the most common applications. Data-driven approaches aim to model and extract new information from the operational data gathered along the operation of a biological process without detailed information of the kinetics of the process. In data-driven approaches, the key is the acquisition of data. In wastewater treatment these comprise lab data (parameters analyzed through analytical methods), online data (mainly data from sensors), on/off Boolean data from equipment and sometimes, data from batch experiments such as respirometry tests in biomass. The output/outcome from the data-driven approaches go from optimization of operational conditions to modeling the effluent composition of a bWWTP. Figure 2.1 illustrates the main components in each modeling approach and main characteristics of the input information and which results can be obtained from both.



INPUT:
- Initial conditions ($S_i$ and X)
- Configuration of the reactor
- Operational conditions
- Dynamic or Steady state?
- Kinetic and stoichiometry information

INPUT:
- Lab data: Parameters measured in lab
- Online data: data recorded from sensors
- Batch tests data i.e. Biomass activity
- Equipment data (on/off) i.e. Pumps

OUTPUT/OUTCOME:
- Effluent composition and distribution of biomass.
- Optimal reactor size
- Optimal operational conditions i.e. high efficency of removal

OUTPUT/OUTCOME:
- Effluent composition
- New insights on the system operation
- Patterns within the different data sets
- Hints on the optimization of the system perfor mance

Figure 2. 1 Comparison between ASM type models and data-driven approaches

Although the outcome could be similar with both approaches; prediction of the effluent, the application and the input information to the models is very different. Data-driven approaches require enough data to perform an analysis task, they are data driven. However, the nature of ASM models is phenomenological, which aims to describe the biochemical interaction between the microbial community in the wastewater system and main pollutants in the wastewater; organic matter, nitrogen, phosphorus and other dissolved nutrients. Both approaches provide useful information from the process performance.

In this chapter, the fundamental concepts underlying activated sludge models, the main problem they face and their evolution are discussed. Along, in this chapter, state of the art applications of data-driven methods in wastewater treatment are discussed.

## 2.2 The activated sludge models

The main bWWTP studied through ASM models comprise: i) organic matter degradation: biological oxygen demand (BOD) reduction, often expressed in terms of COD ii) nitrogen removal through N-DN and, iii) biological phosphorus (P) removal. The nitrification is an autotrophic process i.e. no external source of carbon is needed in this first stage. The nitrification process involves the oxidation of ammonia ($NH_4^+$) further to nitrate ($NO_3^-$), with molecular oxygen as the electron acceptor. In this process, ammonia oxidation to nitrite is mediated by aerobic oxidizing bacteria (AOB), some common species are; *Nitrosomonas europea, Nitrosomonas eutropha, and Nitrosomonas halophila* and *Nitrospira spp.,* while the nitratation process is mediated by nitrite oxidizing bacteria (NOB); *Nitrobacter spp.* (Vlaeminck et al., 2010).

Denitrification is a process by which nitrate and nitrite are reduced to di-nitrogen gas ($N_2$). The anoxic biological denitrification is accomplished with electron donors coming from organic matter (carbon sources: mainly methanol, acetate and ethanol) (Cheremisinoff, 1997). The denitrification can be held-on by different heterotrophic denitrifying bacteria which involves the consumption of nitrite ($NO_2^-$) and/or nitrate ($NO_3^-$). There are numerous genera of denitrifying bacteria identified in activated sludge; *Achromobacter, Escherichia, Neisseria, Acinetobacter, Flavobacterium, Paracoccus, Gluconobacer, Propionibacterium, Pseudomonas, Bacillus, Thiobacillus, Enterobacter* among others (Ni et al., 2016). Some common species studied in pure cultures are; *Pseudomonas denitrificans, Paracoccus denitrificans, Thiobacillus denitrificans, Comamonas denitrificans.* In activated sludge, around 20-80% of all flocculated biomass and in suspension belongs to heterotrophic denitrifying bacteria (Ni et al., 2016).

On the other hand, the EBPR is one of the most complex wastewater treatment processes due to the phosphorus based compounds participation in the metabolism of the microorganisms (internal stored substrates and products) (Smolders et al., 1995). In a biological phosphorus removal system, phosphate accumulating organisms (PAOs) are enriched and they accumulate large quantities of polyphosphate (poly-P) in their cells and thus enhance the biological phosphorus removal from wastewater (Cosenza et al., 2013). The PAOs have a strict requirement of cyclic anaerobic, anoxic and aerobic conditions, which consequently makes EBPR a more complex ,process compared to the nitrogen (N) and organic matter (COD) removal (Zuthi et al., 2013).

ASM type models are the mathematical representation of these complex systems, they require a proper structural identification, knowledge of the process dynamics, a detailed characterization of the substrates, and the optimization and validation of the kinetic and stoichiometric parameters. The main ASM applications comprise: i) prediction of the effluent composition and/or establishment of key parameters in the process (sensitivity analysis), ii) optimization of the operational conditions of the process, iii) study and evaluation of different operational scenarios and process configurations, and iv) process control applications.

Mathematical models for bWWTP, such as the ASM type models, were developed with the aim to understand the phenomena and dynamics in the activated sludge process. All ASM models are based on Monod kinetics to predict the consumption/production of dissolved substrates; chemical oxygen demand (COD), nitrogen, phosphorus and dissolved oxygen (DO) and generation/decay of biomass and other particulate matter ($X_N$, $X_S$)(Monod, 1949). As a result, the model predicts the composition of the effluent over time for the incoming substrates and the newly formed products from the biochemical reactions occurring in the system.

In order to understand the fundamentals of how the models are built, it is necessary to introduce the concept of *control volume* (CV). ASM models are based on mass balances; the concept of the mass balance in a defined control volume approach is illustrated in Figure 2.2. For any mass balance, it is fundamental to define the system boundaries which then define the CV. The boundaries of the CV will define the transformation of the species contained in this system i.e. the CV could be defined either with a recycle to the system in or out of the CV, and as a result, the equations describing the mass balances will change. Within the CV, the species can be transformed into new species (generation) and therefore, the old species will disappear (consumption). Both of these phenomena occur within the boundaries of the CV. The system also experiences perturbations from external sources. In Figure 2.2, the system experiences mass flow into the CV, this mass flow is composed of different species which could or could not interact with the system (reactive or inert matter). The system also experiences loss of mass due to convection. The environment where the ASM are defined comprise three physical phases; solid, liquid and gaseous and therefore, it is a heterogeneous reaction problem (Levenspiel, 1999). The general balance equations in ASM models will be applied to both particulate (Xi) and dissolved (Si) species (Equation 2.1).

$$Acc._{CV} = \dot{M}_{CV_{in}} - \dot{M}_{CV_{out}} + Gen._{CV} - Cons._{CV} \qquad \text{Equation 2. 1}$$

where $\dot{M}_{CV_{in}}$ is the mass flow rate entering the CV and $\dot{M}_{CV_{out}}$ is the mass flow rate leaving the CV, the terms $Gen._{CV}$ and $Cons._{CV}$, are the generation and consumption rates in the CV.

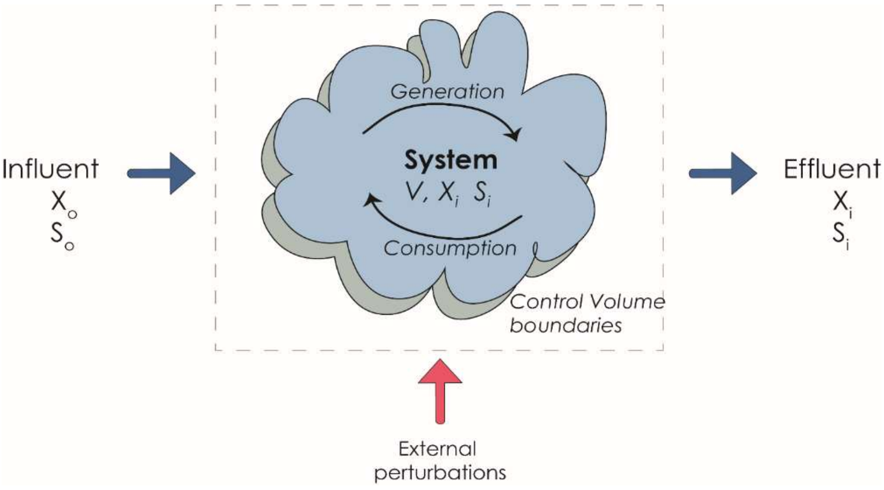


Figure 2. 2 Diagram of the system and control volume (CV). The irregularity of the System aims to illustrate the complexity of the process in study: continuous, batch and semi-discontinuous.

In ASM models, the wastewater is characterized in terms of many dissolved and particulate components (subjected to the type of ASM model) that are used to describe biomass groups, fractions of COD (organic matter slowly and readily biodegradable, soluble and non-biodegradable), nitrogen, phosphorus species and DO. The alkalinity is also included as part of the wastewater characteristics.

The objective of the ASM models is to transform the general mass balance in Eq. 2.1 to mathematical statements that are specific to the quantity of interest i.e. dissolved and particulate species. In ASM models, the consumption of substrates follow Monod kinetics and the stoichiometry is presented using the Petersen or Güjer matrix notation (Henze et al., 2000). Figure 2.3 illustrates the structure of the matrix and the interpretation of the mass balance equation for one particulate species, $X_1$. This interpretation extends to other particulate and dissolved species.

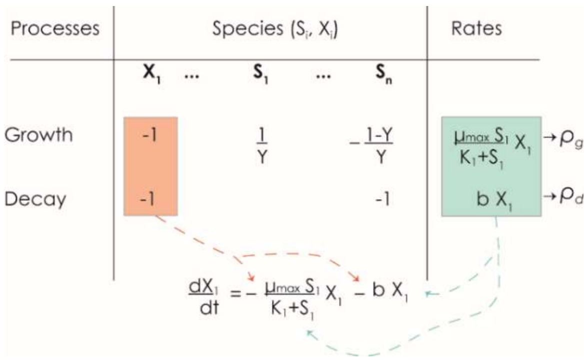


Figure 2. 3 Example of Petersen matrix notation and interpretation for two processes: growth and decay.

The processes studied in ASM1 and ASM3 are N-DN processes and aerobic oxidation of organic matter. In both ASM1 and ASM3 approaches, the nitrification considers a one step nitrification process instead of a separate nitritation and nitratation i.e. the production of both nitrite and nitrate (NOx-N). Table 2.1 summarizes the processes involved in ASM1 and ASM3 for both particulate and dissolved species.

Table 2. 1 Biochemical processes involved in ASM1 and ASM3 models for dissolved and particulated species.

| No. | Process | ASM1 | ASM3 |
|---|---|---|---|
| 1 | Hydrolysis of entrapped organics | X | X* |
| 2 | Hydrolysis of entrapped organic nitrogen | X | |
| Heterotrophic organisms, aer. and den. activity | | | |
| 3 | Aerobic storage of readily biodegradable substrate (SS) | | X |
| 4 | Anoxic storage of SS | | X |
| 5 | Aerobic growth | X | X |
| 6 | Anoxic growth (Denitrification) | X | X |
| 7 | Aerobic endogenous respiration | | X |
| 8 | Anoxic endogenous respiration | | X |
| 9 | Aerobic respiration of cell internal storage product of heterotrophic organisms | | X |
| 10 | Anoxic respiration of cell internal storage product of heterotrophic organisms ($X_{STO}$) | | X |
| 11 | Decay of heterotrophs | X | |
| Autotrophic organisms, nitrifying activity | | | |
| 12 | Aerobic growth of nitrifying organisms, Nitrification | X | X |
| 13 | Aerobic endogenous respiration | | X |
| 14 | Anoxic endogenous respiration | | X |
| 15 | Decay of autotrophs | X | |
| 16 | Ammonification of soluble organic nitrogen | X | |

*In ASM3 only one hydrolysis is acknowledged for the readily biodegradable organic matter.

On the other hand, ASM2 and ASM2d additionally cover the process of biological phosphorus removal from wastewater or EBPR. Table 2.2 summarizes the processes involved in ASM2 and ASM2d.

Table 2. 2 Biochemical processes involved in ASM2 and ASM2d models

| No. | Process | ASM2 | ASM2d |
|---|---|---|---|
| 1 | Aerobic hydrolysis | X | X |
| 2 | Anoxic hydrolysis | X | X |
| 3 | Anaerobic hydrolysis | X | X |
| Heterotrophic organisms: $X_H$ | | | |
| 4 | Growth on fermentable, readily biodegradable organic substrates (SF) | X | X |
| 5 | Growth on fermentation products (acetate) (SA) | X | X |
| 6 | Denitrification with SF | X | X |
| 7 | Denitrification with SA | X | X |
| 8 | Fermentation | X | X |
| 9 | Lysis | X | X |
| Phosphorus Accumulating organisms (PAO): $X_{PAO}$ | | | |
| 10 | Storage of poly-hydroxy-alkanoates ($X_{PHA}$) | X | X |
| 11 | Aerobic storage of polyphosphate (PP) | X | X |
| 12 | Anoxic storage of PP | | X |
| 13 | Anoxic growth | | X |
| 14 | Aerobic growth | X | X |
| 15 | Lysis of $X_{PAO}$ | X | X |
| 16 | Lysis of $X_{PP}$ | X | X |
| 17 | Lysis of $X_{PHA}$ | X | X |
| Nitrifying organisms (autotrophic organisms): $X_{AUT}$ | | | |
| 18 | Aerobic growth | X | X |
| 19 | Lysis | X | X |
| Simultaneous precipitation of phosphorus with ferric hydroxide, $Fe(OH)_3$ | | | |
| 20 | Precipitation | X | X |
| 21 | Redisolution | X | X |

The family of ASM models comprises a thorough work on the phenomenological understanding of biochemical processes occurring in wastewater treatment processes. They establish the fundamentals for further studies and improvements to these models. In this review, these further applications and derived models such as the ASM3+Bio-P developed by the Swiss Federal Institute of Aquatic Science and Technology (EAWAG) (Rieger et al., 2001) or the ASMN by Hiatt and Grady, (2008), which will be discussed in the following section.

### 2.2.1 Application and evolution of ASM models

Both ASM and data-driven approaches are very powerful tools for studying complex bWWTP. In addition, dynamic simulations of WWTP are also useful for selecting operational strategies to improve process stability, effluent quality and save operational costs (Ni et al., 2010).

This work provides a comprehensive survey on the different applications of both approaches. The applications and evolution of the deterministic ASM type models with respect to COD, N and P removal are discussed with special focus on the modifications made to the models over time, but also evolution of the research questions and interests from the wastewater community.

### 2.2.2 ASM1 and ASM3

In the early 1980's, Dold et al., (1980) and Batchelor, (1983) developed kinetic models to describe the most important microbial reactions in a single sludge treatment system for N-DN. Both studies involve Monod type kinetics and acknowledge endogenous respiration. Since these early approaches, the oxygen utilization rate was identified as the most sensitive parameter. Both Dold et al., (1980) and Batchelor, (1983) models were validated with experimental data. Moreover, Batchelor, (1983) applied the model to study the influence of the aeration fraction (AF), modifications of solids and hydraulic retention time and influent concentration of organic matter in the total nitrogen removal efficiency. In this first approach, the set of equations were modified to account for the effect of dissolved oxygen concentration on microbial growth. The main results suggested that low values of AF ($<0.5$) favor denitrification which results in low concentration of nitrate in the effluent.

Some years after, the International Association on Water Pollution Research and Control (IAWPRC) task group developed the Activated Sludge Model No. 1 (ASM1) Henze et al., (1987). Dold and Marais, (1986) studied the background of the ASM1 model proposed by the IAWPRC task group and compared it to their previously developed model.

The main differences between both models are discussed in Dold and Marais, (1986). Key differences comprise: i) the degradation of slowly biodegradable organic matter (SBOM) which according to Dold et al., (1980) is hydrolyzed, adsorbed and stored in the cells, instead in the ASM1, the SBOM is hydrolyzed directly to readily biodegradable organic matter (RBOM). Nonetheless, both models provide similar results. ii) the ASM1 does not attempt to account for the observed difference between the filtered total Kjeldahl nitrogen (TKN) and ammonia concentrations.

The ASM1 considers all biodegradable organic nitrogen to be particulate so that no soluble organic nitrogen is predicted and hence the model always under predicts the observed soluble TKN. In ASM1 a single decay process (lysis) was introduced to describe the sum of all decay processes under all environmental conditions (aerobic, anoxic). Two main reasons rely behind this assumption: i) when the ASM1 was released, endogenous respiration was assumed as the decay of cells and subsequent consumption of the decayed cells to form new biomass (Van Loosdrecht and Henze, 1999). The ASM1 used this death-regeneration concept in the activated sludge model because it could best fit the experimental observations. ii) Another issue was that at time (~1985), computing power was scarce and this simplification saved computation time. After the ASM1 was released, the decay processes were conceptualized into death and lysis, which can occur simultaneously or sequentially.

Since the ASM1 was built, several works have proposed applications and modifications. Up to date, around 300 articles cited the ASM1 by Henze et al., (1987) and on average 10 articles per year cite this study since 1988. Van Loosdrecht et al., (2015) developed a thorough study on the current advances of ASM1, perhaps more in depth than in this work. This study rather focusses to address those articles that were fundamental in the evolution of ASM1 and introduces applications from highly cited studies.

The initial concerns after the ASM1 was released were: the calibration, determination, and validation of the kinetic parameters in the model and the implementation of the ASM1 equations as a computer program. These topics extent up to today's discussion with more complex ASM type models. Just a year after the ASM1 was released, Bidstrup and Grady, (1988) developed a user friendly program in Turbo Pascal that aimed to model different configurations of activated sludge processes while using ASM1, the simulations were developed in steady state. Lesouef et al., (1992), Vanrolleghem et al., (1999) and Petersen et al., (2002), focused on experimental methods and techniques used to calibrate ASM1 kinetic parameters. In Lesouef et al., (1992), growth rate of autotrophic bacteria and the inert wastewater fractions were measured by simple methods, both in laboratory, pilot and full scale. Moreover, different configurations of plants were investigated including conventional recirculation, sludge re-aeration and step feed alternate zone denitrification. Based on several works in the literature, Vanrolleghem et al., (1999) summarized methods for determining kinetic and stoichiometric parameters in the ASM1, while Petersen et al., (2002) developed and evaluated a systematic model calibration procedure and a sensitivity analysis on the influence of model parameters and influent component concentrations on the model output.

Other works focused on different modifications of the ASM1; Oles and Wilderer, (1991) adapted the ASM1 for modeling a sequencing batch reactor (SBR), changes of COD, ammonia and $NO_X$ (sum of $NO_3^-$ and $NO_2^-$) during the SBR process cycle were predicted, for the determination of model parameters special batch experiments were carried out. Griffiths, (1994) proposed a modification for the ASM1 which allows predetermination of denitrification rates, division of the heterotrophs into four fractions, including poly-phosphates (poly-P), here presence of poly-P accumulating organisms is elucidated, similar observations were published elsewhere (Wentzel et al., 1992). Côté et al., (1995), presented an hybrid model that combined ASM1 and artificial neural networks (ANN) to forecast the effluent of an activated sludge process. The error between the experimental data and the ASM1 results were modeled by feed forward ANN. This study was the first to demonstrate the potentiality of ANN in combination with ASM1 to predict effluent composition in activated sludge processes. Argaman et al., (1999) on the other hand, applied ASM1 at steady state, BOD instead of COD was considered as carbonaceous organic matter following a zero order consumption rate, this study verified the consistency of ASM1 with experimental results and proposed a mathematical procedure to calibrate the set of kinetic coefficients in ASM1. Ekama and Wentzel, (2004) proposed a simple model for studying activated sludge processes where N-DN was conducted in excess of phosphorus, the contribution was the combination of ASM1 and ASM2 to model this process, however, the chemical precipitation process was not considered. Insel et al., (2011) studied the stability of N-DN processes in membrane bioreactors (MBR) focusing on oxygen diffusion. The main results from this work indicate that the optimal DO set-points increased with higher biomass concentrations due to higher mass transfer limitation, and they remained operative in a wider DO range. In this study, both ASM1 and ASM2 were applied. Le Moullec et al., (2011) compared different modeling approaches such as computational fluid dynamics (CFD) and compartmental modeling applying ASM1 kinetics. The results showed that the ammonium concentration was not predicted accurately with ASM1 under these conditions. Bürger et al., (2016) provided an extension to a well-known sedimentation model –Bürger-Diehl settler model (Bürger et al., 2011)- to include biological reactions from ASM1 and study denitrification during sedimentation in secondary settlers in WWTPs. Other applications of ASM1 involved the feasibility to use model predictive control (MPC) in N-DN processes (Holenda et al., 2008; Wu et al., 2014). Holenda et al., (2008), demonstrated the feasibility of applying the MPC to control the DO concentration in an aerobic reactor of a WWTP, the ASM1 was used to simulate the case studies and the MPC was adapted accordingly. On the other hand, Wu et al., (2014) studied the MPC controller to optimize the aeration and the external carbon source addition in N-DN systems where the carbon source was limited. Smets et al., (2003) proposed a strategy to reduce the complexity of the ASM1 by linearization. The ASM1 was linearized following a *time-varying* reference trajectory; the model obtained suggested the application of MPC for on-line control strategies. With the aim of improving the ASM1, the IAWPRC group developed the ASM3 in 1999 (Gujer et al., 1999). The most important modification to ASM1 deals with the introduction of endogenous respiration which replaces the lysis (decay) process. When the ASM3 was introduced, the computation power was not a limiting factor (as it was with ASM1), and thus, a more realistic description of decay processes was introduced. Endogenous respiration refers to the basic underlying oxygen consumption of the biomass. A significant fraction of the oxygen consumption occurring under endogenous conditions can therefore be related to consumption of internal substrate. There is a shift of emphasis from hydrolysis to storage of organic substrates. Taking substrate storage into account in the ASM3, affects the sludge yield (it will decrease and become similar to normal bacterial yields). It will further affect the decay or lysis rate (because a significant fraction of the "lysis" is turnover of stored substrate) and the hydrolysis process (for the same reason). Thus, the ASM3 approach, presents a more detailed mathematical representation of the activated sludge process in terms of the endogenous processes.

Several variations of ASM3 have been proposed since this model was developed. Iacopozzi et al., (2007) proposed an extension of the ASM3 by considering two step nitrification, i.e. nitritation by AOB, and subsequent oxidation of nitrite to nitrate by means of NOB. The new model was denoted as ASM3_2N. In ASM3_2N, 16 new kinetic rate expressions are included. The modification also accounts for the growing number of applications in advanced nitrogen removal, i.e. (partial) nitritation processes such as SHARON (Single reactor High activity Ammonia Removal over Nitrite) and anaerobic ammonium oxidation (Anammox) processes (Hellinga et al., 1998; Mulder et al., 1995), became relevant. The ASM3_2N has been applied by Hoang et al., (2012) for modeling partial nitritation and denitrification in a SBR reactor treating leachate. In that case, nitrite accumulation was noticed due to high pH and alkalinity causing inhibition of NOB. A year later, Zhou et al., (2013) modeled a granular SBR reactor comparing ASM3 and ASM3_2N. Improvements to the regular ASM3 for the prediction of total nitrogen were demonstrated. Novak and Horvat, (2012) applied the ASM3_2N and an oxygen electrode model to optimize the position of this device in a N-DN system. A sensibility analysis showed that the model was sensitive to: oxygen consumption per unit of $BOD_5$, specific consumption rate of adsorbed $BOD_5$, volumetric coefficient of oxygen transfers rate and wastewater inflow. The best place for positioning the oxygen electrode was the bioreactor instead of the outlet shaft; this resulted in shortening of the oxic/anoxic cycle by 13% and the daily working time of aerators. Other ASM3 modifications focused on describing simultaneous storage and growth processes occurring in activated sludge systems under aerobic conditions (Gao et al., 2017; Sin et al., 2005; Vázquez-Padín et al., 2010; Zhao et al., 2016). Sin et al., (2005) developed a mechanistic model based on ASM3, to describe simultaneous storage and growth processes occurring in activated sludge systems under aerobic conditions. A second order model was proposed for the description of the degradation of the storage products under famine conditions. Vázquez-Padín et al., (2010) modeled an aerobic granular SBR reactor using a one-dimensional biofilm model. In this model, simultaneous growth and storage of organic substrates by heterotrophic biomass and inclusion of nitrite as intermediate compound were considered. Moreover, different COD to nitrogen ratios were tested. The modeling results showed that the largest fraction of particulate compounds in the granules corresponded to the inert particulate organic material. Zhao et al., (2016) also studied high COD and nitrogen loads from piggery waste through N-DN considering simultaneous storage and growth in the heterotrophic nitrification and aerobic denitrification process. Gao et al., (2017) modeled a cyclic activated sludge system. The ASM3 was modified by combining the process of simultaneous storage and growth, and the kinetics of soluble microbial product and extracellular polymeric substances (EPS). The modified model showed improvements in comparison to ASM3. Xu et al., (2017) combined the ASM3 and the anaerobic digestion model 1 (ADM1) to successfully describe simultaneous C-N and S removal in activated sludge systems.

Along with the multiple variations of ASM1 and ASM3 models, the Anammox process was discovered (Mulder et al., 1995). The potentialities of modeling partial nitritation-Anammox (PN-A) systems arouse (Hulle et al., 2010), and mathematical modeling served as a useful tool for studying this new technology. To date, the dynamic nature of the PN-A process has been studied through simulations based on deterministic mathematical models. The most widely used models for PN-A in literature are the ASM3_2N and variations of the ASM1 to consider the Anammox process (Dapena Mora et al., 2004; Hao et al., 2005; Hao et al., 2002; Hellinga et al., 1998; Magrí et al., 2007). Extending the ASM3 or other ASM type models by Anammox has the advantage of studying intrinsic characteristics of the process, such as: oxygen consumption, inhibition, pH and temperature effects. ASM3 type models also help in the validation of hypothesis and scaling-up (based on bench scale reactors conditions). Below, I discuss some of the studies that extended ASM1/ASM3 to study PN-A processes. The main focus of these studies are; temperature, oxygen diffusion/aeration in biofilms, adaptation of PN-A systems to different carbon to nitrogen ratios and different reactor configurations.

Koch et al., (2000) applied an extended ASM3 with two step nitrification and the Anammox process. The autotrophic nitrogen removal in a rotating biological contactor (RBC) treating ammonium rich wastewater was studied through modeling and microbial analysis. The results showed that AOB and NOB are dominant in the upper aerobic 100-200 $\mu$m of the biofilm. The nitrite is assumed to diffuse into deeper anoxic layers of the biofilm, where it is reduced anaerobically in parallel with ammonium oxidation through the Anammox process. Hao et al., (2005) and Hao et al., (2002) adapted the ASM3 to model a one-stage PN-A process and studied the influence of oxygen consumption and temperature fluctuations on the process performance, respectively. Since the efficiency of PN-A processes depends greatly on good aeration strategies (limiting nitrification to an optimal nitrite: ammonia ratio), several studies have addressed this phenomenon through modeling (Corbalá-Robles et al., 2016; Mattei et al., 2015; Vangsgaard et al., 2012b; Volcke et al., 2005; Wyffels et al., 2004b). Wyffels et al., (2004) studied nitrite accumulation in a membrane bioreactor for the treatment of sludge reject water under continuous aeration at low DO concentrations (0.1 mg DO $l^{-1}$). Nitritation and nitratation in a MBR were modeled and the model was calibrated. The oxygen transfer coefficient and the ammonium loading rate were shown to be the appropriate operational variables to describe the experimental data accurately. Simulation results indicated that stable nitrite production from sludge reject water was feasible at relatively low temperatures of 20°C with an HRT of 0.25 days. Volcke et al., (2005) studied a SHARON-Anammox process through simulation in order to achieve optimal conditions with respect to the ammonium to nitrite ratio. Results showed that it was possible to control the nitrite: ammonium ratio by means of a cascade feedback control. The controller consisted of a master controller, which allowed to maintain the desired nitrite: ammonium set-point at 1.2 at an oxygen set-point between 0–8 g $m^{-3}$, for the slave controller that acted by adjusting the air flow rate, limited between 36-20,000 $m^3$ $h^{-1}$. Vangsgaard et al., (2012) and Volcke et al., (2012) developed models separately in order to study the consumption of substrates considering mass balances in PN-A systems; in the reactor (macro-scale) and in the biofilm (micro-scale), this means, considering the diffusional restrictions. These studies, however, were not validated with experimental data due to the complexity of measuring the substrates in the inner part of the biofilm. Nevertheless, the information within the results of the studies mentioned were important from an operational point of view. On the other hand, Volcke et al., (2012) developed a micro-scale model to study the effect of the granule size distribution on the performance of a granular sludge reactor in which autotrophic nitrogen removal was developed through a one-stage PN-A reactor. The results found by Volcke et al., (2012) show the importance of considering a granule size distribution in the mathematical modeling of these type of processes. Mattei et al., (2015) developed a mathematical model based on the multispecies modeling approach by Wanner and Gujer, (1986). The analysis and prediction of microbial interactions within multispecies biofilms including Anammox pathways were modeled. The results showed that the biofilm never experienced a fully penetrated oxygen profile and the Anammox bacteria could always survive in the inner part of the biofilm. However, prolonged exposure to a DO level of 5 mg $O_2$ $l^{-1}$ would lead to the loss of Anammox activity. Corbalá-Robles et al., (2016) evaluated the effect of aeration pattern in a granular SBR PN-A with flocculent biomass through modeling. The results showed that most of the ammonium oxidation potential would occur by means of the biomass in suspension rather than in granules. The aeration pattern had an important impact on the TN removal: a better performance was suggested for continuous aeration than for intermittent aeration. Other modeling approaches for biofilm PN-A systems were applied in the study of membrane aerated biofilm reactors (MABR) (Lackner et al., 2008; Terada et al., 2007). These studies focused on the analysis of the performance of MABRs in comparison to moving bed biofilm reactors (MBBR). The aeration patterns in the biofilm were compared, i.e. co-diffusion and counter diffusion. The model developed by Terada et al., (2007), considered only three bacterial metabolisms (AOB, Anammox, and NOB); growth and endogenous respiration were considered. The main results from this model demonstrated the high efficiency in TN removal of MABR in comparison to MBBR systems. Moreover, the model allowed to study the distribution of the biomass species and substrates along the biofilm. On a similar approach, the effect of heterotrophic biomass in both MBBR and MABR systems was studied by Lackner et al., (2008). The main results from this work suggest that in presence of organic matter (and as a consequence the presence of heterotrophic bacteria (HB)), the counter diffusion configuration (MABR) efficiency decays for COD:N>2. On the other co-diffusion (MBBR) systems were more stable.

Due to benefits that PN-A systems presented over N-DN processes for the treatment of wastewaters with high concentrations of ammonia and low ones for organic carbon, several works in the literature have explored to which extent PN-A systems could work efficiently and under different organic matter concentrations. Dapena Mora et al., (2004) applied the two step nitrification-denitrification ASM1 extended to an Anammox model for studying an SBR. The main goal of this study was to prove the feasibility of enrichment of Anammox bacteria from municipal sludge employing a SBR. The modeling results revealed that heterotrophs still remain in the system after the start-up of the reactor and can protect the Anammox bacteria from a negative effect of oxygen. Bi et al., (2015) developed a model for describing the nitrogen and organic carbon removal via simultaneous Anammox and heterotrophic denitrification process (SAD), and found that an organic carbon to nitrogen ratio of 1.5-2 was suitable for a batch SAD process. The most influential parameters were the half saturation constants for nitrite of heterotrophic bacteria and anammox bacteria and the anoxic reduction factors of HB and anammox bacteria.

In the past decade modeling the emissions of $N_2O$ in both N-DN and PN-A systems gained attention. In the study of N-DN modeling, the existing approaches focused on the enhancement of ASM1/ASM3 type models to better describe $N_2O$ emissions from denitrification processes. Some approaches consider four separate processes for denitrification and are therefore able to explain the $N_2O$ emissions from N-DN processes (Lu et al., 2018). The model proposed by Hiatt and Grady, (2008) was based to great extent on the ASM1. This model is known as the Activated Sludge Model Nitrogen or ASMN. The nitrification process incorporates two nitrifying populations; the AOB and NOB, however, free ammonia ($NH_3$) and $HNO_2$ are considered as the true substrates, respectively. ASMN takes into account four steps in the denitrification process using individual rate equations. The four processes describe anoxic heterotrophic growth using nitrate, nitrite, nitric oxide, and $N_2O$ respectively, as the terminal electron acceptor. The $\eta_g$ and $\eta_h$parameters from ASM1 are extended to four $\eta_i$ in the ASMN. Four separate $\eta_i$ parameters were used, one for each step in denitrification. These parameters took into consideration the fraction of heterotrophic bacteria accomplishing each step of denitrification and the reduced maximum specific growth rate under anoxic conditions. The ASMN was verified with experimental data and provided accurate results.

Other N-DN models developed by Pan et al., (2013) focused on the improvement of the denitrification modeling based on previous experimental studies, this model is known as the activated sludge model with indirect coupling of electrons or ASM-ICE. The model decouples the carbon oxidation and the nitrogen oxide reduction processes. For this purpose, electron carriers are introduced as new components in the model to link carbon oxidation to nitrogen oxides reduction, the so called *Mred* (reduced mediator) and *Mox* (oxidized mediator), defined as the reduced and oxidized forms of electron carriers, respectively. The main contribution of the ASM-ICE was the accurate prediction of the $N_2O$ accumulation during the denitrification process. This approach was later compared by the same authors in Pan et al., (2015) with the ASMN model for the $N_2O$ accumulation, being differentiated as the direct and indirect electron coupling approaches, respectively. The results showed the great accuracy of the ASM-ICE in the prediction of $N_2O$ accumulation in comparison to the ASMN. Moreover, within the four systems studied, the ASM-ICE was able to represent the experimental data well, while the ASMN only had accurate results in one system. Domingo-Félez and Smets, (2019) developed a simplified approach that aims to describe the electron denitrification rates as current flow in electric circuits. The model has fewer parameters than existing models (ASM-ICE and ASMN) and can be integrated in existing model structures.

In PN-A systems three production pathways of $N_2O$ are known; heterotrophic denitrification and two processes mediated by AOB, that is, $NH_4^+$ oxidation via $NH_2OH$ and autotrophic denitrification (NOB → nitrite and nitrate reduction enzymes) (Lu et al., 2018). Some modeling studies have addressed this issue and the conditions where $N_2O$ is produced. Ni et al., (2013) employed a multispecies one-dimensional biofilm model considering nitric oxide (NO) and nitrous oxide ($N_2O$) productions in PN-A membrane aerated biofilm reactors (MABR), it was found that intermittent aeration can reduce NO and $N_2O$ emissions in MABR systems and simulations showed that over 3.5% of the removed total nitrogen could be featured to NO and $N_2O$ production under operational conditions optimal for total nitrogen removal. Lu et al., (2018) developed a mechanistic model based on ASM1 to describe $N_2O$ production in a granular lab-scale PN-A SBR reactor. The three pathways mentioned previously were considered and kinetic parameters in the extended model were validated with experimental data obtained from batch experiments. The results showed that heterotrophic denitrification became a greater contributor to $N_2O$ emission compared to the oxidation of ammonium via $NH_2OH$ and autotrophic denitrification pathways. On the other hand, Wan et al., (2019) found through modeling studies that the main production of $N_2O$ in a granular single stage PN-A reactor was due to nitrifier denitrification in the outer layer of the granules (pathway of nitrification in which ammonia is oxidized to nitrite followed by the reduction of nitrite to nitric oxide, $N_2O$ and molecular nitrogen ($N_2$) (Wrage et al., 2001)).

Several mathematical models have been developed to understand the phenomenological processes occurring in biological nitrogen removal processes. The research has evolved from adapting the traditional ASM1 and ASM3 to add two step nitrification and four step denitrification, considering the Anammox process, and to study $N_2O$ emissions in these systems and the many applications to different reactor configurations.

### 2.2.3 ASM2 and ASM2d

Modeling approaches previous to the ASM2, that aimed at explaining the kinetics within the bio-P-removal process were first published in the early 1990s (Ante et al., 1994; Smolders et al., 1995, 1994; Wentzel et al., 1992). These first approaches followed the ASM1 fashion for the representation of the systems. Smolders et al., (1994) developed a metabolic model to describe the stoichiometry and kinetics of the EBPR process. In this early approach all relevant metabolic reactions underlying the metabolism, considering also components like adenosine triphosphate (ATP) and nicotinamide- adenine dinucleotide (NADH,) are described based on biochemical pathways. On the other hand, the industrial standard for modeling EBPR is the ASM2 (Gujer et al., 1995; Henze et al., 2000). The ASM2 comprises many biochemical reactions occurring simultaneously such as: organic oxidation, nitrification, denitrification and phosphate release and uptake (Table 2.2) (Kim et al., 2001). The focus of the ASM2 is the representation combined biological processes for COD, Nitrogen and Phosphorus removal and also considers chemical precipitation. The main difference between the ASM1 and ASM2 is that the biomass has internal structure and therefore its concentration cannot simply be described with the distributed parameter of concentration of biomass (the phosphorus participates in the metabolism of the microorganisms), as well, the chemical precipitation process and the total suspended solids (TSS) are introduced in the kinetics of the model.

ASM2 was first adapted by Isaacs et al., (1995), who introduced denitrification by phosphorus accumulating organisms (PAOs). Additionally, Mino et al., (1995) proposed the kinetics of processes associated to glycogen accumulating organisms (GAOs), introducing new parameters such as $X_{Gly}$ which denotes the intracellular stored glycogen and serves as a carbon source for poly-hydroxy-alkanoates (PHA) storage processes. Kuba et al., (1996) and Murnleitner et al., (1997) made progress in establishing metabolic modeling of phosphorus removal in biological wastewater treatment processes. Furumai et al., (1999) applied the ASM2 accounting denitrification by PAOs in a long-term operation of an SBR. In the same year, the ASM2d was developed, this model aims to correct some pitfalls in ASM2 and considered the denitrification by PAOs (Henze et al., 1999). Manga et al., (2001) acknowledged the competition between PAOs and GAOs and suggested kinetic expressions for this competition process for ASM2. A couple of years later, the EAWAG Bio-P or ASM3 Bio-P was developed by Rieger et al., (2001) and Siegrist et al., (2002). The ASM3 Bio-P acknowledged some processes from ASM2d, however it was based on the ASM3. The ASM3 Bio-P neglected the fermentation of readily degradable substrate and accounted for biomass decay as an endogenous respiration process. Four additional state variables were added to the ASM3, and the kinetics of GAOs were not considered. However, the competition between GAOs and PAOs was still subject of further studies, Zeng et al., (2003) developed a model that described the competition of GAOs and PAOs which was derived from an experimental study with mixed cultures of PAO and GAO under anaerobic conditions. The activity of both was studied and kinetic and stoichiometric parameters were identified. Yagci et al., (2004) also modelled the competition between PAOs and GAOs for acetate uptake in an SBR. The stoichiometry and kinetics related to GAOs were presented in detail, and the model is calibrated. The structure of the extended model was similar to ASM2d. The same authors, evaluated the performance of ASM2d with varying phosphate concentrations in the influent of a bench scale SBR (Yagci et al., 2006). The main results suggested that competition between GAOs and PAOs should be added to the ASM2d.

The EBPR in membrane bioreactors was also studied through modeling. Jiang et al., (2008) proposed the ASM2d-SMP (soluble microbial products; SMP) process accounting for two new state variables $S_{UAP}$ and $S_{BAP}$, which are the biomass associated products (BAP) and utilization associated products (UAP).

Later, Cosenza et al., (2013) would integrate the ASM2d-SMP and the model developed by Mannina et al., (2010) to account for the variation in membrane filtration characteristics and the influence on the COD removal in the process. Zuthi et al., (2013) developed a rigorous study regarding the core issues of biological phosphorus removal in activated sludge processes and membrane reactors, which as well provides both modeling approaches for EBPR; metabolic modelling and ASM2 and ASM2d models.

Table 2.3 summarizes more applications of ASM2 and ASM2d, and the conditions under which these studies were performed.

Table 2. 3 Different mathematical modelling approaches for P removal and operational conditions.

| **Approach** | **P [g P m$^{-3}$]** | **COD [g COD m$^{-3}$]** | **T [°C]** | **pH [-]** | **Type of system** | **Ref.** |
|---|---|---|---|---|---|---|
| Based on the ASM1 | 8.8 | 61 | 25 | 7.0 | Activated sludge system: 4 combined bioreactors | (Ante et al., 1994) |
| Anaerobic metabolism of P removal | 15 | 400 | 20 | 7 | SBR reactor: only anaerobic phase was studied | (Smolders et al., 1994) |
| Aer. and Anaer. P removal: ATP prod. involved in the stoichiometry and kinetics | 15 | 400 | 20 | 7 | SBR reactor: both anaerobic and aerobic phases were studied | (Smolders et al., 1995) |
| Based on ASM2: accounting den. by PAO | 4 | 20 | - | - | Pilot scale activated sludge plant | (Isaacs et al., 1995) |
| TU Delft P model ATP/NAHD ratio based on Smolders et al., (1995) | 15 | 400 | 20-25 | ~7 | Bench scale SBR | (Kuba et al., 1996; Murnleitner et al., 1997) |
| Influence of COD concentration and biofilm thickness is studied: Based on ASM2 | 17 | 400-600 g Acetate m$^{-3}$ | - | - | Sequencing Biofilm Bioreactor (SBBR) | (Morgenroth and Wilderer, 1998) |
| Based on ASM1 model. | 14 | 664 | 14.5 | ~7 | Pilot plant | (Nolasco et al., 1998) |
| Modelling of Phosphate removal by precipitation | - | - | - | - | Batch experiments | (Fytianos et al., 1998) |
| Simplified ASM2 and Neural Nets | 6.9 | 296 | 20 | ~7 | SBR reactor | (Zhao et al., 1999) |
| ASM2d: ASM2 + den. by PAO | 6 | 260 | 10-20 | ~7 | Activated sludge | (Henze et al., 1999) |
| ASM2d and ASM3: without fermentation of ready degradable substrate: ASM3+BioP | 4.14-10.01 | 250-380 | 20 | ~7 | Pilot plant | (Siegrist et al., 2002) |
| Based on ASM3: Account of Mg limitation | 63 | 2430 | 27 | 6.8 | Pilot Scale SBR | (Ky et al., 2001) |
| Based on ASM2: linear version of ASM2 | 6.4 | 366 | 25 | ~7 | Bench-scale SBR | (Kim et al., 2001) |
| Based on ASM2 and TU Delft P model: model for denitrifying dephosphatation | 15 | 400 | 20 | ~7 | Bench scale results from Kuba et al., (1996) | (Hao et al., 2001) |
| Based on ASM2 model: Effect of COD on the PAO | 30 | 300 | 23 | ~7 | Bench scale SBR | (Soejima et al., 2008) |
| Metabolic model based on TU Delft P model: competition between PAO and glycogen accumulating organisms (GAO) | 2$^{a}$ | 400$^{a}$ | 18-22 | ~7 | Bench scale SBR | (Oehmen et al., 2010) |
| Based on ASM2: MBR modeling and soluble product approach, based on (Jiang et al., 2008) | 1.5 | 327 | 21 | 7.6 | Pilot Plant MBR | (Cosenza et al., 2013) |
| Based on (Murnleitner et al., 1997; Smolders et al., 1994) models. | 10.5 | 110 | 20 | 7-8.9 | Bench scale SBR | (Acevedo et al., 2014) |

*Continue…*

| | | | | | | |
|---|---|---|---|---|---|---|
| Model for $N_2O$ Production during denitrifying P removal | 7.2[b] | 240[b] | 20-22 | ~7 | Bench scale SBR | (Liu et al., 2015) |
| Extended ASM2: Role of EPS in Biological P removal | 6-20 | 150-400 | 20 | 7.5 | Bench scale SBR | (Yang et al., 2017) |

[a] based on the work of Zeng et al., (2003)

[b] based on Wang et al., (2011)

(*): Processes carried out by PAO: two anaerobic processes (acetate uptake and anaerobic maintenance) and four aerobic processes (PHA degradation, poly-P formation, glycogen accumulation and aerobic maintenance))

Some of the operational conditions in Table 2.3, are far from the operational conditions and composition recommended initially by Henze et al., (1999). The existing models have significant differences among the assumptions and kinetics involved. Therefore, the reactions of the metabolism can only represent the reaction stoichiometry based on the assumed biochemical pathways. As consequence, the parameter calibration for each model will result in very different kinetic values. Recent works are focused on modelling $N_2O$ production in EBPR (Liu et al., 2015; Wisniewski et al., 2018). The production of $N_2O$ in EBPR is mainly featured to accumulation of nitrite and as a consequence, a decreased consumption of phosphate.

In this work, the evolution of ASM models is understood as the improvements and modifications made to the initial ASM models developed by the IWA.

From the applications above, the literature survey shows that new biochemical pathways were discovered which came along with the technology advances for a more detailed study of the bWWTP occurring within. Figure 2.4 summarizes the evolution of ASM models for modeling C, N and P removal from wastewater. Some of the most common commercial and free softwares for modeling ASM type models are summarized in the section A.1 in the Appendix.

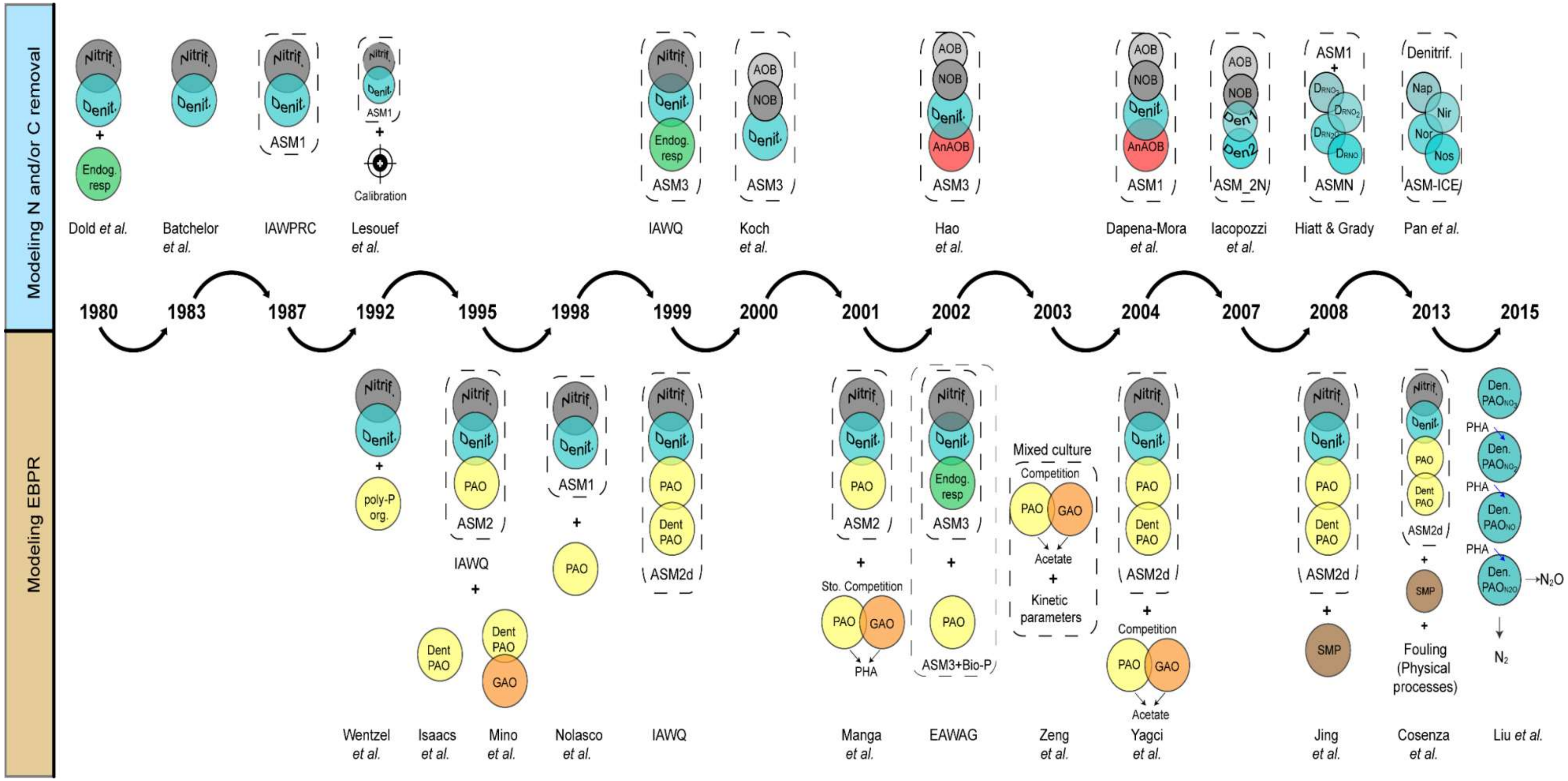


Figure 2. 4 Evolution of Activated sludge models for C, N and P removal in wastewater treatment

### 2.2.4 Challenges

Over the years, several modifications to the initial ASM1, ASM2 and ASM3 were proposed with the aim to overcome limitations identified in these models. From a biochemical perspective, these upgrades range from one to two step N-DN, the integration of $N_2O$ kinetics and Anammox in ASM1 and ASM3, to account for the kinetics of GAOs and denitrifying PAOs in EBPR, e.g. in ASM2 type models. However, some limitations are still relevant when ASM-type models are applied.

The initial ASM-type models were developed to model a conventional activated sludge process. New technologies such as MBR and MBBR, suggested the adaptation of the ASM models to these new technologies, adding complexity to the modeling process. The literature has vastly addressed that there exist large differences in the type the biomass and biokinetic processes in these different systems; due to the stratification of the biomass. In MBR systems, not only the biokinetics are different, but the mechanical processes involved; the filtration process, backwashing and fouling (Fenu et al., 2010; Naessens et al., 2012a, 2012b). For MBR (as for other configurations), the characterization of the biomass is crucial for building a realistic model when based on ASM type models (Cosenza et al., 2013; Jiang et al., 2009). It has been demonstrated that the active heterotrophic biomass in the influent wastewater that, although usually neglected in the conventional activated sludge modeling (initial ASM models), needs to be better addressed when modeling MBRs. In fact, from a theoretical point of view, longer sludge ages lead to decreases in the percentage of active biomass and thus, the higher the SRT, the less negligible the new biomass entering the plant via the influent becomes. Other studies have also addressed the necessity to add solubilization of inorganic solids entering the system when modeling MBR with high SRT (Spérandio and Espinosa, 2008). Further studies have addressed the differences of growth rate for MBR nitrifiers compared to those in conventional activated sludge systems, experiments showed the inhibition of soluble microbial products on the nitrifiers in MBR systems (Jiang et al., 2009). On the other hand, in biofilm systems, a key drawback of the implementation of ASM type models is the oversimplification of the mass transfer. The literature has demonstrated that mass transfer mechanisms are different to those proposed in the initial ASM type models. In specific, the oxygen diffusion is a process that is usually oversimplified in the ASM models, leading to a misinterpretation of “optimal” DO oxygen concentrations. Horn and Hempel, (1997) studied the mass transfer in an autotrophic biofilm, the results from this study showed that the kinetics such as decay, cannot be considered the same along the biofilm and the distribution of the liquid phase in the biofilm is not uniform. Other studies that have addressed diffusional effects are Vangsgaard et al., (2012b) and Picioreanu et al., (2016). Picioreanu et al., (2016) studied the effect of nitrifying biomass distribution on the oxygen affinity coefficient of NOB and AOB. This process was studied in a 3D modelling approach which demonstrated the need of accounting for the distribution of biomass along the geometry of the biofilm. Nopens et al., (2015) suggest the need of the application of population balance models for integrating the effect of granule/floc size distribution and the consumption of nutrients in activated sludge processes.

Along the evolution of the biochemical processes in ASM type models, a clear need for a new modeling framework for each technology has emerged (Rieger et al., 2012). However, still today, several works in the literature still ignore this fact (Hauduc et al., 2009). A general problem of ASM type models is their difficulty to adjust a significant set of kinetic and stoichiometric parameters for a respective system. The settings of the kinetic parameters should be characteristic to each biochemical process and wastewaters (industrial or municipal). However, to achieve an accurate characteristic model, parallel experiments are necessary to determine kinetic and stoichiometric parameters, which is time consuming and expensive. Hauduc et al., (2013) developed a detailed study on the limitations of ASM type models from a biochemical perspective, where more careful selection of the set of kinetic parameters and selection of ASM modifications are recommended.

## 2.3 Data-driven approaches in wastewater treatment

Data-driven methods in the field of bWWTP have focused on three main applications:

i) forecast the effluent/influent composition of a process through regression models (aka. supervised ML)

ii) process understanding and knowledge discovery i.e. looking deeper into the available data to discover new information and valuable correlations within the data (unsupervised ML).

iii) application of intelligent control systems in wastewater treatment facilities such as adaptive neuro fuzzy interference systems (ANFIS).

The complexity of biological process involved in wastewater treatment together with the vast amount of data generated while monitoring the process performance present an opportunity for the application of data-driven methods in bWWTP. A main advantage of data-driven methods over mechanistic models based on the ASM approach is the high adaptability and accuracy when applied for prediction tasks of highly complex processes. Still, the application of data-driven methods is somehow a *mystery* in the water sector. The perception that the community has on methods based on ML, and in general the application of these methods in wastewater treatment, shows that they are almost *magical* methods that give very good results. Current studies that use these methods do not provide a detailed explanation of the methodology behind obtaining these models and their limitations. Thus, data-driven approaches are commonly referred to *black-box* (Hutson, 2018).

Several works indistinctively address *artificial intelligence* or AI, ML and data mining, although they are different concepts and it is important to clarify the relation within these fields. Thus, these concepts are briefly reviewed before the revision of the literature on the applications of data-driven methods in wastewater treatment.

For humans, intelligence is one of our most important features to which the advances in several fields of science and technology we have today can be featured. The field of AI attempts to understand and build intelligent entities. This concept can be understood from two different approaches; reasoning (thinking or acting rationally) and behavior (acting or thinking humanly). For AI to succeed, two elements are necessary: intelligence and an artifact. The computer has been the artifact of choice in this field (Russell et al., 2010). AI has numerous subfields that cover the approaches mentioned previously, an example: robotics is a sub-field of AI which focusses on building machines that behave *humanly* or perform a human task better than us. When we refer to *machine learning* or ML, this field comprises methods built so when programmed in a computer, they are said to *learn from experience* with respect to some class of task and the quality of the learning performance is evaluated by a performance measure. The computer will be said to learn from experience, only if the performance measure improves (Mitchell, 1997). ML draws ideas from different disciplines; AI, probability and statistics. ML is especially useful in *data mining* problems, a field which deals with large databases which contain relations that can be discovered automatically with ML. A simple example applied to bWWTP is to develop a computer program to learn general rules from data to identify rain events in WWTP. The three fields are related to each other, but not the same.

ML provides methods for modeling and extracting knowledge from data. The terms *supervised* and *unsupervised* refer to two branches of ML. In supervised ML, a model is provided with examples of data for training. This past experience will be used to fit the model that relates the predictors (input parameters) to the response (output parameters) (James et al., 2013). Afterwards this model can be applied for further validation with new unseen data. The opposite occurs in unsupervised ML. There, no data is provided to a model to *learn*, since the outcome is unknown in unsupervised ML, these methods are applied for extraction and discovery of new and relevant knowledge from a dataset. Contrary to supervised ML, the problem is not guided by a key variable to be modeled or classified into a category, but rather to find patterns. However, before the application of any supervised or unsupervised method to the data, several previous steps should be followed. These steps are briefly summarized in Figures 2.5 and 2.6, and will be thoroughly discussed and explained in Chapter 3. The main motivation behind the brief introduction and clarification of some concepts so far is to be able to identify the gaps in the methodology in the further studies reviewed in the following section.

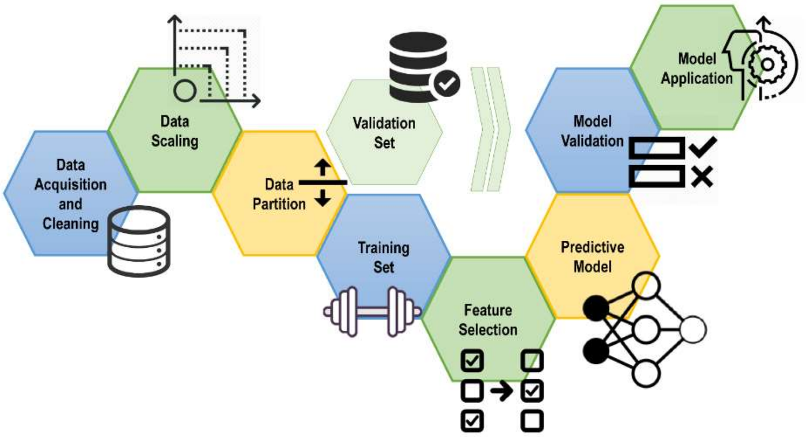


Figure 2. 5 Workflow for building supervised machine/statistical learning models.

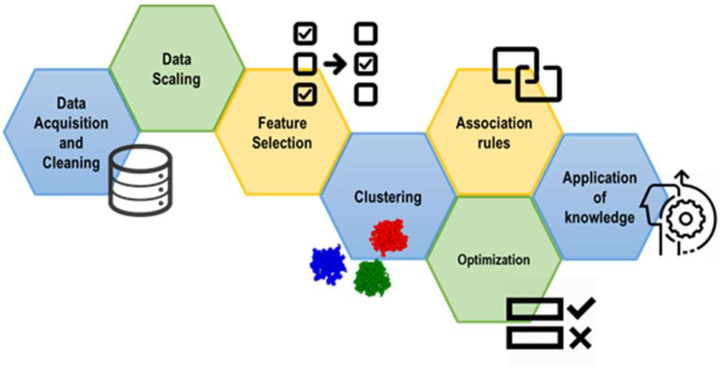


Figure 2. 6 Workflow for the exploration of knowledge with unsupervised ML methods.

### 2.3.1 Data-driven methods in wastewater treatment

In this section, the applications of data-driven methods based on ML in wastewater treatment processes are outlined. Table 2.5 summarizes different studies that applied ML methods in different wastewater treatment related processes.

Table 2. 4 Studies that applied machine learning methods in prediction tasks in wastewater treatment processes and the methods used.

| Applied ML studies to WWT | Feature Selection (FS) | ANN | ANN+Other | Other |
|---|---|---|---|---|
| Côté et al., (1995) | | | ASM1 and ANN to improve the prediction of the effluent | |
| Lee and Park, (1999) | | ANN to predict $PO_4$-P, $NH_4$-N and $NO_3$-N in a WWTP | | |
| Zhao et al., (1999) | | | ASM2 improve the prediction of the effluent | |
| Lee et al., (2002) | | | Prediction of error with ANN from ASM1: species predicted COD, MLSS, Cyanide and SS | |
| El-Din and Smith, (2002) | | Influent of a WWTP is predicted including rain events | | |
| Hamed et al., (2004) | | BOD and SS in the effluent of a WWTP are predicted | | |
| Mjalli et al., (2007) | | BOD, COD and TSS are predicted in the effluent of a WWTP | | |
| Hong et al., (2007) | | Forecast of $PO_4$-P, $NH_4$-N and $NO_3$-N in a SBR bench scale reactor. | | |
| Pai et al., (2007) | | SS and COD were predicted from a WWTP treating hospital wastewaters. | | |
| Ráduly et al., (2007) | | COD, $BOD_5$,TSS and $NH_4$-N were predicted in a WWTP | | |
| Dixon et al., (2007) | | VFA are predicted with ANN in an Anaerobic digestion process | | |
| Akratos et al., (2008) | PCA | BOD removal in a constructed wetland | | |
| Aguado et al., (2009) | | Online P concentration prediction in a SBR and further soft sensor application | | |
| Oehler et al., (2010) | BRT | | BRT and ANN are applied for modeling denitrification process effluent: $NO_3$ $NO_2$ and $N_2O$ | |
| Elmolla et al., (2010) | | COD removal in a Fenton process was predicted | | |
| Güçlü and Dursun, (2010) | | MLSS, SS and COD are predicted in the effluent of a WWTP | | |
| Kashani and Shahhosseini, (2010) | MPCA | COD and VSS concentration in the effluent of 14 SBR reactors was predicted | | |
| Dürrenmatt and Gujer, (2012) | PCA | | GLSR, ANN, RF, and SOMs were applied for modelling two full scale bWWTP and building of soft-sensors based on the information of the models. | |
| Abbasi et al., (2012) | PLS | | | Municipal solid waste is predicted with SVM |

| | | | | |
|---|---|---|---|---|
| Boniecki et al., (2012) | | Prediction of $NH_4$ generated in composting sewage sludge. | | |
| Kusiak et al., (2013b) | | | A pumping system in a WWTP is predicted through MLP, CART, MARS, SVM, and RF. Two parameters are input variables: Energy consumption and flow rates after the pumping system | |
| Kusiak et al., (2013a) | | | Carbonaceous BOD in the influent of a WWTP is predicted through different data mining methods: MLP, CART, MARS and RF | |
| Verma et al., (2013) | | | TSS in the influent of a WWTP is predicted through MLP, MARS, SVM , KNN and RF | |
| Han and Qiao, (2013) | | SVI is predicted in a WWTP through EELM-HRBF-ANN | | |
| Kusiak and Wei, (2014) | BT | | Methane production in a WWTP is predicted through ANN, ANFIS and SVM | |
| Vega De Lille et al., (2015) | | | $NH_4$ prediction based on pH online data and ASM1 .in an Anammox SBR | |
| Guo et al., (2015) | | | TN prediction from a WWTP with ANN and SVM | |
| Bagheri et al., (2015) | | SVI is predicted to study sludge bulking in WWTP | | |
| Zhang et al., (2016) | | System pump modeling with ANN | | |
| Granata et al., (2017) | | | | BOD, COD, TSS and TDS in the effluent of a WWTP are predicted with SVM and CART |
| Xie et al., (2017) | PCA | ANN to predict $NH_4$ from an Anammox reactor | | |
| Asadi et al., (2017) | BRT | | KNN, ANN, MARS and RF were applied for the prediction of BOD, TSS and DO in a WWTP for optimization of aeration. | |
| Alejo et al., (2018) | RF | | SVM and ANN were applied for the prediction of the effluent of a two stage AD process | |
| Torregrossa et al., (2018) | RF | | ANN and RF were applied for evaluating parameters in a WWTP that would have the highest influence in the energy cost. | |
| Huang et al., (2009) | | | | Partial least squares-SVM to predict the effluent of a WWTP: COD, BOD, TN, $NH_4$, |

ML: machine learning; WWT: wastewater treatment; ANN: artificial neural networks; SOM: Self Organizing maps; RF: Random Forest; PCA: Principal component analysis; MPCA: Multiway PCA; BRT: Boosted Regression Trees; ANN: Artificial Neural Networks; SVM: Support Vector Machines; LS-SVM: Least Squared SVM;GLSR; Generalized linear squared regression; MARS: multivariate adaptive regression spline; MLP: Multilayer perceptron; EELM-HRBF: extended extreme learning machine- hierarchical radial basis function.

Most of the studies applied ANN for predictive tasks, the most common software or platform for developing these models is the Matlab Neural Network Toolbox™. Most methods applied belong to supervised ML, except for self-organizing maps and principal component analysis, which are unsupervised learning methods. When compared to the methodology proposed in Figures 2.5 and 2.6, few studies developed feature selection, which as it will be discussed in further chapters in this work, is key to reduce the complexity of data-driven models and obtaining accurate results. Few of the studies discussed focused on explaining pre-processing of the data and the quality of the data for analysis, which is key to select an appropriate method and to evaluate the consistency of the results.
Table 2.5 summarizes some applications of ANN in wastewater treatment systems, focusing on the prediction goal and the software used for analysis.

Table 2. 5 Use of artificial neural networks in modeling different wastewater treatment processes.

| Purpose of study | System | Prediction with ANN | Software | Reference |
|---|---|---|---|---|
| Optimization of effluent prediction by ASM1 model | Activated Sludge | Error of VSS, SS | Not mentioned | (Côté et al., 1995) |
| Effluent prediction with the on-line information. | SBR bench scale reactor, activated sludge system. | $PO_4^{3-}$, $NO_3^-$, and $NH_4^+$ in the effluent. | Matlab | (Lee and Park, 1999) |
| Prediction of the effluent of an SBR P removal system through a simplified ASM2 model and ANN | SBR P removal system | Effluent COD, $PO_4^{3-}$ and $NO_3^-$ | Not mentioned | (Zhao et al., 1999) |
| ASM1 model was combined with ANN to predict the effluent. | Full-scale coke plant wastewater treatment by a activated sludge unit. | ASM1 error of prediction of pH, COD, Qin, and cyanide concentration. | Matlab | (Lee et al., 2002) |
| Predictions of inflow rate in a WWTP | WWTP | Influent flow rate | NeuroShell 2 | (El-Din and Smith, 2002) |
| The effluent of a WWTP was predicted. | WWTP Egypt | BOD and Suspended Solids in the effluent | Not mentioned | (Hamed et al., 2004) |
| Effluent prediction of a WWTP | WWTP | TSS, COD and BOD in the effluent | Matlab | (Mjalli et al., 2007) |
| Real-time estimation of $PO_4^{3-}$, $NO_3^-$ and $NH_4^+$ concentrations in a SBR with online information. | SBR bench scale reactor, P and N removal. | $PO_4^{3-}$, $NO_3^-$ and $NH_4^+$ in the effluent. | Matlab | (Hong et al., 2007) |
| Effluent composition prediction. | SBR in a hospital WWTP | COD and Suspended solids in the effluent. | Matlab | (Pai et al., 2007) |
| ASM3 and ANN were applied for the prediction of a WWTP | WWTP | Ammonium, $BOD_5$, TSS, TKN, and COD | Matlab | (Ráduly et al., 2007) |
| Prediction of the effluent of an anaerobic digester | Anaerobic digestion reactor | VFA concentration in the digester | Clementine | (Dixon et al., 2007) |
| Constructed wetlands modeling. | Constructed wetlands | BOD | StatSoft Statica version 7 | (Akratos et al., 2008) |
| Prediction of P in the effluent for building a soft-sensor in a SBR reactor | SBR bench scale for EPBR | P in the effluent | Not mentioned | (Aguado et al., 2009) |
| Denitrification in soil was modeled. | Soil | TN emissions from soil | R | (Oehler et al., 2010) |
| Antibiotic degradation in aqueous solution by the FENTON process modeled | FENTON process | COD removal | Matlab | (Elmolla et al., 2010) |
| Effluent prediction of a WWTP | WWTP | COD, SS, MLSS | Matlab | (Güçlü and Dursun, 2010) |
| SBR reactor effluent was modeled | Bench scale SBR reactor | COD and VSS | Matlab | (Kashani and Shahhosseini, 2010) |
| Modeling ammonia emissions in composting sewage sludge | Composting sewage sludge process | Ammonia | Statistica v.7.1 | (Boniecki et al., 2012) |
| Prediction and optimization of methane production. | WWTP | Methane production | Not mentioned | (Kusiak and Wei, 2012) |
| Carbonaceous biochemical oxygen demand (CBOD) in the influent is predicted since is a parameter not frequently measured. | WWTP | CBOD | GESCONDA | (Kusiak et al., 2013a) |

*Continue…*

| | | | | |
|---|---|---|---|---|
| A pumping system in a WWTP was modeled following | WWTP | Energy consumption and wastewater flow | Not mentioned | (Kusiak et al., 2013b) |
| TSS in the effluent of a WWTP is predicted due to infrequent measurements | WWTP | TSS in the effluent | Not mentioned | (Verma et al., 2013) |
| In order to quantify sludge bulking, Sludge Volume Index (SVI) and BOD of wastewater treatment process is predicted | WWTP | SVI and BOD | Not mentioned | (Han and Qiao, 2013) |
| Methane production prediction in a WWTP | WWTP, anaerobic digester | Methane production | Matlab | (Kusiak and Wei, 2014) |
| Ammonium concentration in a SBR reactor using pH online measurements was performed. | Anammox reactor | Ammonium in the effluent | Matlab | (Vega De Lille et al., 2015) |
| TN in the effluent of a WWTP was predicted | WWTP | Total Nitrogen effluent | Matlab | (Guo et al., 2015) |
| Prediction of SVI through ANN and GA is done. GA was used in order to optimize the weights and thresholds of the ANN | WWTP | SVI | Matlab | (Bagheri et al., 2015) |
| Least squares support vector machines and ANN are applied to predict different types of carbon dioxide emissions and from which industry come from. | Different industries emissions | $CO_2$ emissions | Matlab | (Sun and Liu, 2016) |
| Improving the performance of wastewater pumping systems | Wastewater pumping systems | Pumped wastewater flowrate and energy consumption | Not mentioned | (Zhang et al., 2016) |
| Prediction of ammonia in the effluent of an Anammox reactor | Bench scale Anammox reactor | Ammonia in the effluent | Not mentioned | (Xie et al., 2017) |
| Optimization of the aeration system (blowers) in a large scale WWTP | Aeration system in a WWTP | Blower energy consumption | Not mentioned | (Asadi et al., 2017) |
| Prediction of a two stage anaerobic digestion process | Bench scale anaerobic digestion system | Ammonia production in the effluent | WEKA and R | (Alejo et al., 2018) |
| Generate high-performing energy cost models for WWTP, using a database of 317 WWTP located in north-west Europe. | WWTP information of consumption | Energy cost | Not mentioned | (Torregrossa et al., 2018) |

The highly accurate results obtained with ANN in comparison to other statistical methods, made ANN an attractive tool for predicting tasks in wastewater treatment. Asadi et al., (2017), Kusiak et al., (2013a, 2013b), Kusiak and Wei, (2014, 2012), Verma et al., (2013) and Zhang et al., (2016) applied different supervised ML methods for aeration optimization, carbonaceous biochemical oxygen demand in the influent, pumps energy consumption and water inflow, methane production in both studies total suspended solids and pump energy consumption and flow rate, respectively. In all these studies, different supervised statistical learning and ML methods were tested where in most of them, ANN delivered the highest performance, achieving squared correlation coefficients higher than 0.9 ($R^2$>0.9). These studies did not clearly justify the application of these methods. Was it based on the nature of data-generation process? Was it based on the (close to) normal distribution of its parameters? Since the calibration of the parameters and selection of the training datasets were not properly described, it is not possible to know whether the ANN was overfitting. No clear description on which *tuning* (adjustment of hyperparameters in the ANN model) methods were applied when obtaining the models.

Interesting results can be found in the literature on the application of other supervised ML models such as: random forest, support vector machines, boosted tree regression, generalized least squares regression, least squares support vector machines, multivariate adaptive regression spline, among others (Abbasi et al., 2012; Alejo et al., 2018; Dürrenmatt and Gujer, 2012; Granata et al., 2017; Guo et al., 2015; Sun and Liu, 2016). Abbasi et al., (2012), Guo et al., (2015) and Alejo et al., (2018) applied support vector machines for the accurate ($R^2$>0.9) prediction of municipal waste generation, ammonia concentration in the effluent of an anaerobic digestion system and the total nitrogen concentration in the effluent of a WWTP, respectively. Dürrenmatt and Gujer, (2012) applied generalized least squares regression, self-organizing maps ANN and random forest to predict COD and ammonia in a WWTP and gained knowledge from the data analysis in the system. Granata et al., (2017) applied regression trees and support vector machines for predicting TSS, COD and $BOD_5$ in storm water. Support vector regression showed a better performance than regression trees.

Some applications that combined deterministic models (ASM type models) and ML methods were also developed (Côté et al., 1995; Lee et al., 2002; Vega De Lille et al., 2015; Zhao et al., 1999). Côté et al., (1995) improved the accuracy of the ASM1 by predicting the remaining errors of the optimized mechanistic model, 5 variables were considerably improved; VSS, COD, $NH_4$, DO and returned activated sludge (RAS). Lee et al., (2002) applied four different strategies for modeling a coke-plant WWTP; only ANN, ASM1 in parallel and in series with ANN and only ASM1. The results showed that the parallel hybrid modeling approach achieved much more accurate predictions with good extrapolation properties as compared to the other modeling approaches even in the case of process upset caused by shock loading of toxic compounds. The accuracy increased from around $R^2=0.7$ using only ASM1 to $R^2=0.95$ using the hybrid model for MLVSS, COD and cyanide concentrations. In a similar approach, Zhao et al., (1999) proposed a simplified ASM2 and applied ANN for improving the accuracy in the prediction of the $PO_4^{3-}$ and $NO_3^-$. Vega De Lille et al., (2015) first applied ANN in an Anammox SBR to predict the ammonium concentration in the effluent of the reactor using online pH measurements. The lab data was used to calibrate the ASM type model of the system instead of feeding it to the ANN. By doing this, the amount of data for ANN increased. Once the ASM was successfully calibrated, the simulation results regarding the ammonium concentration were used as the target data for training the ANN. The previous action improved the learning capacity of the networks by considerably increasing the amount of data in comparison with the available experimental measurements, and the accuracy in the prediction of ammonia increased to 0.99 ($R^2$).

Although data-driven methods have been applied in a variety of bWWTP with highly accurate results, the main concern that remains is the lack of detail in the methodology when the results are reported. Steps such as data pre-processing, data normalization, amount of data considered for analysis are often missing. The main application of these methods in literature has focused on software tools for developing these models without explicit information on data pre-processing. The relevance of these preparatory steps is key in order to develop highly accurate and characteristic models, and more important, to comprehend the limitations of the models proposed (Blum and Langley, 1997; Kohavi and John, 1997).

Clearly, the motivation behind the application of data-driven methods and their popularity in bWWTP is its highly adaptable nature and (usually) computationally faster than other methods, such as ASM. By adaptability, we understand that they can handle the dynamic behavior and complexity of the process well when enough data is provided for training.

### 2.3.2 Challenges

Although the extent to which ML methods have been applied in different bWWTP is wide, important issues and limitations have been identified and are listed below.

- Most of the studies previously reviewed which applied ML methods for modeling bWWTP have the main disadvantage that several steps prior to building the predictive models are not sufficiently explained. The latter leads to a misinterpretation of results and wrong conclusions from the system operation. For proper data analysis, it is fundamental to know the data and carefully pre-process it, since both the amount of data and the pre-processing steps performed are key to extracting relevant information from a system.
- The nature of the data used for analysis is mostly not described in detail. In wastewater treatment processes, heterogeneous datasets are generated: online, off-line and Boolean type parameters. However, the detail on the nature of the data sources, the amount and the quality for analysis, is still missing in the literature.

- Most of the studies in the literature that built predictive models based on supervised ML methods (most of the applications), applied input parameters similar (if not the same) as for ASM type models. Contrary to ASM models, data-driven methods aim to cover most of the data available so new information from the process can be discovered. The great variety of parameters monitored in wastewater treatment can bring relevant information to data-driven models and to data analysis. Parameters such as oxidation reduction potential (ORP), conductivity, acid capacity, turbidity, among others, that could be applied for building more robust models, mostly been overlooked so far.
- The diversity of data sources in wastewater treatment is clear. However, a combination of these data sources for extraction of knowledge is not yet studied. For example; the combination of batch experiment data or online and lab data to study process performance and operation has not been investigated so far.

## 2.4 Outlook and chapter conclusions

A survey through highly cited and recent novel research papers demonstrated that mathematical modeling and ML approaches are important tools for the study, performance, optimization and design and prediction of biological wastewater treatment processes under different operational conditions and configurations. In this work, carbon oxidation and biological nitrogen and phosphorus removal are studied and reviewed from a modeling perspective. The application of mathematical models has started in the beginning of the 1980's, and ever since, the increase of the publications and their use (citations) has increased rapidly over the years.

Initially, N-DN was modeled by ASM1, where 8 biochemical reactions were described, afterwards, an improvement to this model came along with the ASM3 with the main difference of considering endogenous respiration. During the last decade, interest has been directed towards modeling of $N_2O$ emissions and some important modifications to the models, considering more of the intermediate steps in denitrification in ASMN and ASM-ICE. With the discovery of Anammox, the need to incorporate these processes into the ASM1 and ASM3 arose. Different applications were developed, mainly to better understand the dynamics of the bacterial groups and substrates in Anammox based systems for developing efficient operation strategies. As for the models that include the removal of phosphorus, the models evolved in the direction of the inclusion of different processes and compounds, such as denitrifying PAOs, GAOs, and competition between PAOs and GAOS. Nowadays, the production of $N_2O$ by denitrifying PAOs is studied. On the other hand, different works that used data-driven methods in modeling bWWTP have applied these methods mainly to predict the influent and effluent of these systems. However, the methodology, background, nature of data sources and limitations of the models are poorly described in most of the studies. The main software for data-driven methods implementation is the Neural Networks toolbox in Matlab and few to none of the limitations of the approaches are addressed. One of the main goals of this work were to overcome the limitations mentioned in the previous section, for which Chapter 3 shows a detailed study of the steps towards the application of data-driven processes focused on wastewater treatment. Due the increasing interest of the wastewater treatment community in the application of these methods, a benchmark for the application of these models is necessary.

# 3 Data-driven methods in wastewater treatment

## 3.1 Introduction

The current trend towards the use of computer-controlled equipment and more sophisticated instrumentation comes along with large amounts of information generated and recorded commonly in SCADA systems. SCADA platforms are used to monitor and manage a plant or equipment in different industry sectors like telecommunications, water and waste control, energy, transportation, among others, and play an important role in computer based control systems (Dieu, 2001). In wastewater treatment, plant automation came along with the birth of mass production of programmable logic controllers or PLC and of instrumentation control and automation technology (ICA.). Both were important to improve operation for water and wastewater systems and to satisfy both quality of the effluent and efficiency of the operation (Ecob et al., 1995; Olsson, 2012). Both, SCADA systems and ICA are vital nowadays for any WWTP operation; PLC's and remote technical units (RTU's) interpret information from connected sensors and transmit it to the SCADA master. In turn, the PLC and RTU receives control commands in protocol format from the SCADA master, and forward these commands to the appropriate control devices (Ecob et al., 1995). This allows the SCADA master to control specific operational processes all through the network from a single location. Another key development involves the computing and storage capacity, that allows the storage of the large amount of information generated in a WWTP. The vast amount of information generated and recorded in SCADA systems from WWTP involves complex and heterogeneous data sources; on-line from sensors, on/off control data and off-line from laboratories. These large databases generated allow operators and engineers to monitor individual equipment and process performance as well as the water quality to comply with environmental regulations. However, further extraction of knowledge from these databases is challenging without the aid of advanced statistical tools and most importantly, the methodology followed to properly apply these methods.

This chapter aims at covering different topics related to data-driven methods and their application to bWWTP datasets.

## 3.2 Data acquisition and management in wastewater treatment

### 3.2.1 Description data and sources for its generation

Generation of data in wastewater treatment arose from the need of monitoring and controlling the quality of water and removal of pollutants in bWWTP to meet with the environmental regulations (Hreiz et al., 2015; Rieger et al., 2010). In bWWTP from different scales (lab, pilot or full-scale), large amounts of data are generated daily. As briefly described in the previous sections (see Section 2.3 and 3.1), these databases can be classified into; online from sensors and analyzers, on/off online data from controllers (on/off equipment data), and off-line data from laboratory measurements. The measurement interval of online sensor data ranges from seconds to hours, while on/off control data provides Boolean values, which produce non-linear effects on models. Finally, parameters measured in laboratories (off-line data) such as; organic carbon (COD), phosphorus species ($PO_4$-P), ammonium ($NH_4$-N), among others, are often measured few times a week or month to monitor the water quality and validate sensor information (if available). Figure 3.1 illustrates examples of the different data sources; online from sensors, on/off data from controllers and laboratory data.

Figure 3.1 clearly shows the significant differences between the data sources. Data is the core for data-driven methods. The amount and quality of the data will have a significant impact on the results of data-driven methods and at the same time, the applicability of these methods will be subject to both the amount and quality of data (James et al., 2013). Therefore, the understanding and characteristics of different sources of data in bWWTP is crucial for the selection of appropriate data-driven methods for further knowledge extraction or predictive analytics. For example, in supervised ML, the accuracy of a predictive or regression model will not only depend on the method and pre-processing of the data, but the amount of examples or past-experience provided to the model to train. However, due to the differences in the datasets previously explained, the selection of the dataset for building data-driven models is fundamental and will depend on the goal of the study; prediction of an online or off-line parameter. On one hand, online datasets i.e. a group of parameters measured online, provide enough data for building the models (enough past data to train). However, off-line laboratory data can be crucial for monitoring the process performance and therefore should be considered in the model. This leads to the need for a combination of both datasets into one, the resulting dataset will be unbalanced or incomplete, with both weekly and daily values from laboratory and online parameters, respectively.

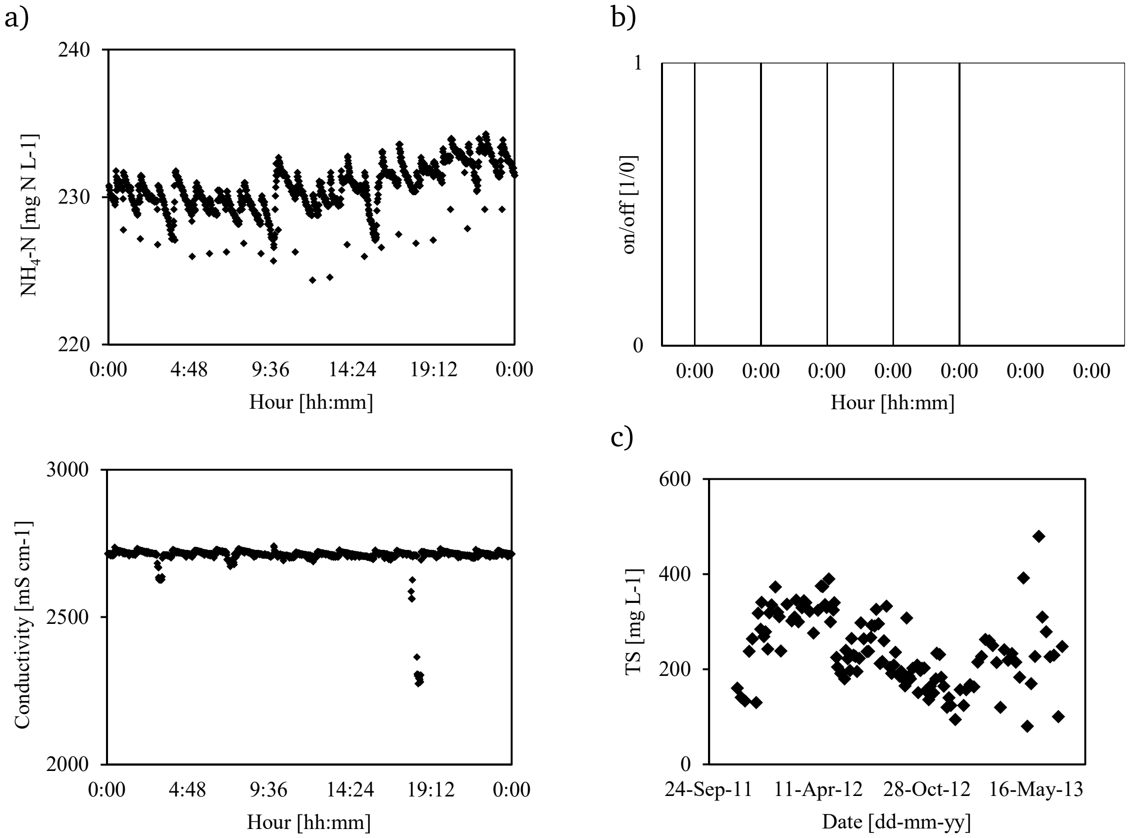


Figure 3. 1 Different data sources found in bWWTP: a) Online sensor data for $NH_4$ concentration (top) and conductivity (bottom) monitored every minute from a full-scale partial nitritation-anammox reactor in hourly resolution. b) on/off data from a feeding pump in a sequencing lab-scale reactor in daily resolution c) total solids concentration measured once a week to month in a full-scale reactor.

Supervised ML provides powerful methods for building predictive models, even with a low amount of observations (Alejo et al., 2018). However, the amount of data is key for the application of more robust and complex methods such as deep convolutional neural networks (CNN), where more parameters in the model are adjusted. Figure 3.2 illustrates the results from three regression methods built to predict the effluent of a full-scale PN-A reactor. I. In this work around 150 data-points were available for training. Neural networks and support vector machines (SVM) were by far more accurate than CNN. Figure 3.2 clearly shows, how the amount of data for training can influence the accuracy of some methods in supervised ML. Due to the nature of CNN, the amount of parameters to adjust in these networks are more and therefore, more examples provide more training iterations and more weights can be updated. However, the amount of training examples as seen here, are not enough to produce an accurate model with CNN.

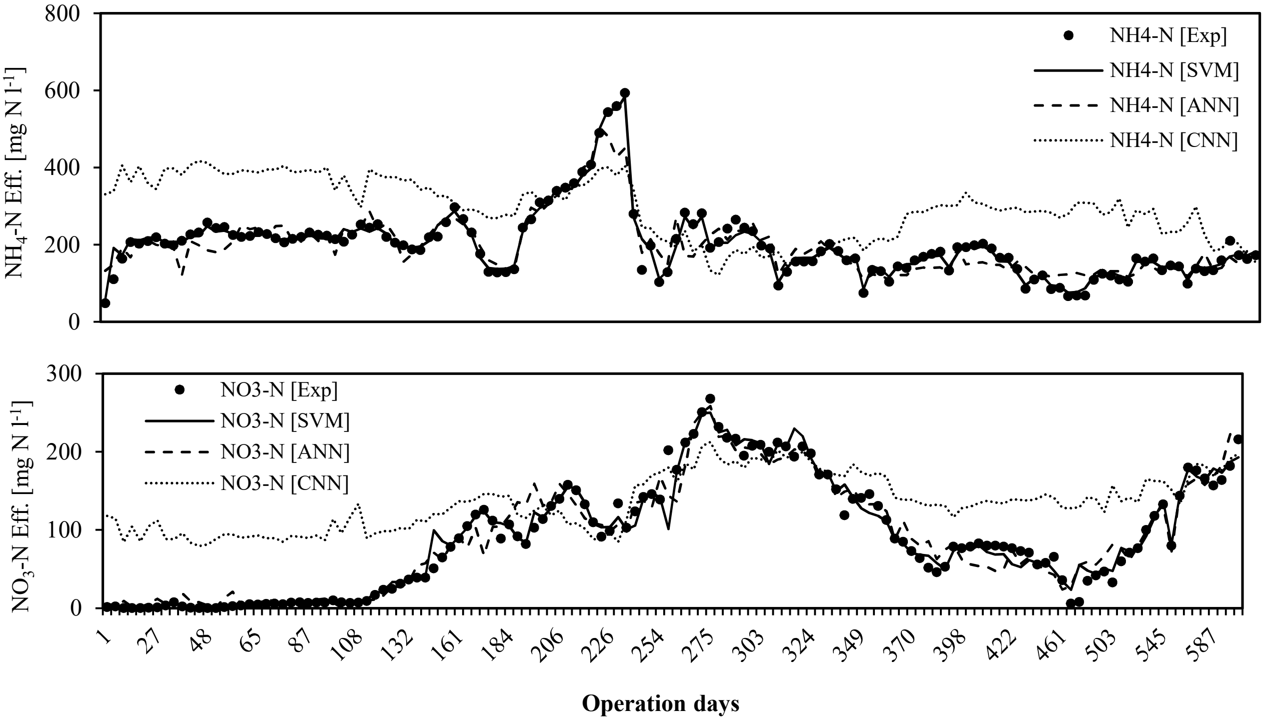


Figure 3. 2 Prediction of nitrogen species in the effluent of a full-scale reactor. Comparison between three approaches: convolutional neural networks (CNN), artificial neural networks (ANN) and support vector machines (SVM).

### 3.2.2 Data acquisition and integration

The first step in data analysis is the data acquisition or collection. Data is collected from different sources (described in a previous section). The data files need to be in a computer readable format depending on the software used for analysis. Tabulated formats commonly used are; comma separated values (CSV/.csv), generic tabular data (.dat), text (.txt), tab separated values (TSV/.tsv), excel spreadsheets (XLSX/.xlsx), among others.

After data acquisition, integration of data is performed (commonly). Data integration involves the combination of different data sources, gathering all the data elements together is not an easy task when the data comes from different sources and they have to be merged in a single dataset. Matching the schema from different datasets involves the removal of inconsistent and duplicated variables as well as redundant and correlated features.

In WWT, the data integration step often involves merging online and laboratory datasets, for which, a time span must be defined to match both sources, i.e. hourly, daily, weekly values, etc. The resulting dataset will be heterogeneous and incomplete i.e. with missing data. The problem of missing data requires further elaboration and statistical analysis. Section 3.2.4 is dedicated only to this issue and the existing statistical methods to deal with missing data or missing values i.e. imputation.

### 3.2.3 Data Cleaning

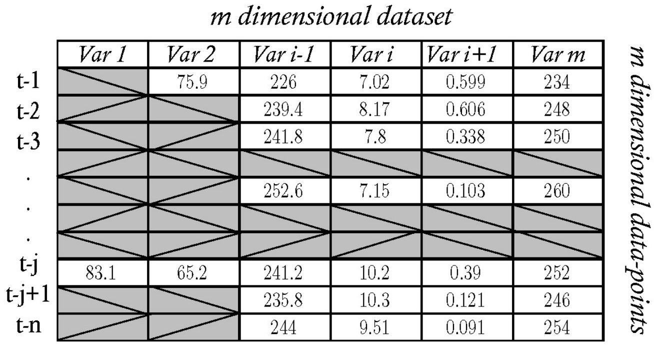
Data cleaning in this work aims at finding and removing redundancy and noise in the data. Redundancy refers to repeated feature and features that can be derived from another variable or set of them and which contribution is not relevant. Inconsistencies in dimension or feature names can cause redundancies also. Redundancy in numerical feature can be analyzed with correlation matrixes while looking at correlations near to 1, whereas redundancy in categorical feature is often studied with the Pearson's Chi-square ($\chi^2$) test. Redundancy should be prevented, it causes an increment of the data size, resulting in longer times for running data-driven methods. Removing noise from WWTP data is a complex step, the removal of outliers/anomalies/noise from data is always subject to the expert knowledge of the process (process engineers, plant operators). The nature of bWWTP processes is dynamic due to its biological components and the variability of the influent wastewater composition. As a result, the data is equally complex and dynamic. A good practice for detecting noise in bWWTP is studying the dispersion, i.e. the degree to which numerical data tend to spread, or variance of the data. The most common measures of data dispersion are range, the five-number summary (based on quartiles), the interquartile range, and the standard deviation. Boxplots can be built based on the five-number summary and are a useful tool for identifying outliers. However, the removal of noise will highly depend on the expert knowledge for verification. In this work, the anomalies/outliers/noise were detected through clustering techniques and the described methods above. The noise was only removed when verified by expert knowledge.

### 3.2.4 Missing data and multiple imputation

Even in well-designed experiments, missing data occur in almost all research. Missing data often restrict the inference power, producing biased estimates and leading to invalid conclusions. Most statistical methods require a complete dataset to be able to make inferences and extract knowledge from it. Unfortunately, the ubiquity of missing data limit this analysis. However, there are some tools and concepts that must be considered and are helpful when dealing with missing data. The concepts and rationale behind missing data were first introduced by Rubin, (1976) who has spawned great amount of statistical literature on this topic up to date and its implications in different fields of research. However, these concepts have been overlooked in the context of data analysis in bWWTP which will be focus of study in this section and reappear in different chapters of this work.

#### i Rationale behind missing data

All datasets consist of a series of variables or features which provide information on a series of *items* or *observations* that can be numerical or categorical (characters) (Carpenter and Kenward, 2012). To explain the nature of missing data or missing values further, the structure of a dataset is described. Figure 3.3 shows an example of a small m-dimensional dataset (m features) and a set of m dimensional data-points. Each m-dimensional data-point in the dataset is a *tuple* (*t-j*) i.e. a finite ordered list or sequence of elements. In Figure 3.3., the dataset consists of n-tuples, each tuple has m dimensions (equal to the number of m features). In bWWTP, the initial row commonly describes the variables or features names in a dataset, in our example; *Var 1*, *Var 2*, *Var i*,…,*Var m*, followed by the m dimensional data-points or tuples in the rows below.



*m dimensional dataset*

| | *Var 1* | *Var 2* | *Var i-1* | *Var i* | *Var i+1* | *Var m* |
|---|---|---|---|---|---|---|
| t-1 | ▨ | 75.9 | 226 | 7.02 | 0.599 | 234 |
| t-2 | ▨ | ▨ | 239.4 | 8.17 | 0.606 | 248 |
| t-3 | ▨ | ▨ | 241.8 | 7.8 | 0.338 | 250 |
| . | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ |
| . | ▨ | ▨ | 252.6 | 7.15 | 0.103 | 260 |
| . | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ |
| | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ |
| t-j | 83.1 | 65.2 | 241.2 | 10.2 | 0.39 | 252 |
| t-j+1 | ▨ | ▨ | 235.8 | 10.3 | 0.121 | 246 |
| t-n | ▨ | ▨ | 244 | 9.51 | 0.091 | 254 |

*m dimensional data-points*

Figure 3. 3 Example of a common dataset structure in bWWTP. Missing values are represented as grey and hatched cells. This dataset is composed by m variables (*Var m*) and n tuples (*t-n*).

Figure 3.3 clearly illustrates an incomplete dataset i.e. incomplete tuples or missing values. Missing data (or missing values) is defined as the data values that are not stored but planned (missing observations). In our example, each tuple was intended to be complete (the side rows contain values), however, for different reasons, these values are missing. The rationale behind the missing data is wide and varies from field to field, however, listed below are few common reasons of missing values in bWWTP (there are definitely more):

- Irregular sampling, a variable is occasionally recorded for validation/monitoring purposes; twice a week to few times a month. Therefore, in a dataset containing daily values, most days will be empty.
- The data values are under or over the detection limit of the equipment or sensor, as a result the values are not recorded.
- Variables that are measured only over a period of time, result in missing values.
- Typographical errors when saving data (mainly from laboratories), characters such as: **#**, **&**, **/**, **\$**, **-**, etc.

The problem of missing data is relatively common in almost all research fields and can have a significant effect on the conclusions that can be drawn from the data analysis. There are two main problems that rise from missing data; loss of efficiency and bias. Loss of efficiency is inevitable since it is impossible to infer from a missing value (unknown measurement) and bias, since the subset of complete tuples may not be characteristic of the whole dataset of study. In Figure 3.3 only one tuple is complete, and it would be impossible to infer from only one data point and extend this knowledge to the complete dataset. The extent of such bias depends on the statistical behavior of the missing data. Missing data can be statistically classified into three categories; missing completely at random, missing at random and missing not at random. A detailed description of these types of missing data is in the section A2 in the Appendix.

### ii Imputation and multiple imputation

Obtaining a complete dataset is often challenging, especially in the field of bWWTP. Due to the reasons previously explained, it is inevitable avoiding to deal with missing data or incomplete datasets. There are two methods for dealing with missing data; removing the tuples with missing data or filling in the missing values (i.e. imputation). Eliminating tuples with missing data is a common practice and likely encountered in data-analysis software tools such as Matlab, R, or Python. Deleting missing values requires the MCAR mechanism and will produce biased parameter estimates when this assumption does not hold. Even if the MCAR assumption is plausible, eliminating data is wasteful and can dramatically reduce the inference power. Consequently, this practice is not recommended unless the proportion of missing data is trivially small. Moreover, in a report from the Task Force on Statistical Inference echoed this sentiment, stating that *"the two popular methods for dealing with missing data that are found in basic statistical packages, list wise and pairwise deletion of missing values, are among the worst methods available for practical applications"* (Wilkinson and Task Force on Statistical Inference, 1999).

It is evident that the primary benefit of list wise deletion (complete deletion of tuples with missing values) is convenient. Restricting the analysis to the complete cases eliminates the need for specialized software and complex missing data handling techniques.

Statistical methods whose main purpose is dealing with missing values is called imputation. There are two general methods for imputation; single and multiple imputation. Single imputation generates a single replacement value for each missing data point. Multiple imputation (MI) on the other hand, imputes missing values with plausible estimates of the missing values for multivariate datasets.

The most common single imputation method is the arithmetic mean imputation or mean substitution, the method suggests the replacement of the missing values with the arithmetic mean of the observed values of the variable of study, i.e. $\overline{\boldsymbol{Var_{\iota,O}}}$ (See Appendix, A2). However, the arithmetic mean imputation considerable distorts the resulting analysis even when the missing data follows the MCAR mechanism. The main drawback of arithmetic mean imputation is its attenuation effect on covariance and correlations between variables in a dataset, since the imputed data will not follow the correlation of the observed data between two parameters, this has a greater effect when there are associations between two variables.

MI on the other hand, was designed to take the errors of the estimation process of Expectation Maximization (EM) algorithms also used for imputation (Carpenter and Kenward, 2012). The advantages are that MI is a less biased imputation method, at the cost of being computationally more expensive. MI is a Monte Carlo approach in which multiple values are generated from the observed data in a way that the incomplete data is filled by repeatedly solving the observed data, the methods to complete or fill the empty values can go from; linear regression methods to decision trees and random forests (Buuren and Groothuis-Oudshoorn, 2011). The main difference with EM is that MI performs several imputations that yield several complete datasets. This repeated imputation can be done thanks to the use of Markov Chain Monte Carlo methods, as the several imputations are obtained by introducing a random component, usually from a standard normal distribution. In a more advanced fashion, MI also considers that the parameter estimates are in fact sample estimates. Thus, the parameters are not directly estimated from the available data but, as the process continues, they are drawn from their Bayesian posterior distributions given the data at hand.

### 3.2.5 Data scaling

Data scaling is one of the most important steps in the pre-processing of data before the application of data-driven methods. Once imputation is performed, the dataset is complete and ready to analyze it with data-driven methods. However, scaling is crucial for some methods in ML. The scaling process allows the transformation of features from different domains to a domain where they are distributed over the same range and order of magnitude. In bWWTP, the monitored parameters are usually characterized with marked skewness or kurtosis. Therefore, it is recommended to transform or scale them for a close normal distribution. The basic idea behind scaling these parameters, is that the expected mean value between each variable should be one (Salama et al., 2010). There are different scaling techniques, most of them are related to the application of the statistical mean and standard deviation of the variables, i.e. subtract the mean from each value, and divide the result by the standard deviation. This process is called standardizing a statistical variable and results in a set of values whose mean is zero and standard deviation is one (Witten et al., 2011). Other scaling methods involve adjusting the variables in the range between 0 and 1, these can be unit range (normalize by the maximum value of each variable) and unit variance (zero mean and unit variance) (Aksoy and Haralick, 2001). It is important to notice that normalization is commonly used interchangeably with standardization. Although, both concepts involve the transformation of data scale, the definitions of both are different. Normalization is the process of rescaling features in the range of 0 to 1 (being 0 the lowest and 1 the highest value). Standardization refers to rescale data to have a mean of 0 and a standard deviation of 1(unit variance).
The datasets studied in this work were standardized, the mean and standard deviation were applied. Eq. 3.1 describes the formula applied for this scaling technique. Given a certain dataset $\{\boldsymbol{Var_1}, \boldsymbol{Var_2}, \boldsymbol{Var_i}, \dots, \boldsymbol{Var_m}\}$ where $i = 1,2, \dots, m$. For a certain variable $\boldsymbol{Var_i}$, the standardization of each observation in $\boldsymbol{Var_i}$ is defined as follows,

$$Var_{i,j_{sc}} = \frac{Var_{i,j} - \mu_i}{\sigma_i} \qquad \text{Equation 3. 1}$$

where $\mu_i$ and $\sigma_i$ are the statistical mean and standard deviation of $\boldsymbol{Var_i}$, respectively. The importance of scaling relies on two reasons: i) most data-driven methods from both supervised and unsupervised ML are sensitive to raw datasets without scaling ii) to avoid the domination of one feature over others, usually caused by differences in orders of magnitude among the feature in a dataset.

## 3.3 Data-driven methods based on machine learning

Statistics provide us tools to build, or more accurately, to find associations between variables in a dataset. These associations are conquered by means of models. For example, given variables; $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_n\}$, where $n$ is positive and finite. It is possible to find a correspondence between $X$ and $Y$, $X \rightarrow Y$. $X$ is called the input, independent variable or feature and $Y$ is the output or dependent variable. There exists some association between $X$ and $Y$,

$$Y = f(X) + \epsilon \qquad \text{Equation 3. 2}$$

Where $f$ is some unknown function of $X$ and $\epsilon$ is an error not associated to $X$. $f$ represents the systematic information that $X$ provides about $Y$. The relation between $X$ and $Y$ drawn from $f$ is relevant for two reasons; prediction and inference. Prediction is when knowing the exact form of $f$ is less relevant than knowing the value of $Y$. In this case, $f$ can be treated as *black-box* because the exact form of $f$ is not relevant, but rather obtaining accurate values of $Y$. On the contrary, when the form of $f$ is relevant; our interest is to understand how $Y$ changes with $X$, we are interested in the inference. ML refers to the set of methods that allows the estimation of $f$ and inferring from $f$, the field of ML is often considered as the evolution of statistical methods (like linear regression) to more complex methods; regression trees, ANN, among others. ML methods require two components, a *machine* and the ability to *learn*. The machine chosen by excellence is the computer, it allows to build computer *learning* programs (ML algorithms.). But how does a computer program learn? Mitchell, (1997) refers to a computer program that learns when this program is able to learn from experience with respect to some task, this is, the performance of learning from the task improves with the experience. However, how this computer *learning* program differentiates with a regular computer program? Figure 3.4 aims at illustrating this difference; Figure 3.4a is a regular computer program and Figure 3.4b is a *learning* program.

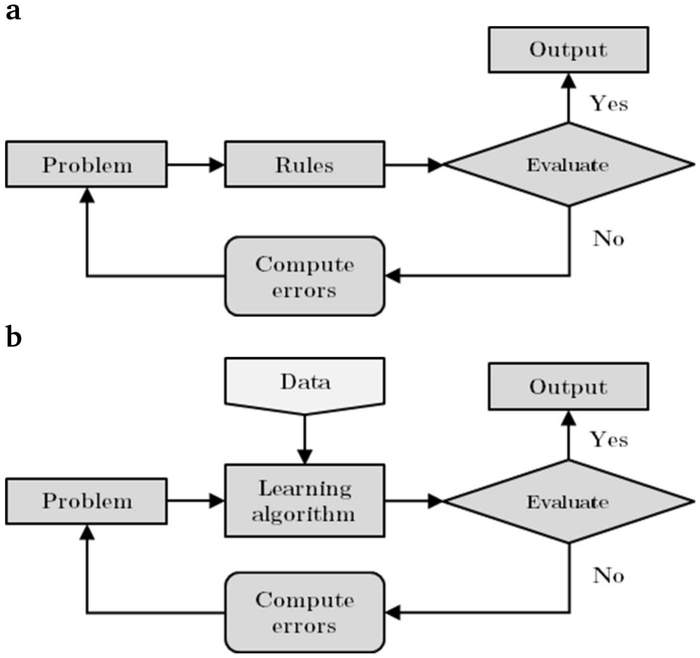


Figure 3. 4 Difference between a regular (a) and a *learning* (b) computer program

A simple example is a regression problem. In bWWTP, it is often desired to automatically predict the effluent composition of a particular process (like an activated sludge system). ASM models are complex and require time and often a measuring campaign to be able to build and validate an accurate model to predict the effluent of the activated sludge system. Given the conditions where building an ASM model is not feasible, an alternative approach would involve creating a computer program based on rules on expert knowledge (Figure 3.4 a). The problem is not trivial, the computer program would be composed by an extensive amount of *if-then* rules to predict the effluent of the activated sludge process and probably will be less accurate than an ASM model. Another alternative is to build a computer *learning* program, a program that is able to i) use data of the WWTP and ii) based on a performance measure (for example, root mean squared error) is able to improve the prediction of the effluent when more data is provided for *training* (Figure 3.4b). As a result, the program improves its prediction of the effluent when more historical data is fed; it allows to adjust the program parameters, re-compute the measure of performance and repeat iteratively until a threshold is met (minimum error between the predicted and

experimental value). When a program is built under these conditions, we refer to this program as ML algorithm.
ML methods can be classified based on the amount of whether they require training data or not. Following this criteria, ML can be classified in four categories; supervised, unsupervised, semi-supervised and reinforcement learning. The example above is part of supervised ML. In this work, methods from supervised and unsupervised ML to analyze data from different bWWTP to find patterns and accurately predict the effluent of highly dynamic and complex processes were applied.

## 3.4 Supervised machine learning

The applications in which the training data comprises samples involving pairs of input data and target are known as supervised learning problems (Bishop, 2006). There are two (main) types of supervised learning problems: classification (categorical features) and regression (numerical features). Classification involves qualitative prediction, predicting a class label; colors, species, weather (rainy, sunny). In bWWTP, categorical features are less common than numerical features. Regression involves quantitative prediction, the literature on the applications of ML in bWWTP has mainly focused on regression and the development of predictive models to forecast the effluent of different processes in different scales (Chapter 2). In both classification and regression problems, more than one input and output variable can be obtained.
A set of steps to obtain an accurate regression model is independent of the method applied; regression trees, logistic regression, neural networks, etc. Figure 3.5 summarizes these steps.

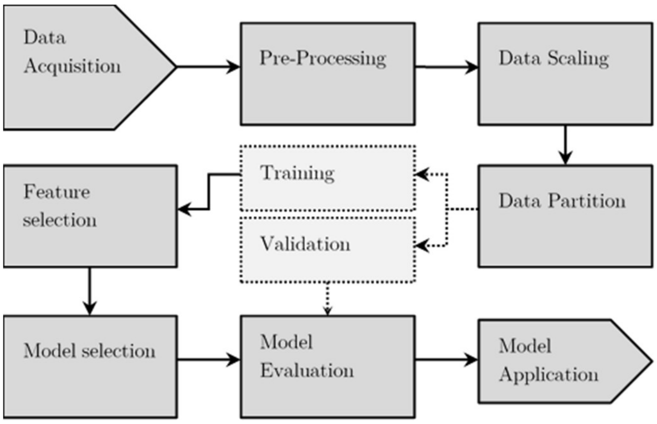


Figure 3. 5 Steps for building a supervised machine learning model

The pre-processing and data scaling are explained in detail in section 3.2. Data partition refers to the selection of an adequate training dataset. Feature selection is discussed in detail in Section 3.6. When building a supervised ML model, many models are usually evaluated and the one with the best performance in the validation set is selected. In this work, different supervised ML methods were applied for prediction purposes. Table 3.1 summarizes which methods from supervised ML were used along this work.

Table 3. 1 Distribution of supervised machine learning methods applied in this work.

| Study | Method | Purpose |
|---|---|---|
| Nitrite accumulation in winter seasons in a WWTP (Chapter 4) | Random forest | Prediction of nitrite accumulation in winter seasons |
| Effluent prediction of full-scale SBR systems (Chapter 5) | Deep neural networks, support vector machines and ensemble learning | Develop predictive models with feature selection optimization |
| Extraction of relevant parameters in PN-A systems (Chapter 6) | Random forest and recursive feature elimination | Combination of heterogeneous datasets to extract relevant information from lab and full-scale PN-A systems |
| Advanced wastewater treatment, modeling extreme low levels of phosphorus (Chapter 7) | Ensemble learning based on support vector machines and convolutional neural networks | Prediction of extremely low levels of phosphorus |

The nature of the data analyzed in bWWTP is highly dynamic due to the; variability in the influent composition and complexity of the biochemical reactions occurring within the processes. Thus, the criteria on the selection of the methods from supervised ML to model these systems was based on their adaptability to dynamic data, their ability to handle high non-linearity and perform appropriately with limited data for training. The methods described in this section meet these requirements and were used in this work to model different bWWTP.

### 3.4.1 Data partition; training and validation datasets

To build a supervised ML model it is necessary to provide a certain amount of examples for the model to learn. After selection of the most relevant features, the next step towards building a supervised ML model is data partitioning into training and validation datasets. The training dataset is comprised of the examples or past experience for the model to learn (to adjust parameters in the model). Once a threshold is met, the model will be evaluated with unseen data i.e. the validation dataset. The partitioning of data can be arbitrary, however, it is recommended that when the amount of data is considerable (thousands of observations), then it is common to hold out one-third of the data for validation and use the remaining two-thirds for training (Witten et al., 2011). A good practice to select the training and validation datasets is *k-fold cross-validation* (Fushiki, 2011). The *k-fold cross-validation* is a resampling procedure used to partition the datasets. The procedure has a $k$ parameter, that refers to the groups that the datasets will be split into. Therefore, the procedure is called *k-fold* cross-validation. A $k$ value of 5 or 10 is recommended, these values are found through experimentation to generally result in a model skill estimate with low bias and modest variance (Witten et al., 2011). In this work, *k-fold* cross validation is used for the selection of the training and validation datasets.

### 3.4.2 Support vector machines and nü support vector machines

Support Vector Machines (SVM) are discriminative classifiers formally defined by a separating hyperplane. In this work, SVM were applied for prediction purposes in Chapters 5 and 7. Given labeled training data, the method computes an optimal hyperplane (the model) which categorizes (or predicts) new examples (Cortes and Vapnik, 1995). An SVM model is a representation of the examples as points in space, mapped via a non-linear kernel function so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Typical kernel functions include linear, polynomial, spline, and radial basis. SVM can also be applied to regression problems, then are referred as support vector regression or SVR. Let $x$ to denote the input vector of the SVM and $z$ to denote the feature space vector which is related to $x$ by a transformation, $z = \phi(x)$. Let the training set $\{x_i, y_i\}$, consist of $m$ data points where $x_i$ is the $i^{th}$ input pattern and $y_i$ is the corresponding target value, $y_i \in \mathbb{R}$. The function $f(x)$ is represented using a linear function in the feature space,

$$f(x) = \omega \cdot \phi(x) + b \qquad \text{Equation 3. 3}$$

where b denotes the bias and ω is the “flatness” parameter (Smola and Scholkopf, 2004). As in all SVR designs, we define the kernel function; $k(x, \hat{x}) = \phi(x) \cdot \phi(\hat{x})$, where “·” denotes the inner product in the $z$ space. Thus, all computations were done using only the kernel function. This inner-product kernel helps in taking the dot product of two vectors in the feature space without having to construct the feature space explicitly. The goal of SVR is to estimate a function $f(x)$ that is as “close” as possible to the target values $y_i$ for every $x_i$ and at the same time, is as “flat” as possible for good generalization. Flatness in the case of Equation 3.4 means that one seeks a small $\omega$. One way to ensure this is to minimize the norm, i.e. $\|\omega\|^2 = \omega \cdot \omega$. This results in a convex optimization problem:

$$\text{Minimize:} \quad \frac{1}{2}\|\omega\|^2 \qquad \text{Equation 3. 4}$$

$$\text{Subject to:} \quad \begin{aligned} y_i - \omega \cdot \phi(x) - b &\leq \varepsilon \\ \omega \cdot \phi(x) + b - y_i &\leq \varepsilon \end{aligned} \qquad \text{Equation 3. 5}$$

The tacit assumption in 3.5 is that a function $f$ actually exists and this function approximates all pairs $(x_i, d_i)$ with $\varepsilon$ precision. However, a more general expression when some errors are allowed, a “soft margin” loss function can be defined with slack variables $\xi_i$ and $\xi_i^*$. The formulation then can be modified to,

$$\text{Minimize:} \quad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \qquad \text{Equation 3. 6}$$

$$\text{Subject to:} \quad \begin{aligned} y_i - \omega \cdot \phi(x) - b &\leq \varepsilon + \xi_i \\ \omega \cdot \phi(x) + b - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned} \qquad \text{Equation 3. 7}$$

The constant C>0 determined the trade-off between the flatness of $f$ and the amount up to which deviations larger than $\varepsilon$ are tolerated. The key idea on solving the inequalities in the problem proposed in SVM (Equations 3.6 and 3.7) is to construct a Lagrange function from the objective function (Equation 3.6), which is referred to the primal objective function, the constraints to the objective function is introduced by a set of dual variables (Equation 3.7). A detailed mathematical development of this problem can be found in Smola and Scholkopf, (2004). From the introduction of the SVR problem, key parameters to consider for further model development are to characterize the hyperparameters in SVR; C, kernel function hyperparameters and $\varepsilon$. The regularization parameter (which prevents a method to overfit) in this type of SVR is C, which often takes a value of 1, however, larger values of C would lead to larger margins of error, which would result in high performance in the training stage, however poor results in validation.

The $\nu$-support vector machines is a class of SVM developed by Schölkopf et al., (2000). It can handle both classification and regression tasks. In this work, the focus is on regression tasks, i.e. $\nu$-SVR. In $\nu$-SVR introduces a new hyperparameter; $\nu$. At each point $x_i$, an error of $\varepsilon$ is allowed. Everything above $\varepsilon$ is captured in the slack variables which are penalized in the objective function via a regularization constant C, chosen a priori. The size of $\varepsilon$ is traded off against model complexity and slack variables via a constant $\nu \geq 0$:

$$\text{Minimize:} \quad \frac{1}{2}\|\omega\|^2 + C\left(\nu\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right) \qquad \text{Equation 3. 8}$$

$$\text{Subject to:} \quad \begin{aligned} y_i - \omega \cdot \phi(x) - b &\leq \varepsilon + \xi_i \\ \omega \cdot \phi(x) + b - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned} \qquad \text{Equation 3. 9}$$

The main contribution of $\nu$-SVR to the regular SVR, is that $\nu$-SVR automatically computes $\varepsilon$, since after the Lagrange construction of Equation 3.8 and 3.9, the $\varepsilon$ parameter is eliminated. The meaning of the $\nu$ parameter in this type of SVR is an upper bound on the fraction of errors and is a lower bound on the fraction of support vectors (Chang and Lin, 2002).

### 3.4.3 Artificial and convolutional neural networks

The brain's architecture was inspiration on how to build intelligent machines. This is the key idea that inspired artificial neural networks or ANNs. McCulloch and Pitts, (1943) first introduced a simplified computational model of how biological neurons work together in animal brains to perform complex operations using propositional logic, this is the first artificial neural network architecture. The Perceptron is one of the simplest ANN architectures introduced by Rosenblatt, (1958), the perceptron represent a neuron and is based on an artificial neuron called a linear threshold unit (LTU): the inputs and output are now numbers and each input connection is associated with a weight. The output of the perceptron is computed through a step function or Heaviside function ($h$). Figure 3.6 illustrates the architecture of a perceptron.

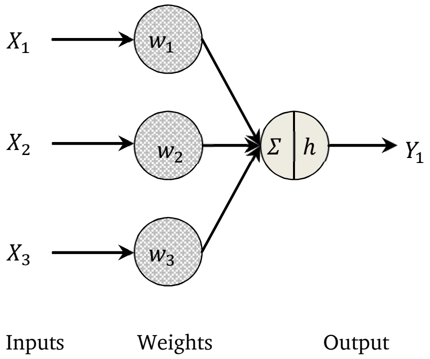


Figure 3. 6 Perceptron architecture and components. Σ denotes the LTU or weighted sum of the inputs *h* the step function.

The LTU computes a weighted sum of its inputs (Equation 3.10), then applies a step function to that sum and outputs the result (Equation 3.11).

$$z = w_1x_1 + w_2x_2 + w_3x_3$$

Equation 3. 10

$$h_w(x) = step(z)$$

$$h(z) = \begin{cases} 0 & z < 0 \\ 1 & z > 0 \end{cases}$$

Equation 3. 11

The hyperparameters in the perceptron architecture are the weights. The Hebb's rule was the first method applied to optimize the perceptron model, which computes the weight in a next step while minimizing the error between the experimental and predicted value. The rate of update of the weight is also defined by a learning rate. The decision boundary of each output neuron is linear, so Perceptrons are incapable of learning complex patterns. To solve these problems, multiple perceptrons can be stacked, thus resulting in a multilayer perceptron or ANN.

In more complex neural networks (NN), learning and training phases involve automatic parameter estimation in a flexible way so that the NN is more characteristic. The basic processing elements of NN are still called neurons or nodes. Figure 3.7 illustrates the basic architecture of a feed forward neural network or FF-NN. When a network is composed by only one hidden layer it is known as an ANN, however, when the amount of hidden layers is higher than two, then it is referred as deep NN or DNN.

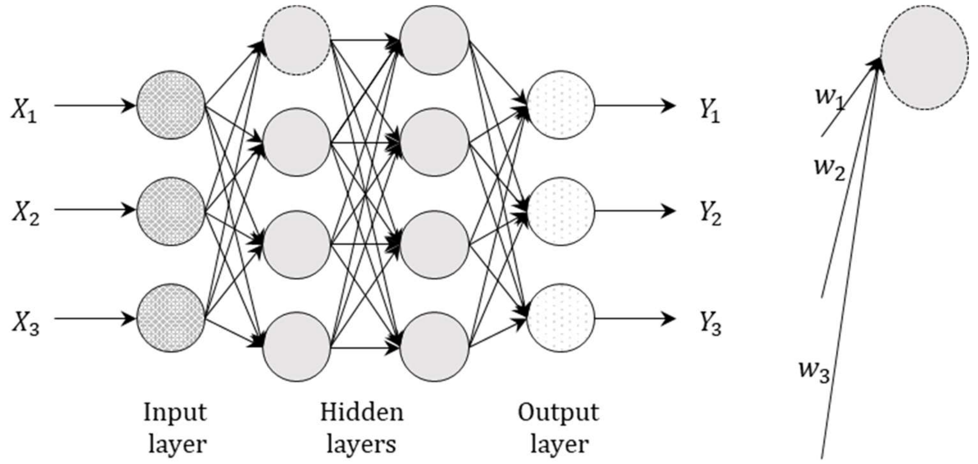


Figure 3. 7 Basic architecture of a deep neural network. In color, a typical artificial neuron and its components are illustrated. The arrows represent the signal from inputs (x1, x2 and x3) to outputs (o1 and o2).

In a simplified mathematical model of the neuron, the effects of the synapses are represented by connection weights ($w_1, w_2, w_3$) (Figure 3.7) that modulate the effect of the associated input signals, and the nonlinear characteristic exhibited by neurons is represented by a transfer function ($f_1$), this transfer function is analogous to the Heaviside step function. However, due to the optimization criteria of the weights in DNN, the Heaviside function is replaced by a sigmoid, hyperbolic tangent function or a rectified linear unit (ReLU), the main characteristic of these functions is that they are derivable in their domain. The neuron impulse is then computed as the weighted sum of the input signals, transformed by the transfer function. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm. Referring to Figure 3.7, the signal flow from inputs $x_1,\ldots,x_n$ is considered to be unidirectional (i.e. feed forward), which is indicated by arrows. Every output of a neuron can be computed through Equation 3.12

$$Y_i = f\left(\sum_{j=1}^{n} w_j X_j\right)$$ Equation 3. 12

The method commonly applied for computing and optimizing these weights is the backpropagation algorithm. This algorithm is similar to the gradient descent method using reverse mode auto-differentiation. The main idea behind the backpropagation algorithm is that for each training sample, the algorithm feeds it to the network and computes the output of every neuron in each consecutive layer (this is the forward pass). Then it measures the network's output error (i.e., the difference between the desired output and the actual output of the network), and it computes, how much each neuron in the last hidden layer contributed to each output neuron's error. It then proceeds to measure how much of these error contributions came from each neuron in the previous hidden layer — and so on until the algorithm reaches the input layer. This reverse pass efficiently measures the error gradient across all the connection weights in the network by propagating the error gradient backward in the network (hence the name of the algorithm) (Géron, 2019).

Convolutional neural networks (CNN) are neural networks inspired by the neurons in the visual cortex of the brain. Many neurons in the visual cortex have a small local receptive field, meaning they react only to visual stimuli located in a limited region of the visual field. Thus, the CNN are capable to retain just relevant features or characteristics in an image. A CNN, contains layers that are only partially connected to the lower layers. Lecun et al., (1998) first introduced CNN for use for digital and hand writing recognition.

### 3.4.4 Decision trees

The basic setup of Decision Trees is supervised ML and they are built based on if-then rules. The most common algorithms is the Classification And Regression Tree or CART, introduced by Breiman et al., (1984). They are binary decision trees (split always in two subsets). Decision trees are very powerful for both regression and classification tasks and they can be combined to build up even more powerful methods, such as, random forests. Due to the capability of decision trees and further applicability in this work, they are thoroughly explained here. Figure 3.8 shows the main structure and components of a binary decision tree or CART. Decision trees can be applied for classification and for regression. The main idea is to split the training data into subsets based on rules. In the decision tree in Figure 3.8, the first node (the root node) will split the total amount of training data (all samples) using the feature $X_i$ and a threshold: $X_i \leq a$. $X_i$ is a predictor and $a \in \mathbb{R}$ is a value in the domain and range of $X_i$. The total amount of samples will split into two subsets in the root node (depth 0). If the rule is met for a new sample ($True$), the suggested value of the output (feature), $Y_i$, is defined by the left node, this node is a leaf node because it does not split further, it is pure (gini or mse=0). A node is pure (gini or mse=0) if all training instances belong to the same value or class. If the new sample is greater than $a$; $X_i > a$, the decision must move to the right child node (depth 1), and another rule prevails in this node; $X_i \leq b$. If $True$, then the value of $Y_i$is defined by the left node (also a leaf node) (depth 2), on the contrary the new sample takes the value $Y_i$ on the right node, a leaf node. Each node holds relevant information; the rule of splitting (root and child nodes), number of samples, the performance measure (gini index, mse or entropy) and the value of the output $Y_i$ in the node.

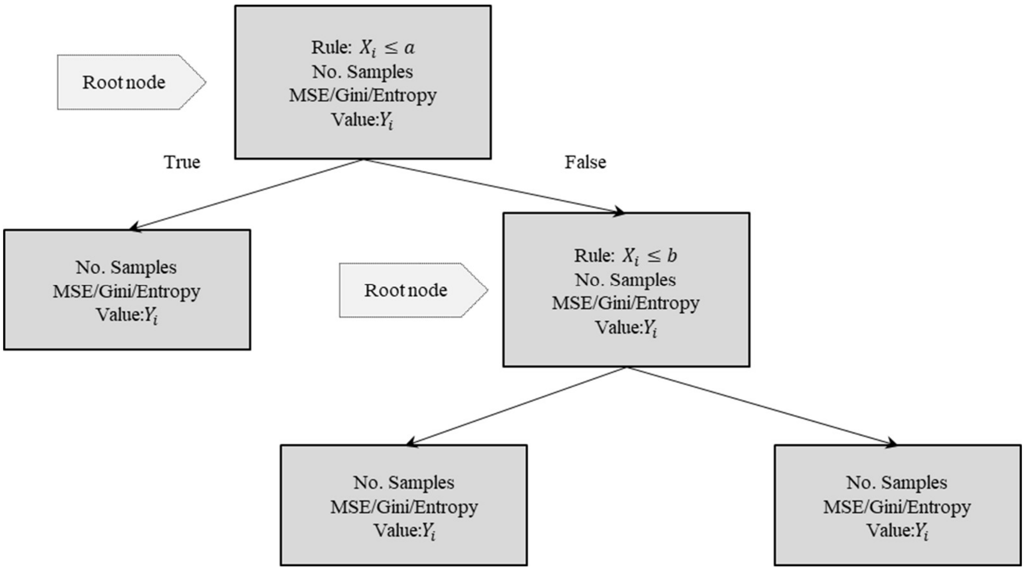


Figure 3. 8 Basic architecture and main components of a binary decision tree.

A decision tree can be either a classification or a regression tree, which depends on the performance measure of the tree. A classification decision tree (classification into classes, categorical feature) measures the gini index, which is a measure of a feature's impurity. A node is pure if the gini index is equal to 0. The gini index is defined in Equation 3.13

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \quad \text{Equation 3. 13}$$

$p_{i,k}$ is the ratio of class $k$ instances (samples) among the training instances in the $i^{th}$ node. The selection of the rule ($t_k$) and the feature for the root node and child nodes aims at finding the purest subsets. The algorithm minimizes the following function,

$$J(k, t_k) = \frac{m_{left}}{m} \mathrm{G}_{\mathrm{left}} - \frac{m_{right}}{m} G_{right} \qquad \text{Equation 3. 14}$$

$m_{left/right}$is the number of instances in the left/right subset and $\mathrm{G}_{\mathrm{left}}$measures the impurity of the left/right subset.

A regression tree predicts a numerical value instead of a class in each node. The algorithm splits each region in a way that makes most training samples as close as possible to the predicted value.

In a regression tree, the aim is to minimize the mse while splitting the training data (Equation 3.15).

$$J(k, t_k) = \frac{m_{left}}{m} mse_{left} - \frac{m_{right}}{m} mse_{right} \qquad \text{Equation 3. 15}$$

Where $mse_{node} = \sum_{i \in node} \left(\hat{y}_{node} - y^{(i)}\right)^2$ and $\hat{y}_{node} = \frac{1}{m_{node}} \sum_{i \in node} y^{(i)}$. Here $y^{(i)}$is the actual prediction, $m_{node}$is the number of observations in the node and $\hat{y}_{node}$ is the mean response of the training observation in the node evaluated.

Decision trees are intuitive and their decisions are easy to interpret, these models are often called white box models. Additionally, decision trees require very little data preparation, they do not require feature scaling or centering. However, one of the main drawbacks of decision trees is that they can easily overfit the data if not restricted. To avoid overfitting (regularization), a good practice is to limit the depth of the tree and reduce the risk of overfitting. Decision trees make orthogonal decisions (perpendicular to an axis), are sensitive to rotation of the training dataset i.e. are sensitive to small changes in the training dataset. As a result, they do not generalize well. An ensemble of decision trees could limit this instability while combining the prediction of each individual tree. Ensemble learning will be subject of study in the next section and applied in Chapter 4 to produce highly accurate models for modeling nitrite accumulation in a WWTP.

### 3.4.5 Ensemble learning: random forest and gradient boosting

Ensemble methods are composed of a group of weak learners which usually perform better than a single one, thus: *wisdom of the crowd*. The group of weak learners or predictors is called an ensemble; a technique is called *ensemble learning,* and an *ensemble learning* algorithm is called an ensemble method. There exist different types of ensemble methods; voting ensembles, *bagging*, *pasting*, gradient boosting and random forest. They differ from each other depending on the training process or the output computing. In this work gradient boosting and random forest ensembles for prediction and feature selection were applied in Chapter 5 and Chapter 6, respectively.

Briefly, voting ensembles combine different methods; support vector machines, decision trees, logistic regression, among others. All the predictors will aim at predicting the same output, the output of the voting ensemble will be the one that gets the most votes. For example, a classification voting ensemble built to predict with two possible outputs; A or B. If most of the predictors classify a new sample to belong to class A, then the output of the voting ensemble will be A.

*Bagging* (bootstrap aggregating) uses the same training method or predictor(Breiman, 1996). Each predictor in the *bagging* ensemble will be trained with random subsets of the training dataset. In *bagging*, a sample can randomly appear in two of the training subsets or more, the sampling is *with replacement*. When the sampling of the training subsets is without replacement, the ensemble method is called *pasting*. The output of the *bagging* and *pasting* ensembles will be the average of the predictors in the ensemble.

Gradient boosting is a type of ensemble method that combines several weak learners sequentially, by adding predictors to the sequence tries to fit the new predictor to the residual errors made by the previous predictor. The output of a gradient boosting ensemble will depend on the sequence of the predictors output.

Random forest is an extremely powerful method for classification and regression tasks and was also introduced by Leo Breiman (Breiman, 2001). This method could be considered a *bagging* ensemble composed by hundreds to thousands of decision trees, the sampling of the training samples however, is slightly different in random forests. Same as *bagging*, the training subsets are composed of random samples from the training dataset. In addition, in random forest, random features are selected for each subset. Random forest achieves higher accuracy with low bias and variance than other popular tree structured algorithms like CART(Wang et al., 2016). Compared to *bagging* ensemble, random forest is a more convenient and optimized ensemble for decision trees, it introduces randomness when growing the forest and usually generalizes better.

In decision trees, the best feature to split the training data among all feature is searched, in random forest, the best feature among random subsets of feature is searched. The resulting ensemble is more diverse. Comparable to decision trees, important features appear closer to the root node, while unimportant ones appear closer to the leaf nodes or not at all. An estimate of the feature importance can be obtained by computing the average depth at which it appears across the trees. In random forest, the same can be done for all trees in the forest, this internal measure of the random forest is known as feature importance, score of importance or ranking of importance.

The ranking of importance is commonly used as a feature selection method, more on this subject in section 3.6.

## 3.5 Unsupervised machine learning

The purpose of this section is to introduce the fundaments and key ideas behind unsupervised ML. In unsupervised ML, prediction is not the goal of study, rather to find representations of the multidimensional data (>3D), i.e. patterns. Unsupervised ML or unsupervised learning are a set of statistical intended tools to discover patterns from the population analyzed (dataset). The interpretation of these patterns however require knowledge of an expert validate the results. In this work, two particular methods were studied and therefore are explained in this section: hard clustering and principal component analysis. A good practice involves the application of unsupervised learning in the exploratory analysis phase of data analysis and sometimes in parallel to supervised learning methods, however it strongly depends on the goal of study. Unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise (Bousquet et al., 2011). In this work, unsupervised ML methods were used to explore patterns in heterogeneous datasets of different bWWTP, more specific, clustering methods are applied in Chapters 4 and 6, principal component analysis (PCA) was mainly applied to illustrate the results, due to its ability to illustrate high dimensional data in two dimensions, thus, dimensionality reduction.

### 3.5.1 Principal component analysis

PCA is a process by which principal components are computed and used to understand the data. The principal components show the dispersion of data, while computing the largest variance among a multidimensional dataset. Through the computation of the principal components, it is possible to find a low-dimensional representation of the dataset while preserving the variance information, thus it is possible to illustrate the multidimensional dataset in terms of the principal components in a lower dimensional set-up. The principal components are the eigenvectors of the covariance matrix of a dataset $\boldsymbol{X} = \left(x_1, x_2, \ldots, x_p\right)$ where $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{p}\times\boldsymbol{N}}$ and composed by $p$ features and $\boldsymbol{N}$ observations, also $\boldsymbol{X}$ has a mean of zero (scaled). The covariance is a descriptive measure of the linear association between two variables. A positive value of the covariance indicates an increasing linear relationship whereas a negative value indicates a decreasing linear relationship. When the covariance approaches zero, no linear relationship is found. However, the covariance does not measure the strength of association between two variables, the correlation does. The covariance between feature $x$ and $y$ is defined in Equation 3.16.

$$Cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N} \qquad \text{Equation 3. 16}$$

Table 3.2 shows the covariance matrix of the dataset $\boldsymbol{X}$. Since $Cov(x,y) = Cov(y,x)$ and $Cov(x_1,x_1) = Var(x_1)$, the covariance matrix of the dataset $\boldsymbol{X}$ is symmetric.

Table 3. 2 Covariance matrix of dataset *X*

| | | | |
|---|---|---|---|
| $Var(x_1)$ | $Cov(x_1,x_2)$ | … | $Cov(x_1,x_p)$ |
| $Cov(x_2,x_1)$ | $Var(x_2)$ | … | $Cov(x_2,x_p)$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $Cov(x_p,x_1)$ | $Cov(x_p,x_2)$ | … | $Var(x_p)$ |

The principal components of the covariance matrix i.e. the eigenvectors of the covariance matrix, are often computed through single value decomposition (it is more efficient), the dataset $\boldsymbol{X}$ can be decomposed in three matrixes; $\boldsymbol{X} = U\Lambda\mathrm{V}$. $U$ and $V$ are unitary matrixes ($UU^T = I$) and $\Lambda$ is a diagonal matrix. The values in $\Lambda$ are the single values, $U$ contains the single vectors on the left and $V$ the single vectors on the right of $\Lambda$. By ordering the single values or eigenvalues in $\Lambda$ in a descending order, we can find the corresponding eigenvector or principal components. The first principal component will hold the largest variance, the second component the second largest and so on. Moreover, since $\boldsymbol{X}$ is symmetric, then the eigenvectors are orthogonal to each other. The eigenvectors will show the directions of the spread of data and the eigenvalues will indicate the magnitude of this. The principal components will be a normalized by linear combination of the features in ***X***, Equation 3.17 shows the first component $z_1$.

$$z_1 = \phi_{11}x_1 + \phi_{21}x_2 + \cdots + \phi_{p1}x_p$$

Subject to: $$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Equation 3. 17

Where $\phi_{i,p}$ is referred as *loading*. The dimension $k$ of $Z = (z_1, z_2, \dots z_k)$ will always be lower than $p$ (dimension of $\boldsymbol{X}$).

### 3.5.2 Hard clustering methods

The goal of clustering is to group the elements of a dataset according to a similarity measure. A dataset $\boldsymbol{X}$ can be understood as a multidimensional space and with clustering we aim at splitting this space into cohesive groups. The main objective of clustering is to find $K$ clusters that satisfy the double property of maximum cohesion and maximum separation. Mathematically it is easier to employ the inverse of a similarity measure, this is, a distance function, the most common vectorial distance measure is the Euclidean distance. The Minkowski distance is a generalization of the Euclidean distance when $p = 2$ (Equation 3.18).

$$d_p(\bar{x}_1, \bar{x}_2) = \left( \sum_{i=1}^{N} \left| \bar{x}_1^{(i)} - \bar{x}_2^{(i)} \right|^2 \right)^{\frac{1}{p}}$$

Equation 3. 18

There are two main classifications of clustering methods; hard and soft clustering. In hard clustering techniques each sample of the dataset is assigned to only one cluster. *k-means* and hierarchical clustering are hard clustering methods. Soft clustering techniques are based on a probabilistic approach, in these methods, the probability of a sample to belong to a determined cluster is computed:

$$c(\bar{x}_i) = (p(\bar{x}_i \in C_1), p(\bar{x}_i \in C_2), \dots, p(\bar{x}_i \in C_k))$$

Equation 3. 19

Fuzzy c-means and expectation maximization methods are soft clustering techniques and measure the degree of membership of a sample to the clusters, however, these methods were not applied in this work.

### k-means

The simplest implementation of the principle of maximum internal cohesion and maximum separation is k-means clustering. k-means tries to minimize the total average intra-cluster distance between sample $\bar{x}_i$ assigned to a cluster $K_i$ and its centroid ($\mu_j$), known as inertia ($S(t)$).

$$S(t) = \sum_{k=1}^{K} \sum_{\bar{x}_i \in K_j} \left\| \bar{x}_i - \mu_k^{(t)} \right\|^2$$

Equation 3. 20

$S(t)$ cannot be considered as an absolute measure because its value is highly influenced by the variance of the samples. The first step in k-means is selecting random centroids and assign each sample in the dataset to the cluster whose centroid has the smallest distance from $x_i$:

$$c\left(\bar{x}_i; M^{(t)}\right) = argmin_j^{(t)}\ d(\bar{x}_i, \bar{\mu}_j^{(t)})$$

Equation 3. 21

Once the assignments are completed, the centroids are recomputed as arithmetic means:

$$\mu_j^{(t)} = \frac{1}{N_{K\,j}} \sum_{\bar{x}_i \in K_j} \bar{x}_j$$

Equation 3. 22

This process continues until the centroids stop changing ($S(0) > S(1) > \cdots > S\left(t_{opt}\right)$). However, the initial centroids selection will highly influence the computational time until the optimal centroids are found. Most statistical softwares such as R and Python perform a variable number of initializations and select the one whose initial inertia is the smallest. Although high amounts of clustering algorithms have been created after k-means, this method is still widely applied. The main advantages of k-means are: i) easy to implement and interpret the results ii) allows the organization of data into sensible groupings iii) is adaptable to new data iv) the data to be analyzed does not require labels that tag the examples with prior identifiers.

### Selection of the optimal number of clusters

One of the biggest drawbacks of k-means and similar algorithms is the explicit request for the number of clusters. Commonly it is necessary to evaluate different metrics for finding an appropriate number of clusters. One method is the *elbow-criterion*; this method is a graphical method where the inertia or within cluster sum of squares (*Wk*) is associated to an increasing amount of clusters. When the number of clusters is very small, the density is proportionally low, hence the cohesion is low and, as a result, the inertia is high. Increasing the number of clusters forces the model to create more cohesive groups and the inertia starts to decrease abruptly. If we continue this process, we will observe a very slow approach towards the value corresponding to a configuration where the number of samples is equal to the number of clusters. The *elbow-criterion* suggests to pick the number of clusters corresponding to the point that separates the high-variation region from the almost flat one (the curve appears like an elbow). In this way, we are sure that all clusters have reached their maximum cohesion without internal fragmentation. However, many clustering analysis problems also require additional metrics, the silhouette score will also be studied for selecting the number of clusters (Rousseeuw, 1987).

### DBSCAN

DBSCAN stands for Density-Based Spatial Clustering and Application with Noise and is a density-based clustering algorithm of the type hard-clustering. This method was introduced by Ester et al., (1996). The main goal is to identify the clusters in a dataset of any shape containing noise and outliers. Three main advantages of DBSCAN exist; i) unlike K-means, DBSCAN does not require the user to specify the number of clusters to be generated (other parameters are optimized that will lead to a certain amount of clusters) ii) DBSCAN can find cluster with arbitrary shape; spherical, linear, elongated, etc. iii) DBSCAN has good efficiency on large databases and can identify outliers.

The key idea is that for each point of a cluster the neighborhood of a given radius has to contain at least a minimum number of points, i.e. the density in the neighborhood has to exceed some threshold. Two important parameters are optimized in DBSCAN; epsilon ($\epsilon$) and minimum points. The parameter $\epsilon$ defines the radious in which a certain minimum amount of neighbor points are found. Additionally, the points can be classified through DBSCAN in three categories; a border point, a core point and an outlier. A core point is understood as the point which neighbors are greater or equal to the minimum points, a border point would be if the number of neighbors is less than the minimum points but it belongs to the radius $\epsilon$, finally DBSCAN classifies a point as an outlier when it is neither a core or a border point.

K-nearest neighbors was the method applied for computing the value of epsilon (Altman, 1992). The neighbor's method returns two parameters, one which contains the distance to the closest neighbor's points and the other which contains the index for each of those points. The aim is to determine the "knee", which corresponds to the optimal *eps* parameter in the plot of these two parameters.

## 3.6 Dimensionality reduction

The goal of dimensionality reduction is to transform a high-dimensional dataset into a lower dimensional one. Dimensionality reduction is applied mainly to i) remove redundant information from the dataset, ii) reduce the computational time and storage space required iii) improves the interpretation of the dataset since it becomes easier to visualize in a very low dimension such as 2D or 3D (with PCA). In this work, PCA and feature selection were applied to reduce the dimension of the datasets studied, however PCA was applied only for the illustration of multidimensional datasets.

### 3.6.1 PCA for dimensionality reduction

From section 3.5, the first component direction describes the greatest variability of the data. For example, for a two dimensional dataset with feature $x_1$ and $x_2$, the first principal component would be defined by Equation 3.23

$$z_1 = \phi_{11}(x_1 - \bar{x}_1) + \phi_{21}(x_2 - \bar{x}_2) \qquad \text{Equation 3. 23}$$

The loadings $\phi_{11}$and $\phi_{21}$would explain most of the variability of the data. The idea behind PCA for dimensionality reduction involves constructing the first M principal components, for our purposes, the first two principal components; $z_1$and $z_2$. In this work, PCA is applied for data visualization in clustering analysis over the two first principal components.

### 3.6.2 Feature selection

There is a vast amount of feature selection techniques and new being developed day by day. There are three main approaches to feature selection in supervised ML; *wrapper, filter and embedded methods* (Chandrashekar and Sahin, 2014). *Wrapper* methods evaluate subsets of feature, which allow to detect possible interactions between them, by using learning algorithms. *Filter* methods are used as a pre-processing task to rank features wherein highly-ranked features are selected and applied to a predictor. However, no learning algorithm is used. Finally, *embedded* methods are a combination of *wrapper* and *filter* methods, where the learning task and the ranking task cannot be separated, such as in *random forest* methods (Granitto et al., 2006). Previously mentioned in Section 3.4, the random forest feature selection process is based on the *importance score.* The hundreds to thousands of decision trees are independently built, and therefore, the trees are de-correlated. Each tree is built over binary decisions and a heuristic is involved to build the tree (*wrapper* method characteristic). As a result, the best features can be extracted from the average ranking, the features at the top of the list being the most relevant (Section 3.4). Random forest can be combined to recursive feature elimination to perform feature selection. In recursive feature elimination, each predictor is ranked using the *importance score* and the performance of each predictor is retained in multiple iterations for prediction tasks. The top ranked predictors are retained, the value of the best performances are determined and are selected as best features. In this work, random forest *importance score* and recursive feature elimination were used as dimensionality reduction task for regression models.

## 3.7 Implementation of methods

Two programming languages were used for the implementation of the methods described in this chapter; Python 3.7 and R 3.6 (Oliphant, 2007; R Foundation for Statistical Computing, 2016). Jupyter notebooks in Anaconda platform and R-Studio where used as development tools for the implementation of the methods. Different libraries were used for different purposes. Table 3.3 summarizes the name of the libraries, the language and the scope.

Table 3. 3 Libraries used in the analysis of WWT dataset in this work.

| **Method [*Library*]** | **Scope** | **Language** | **Reference** |
|---|---|---|---|
| k-means [*stats*] | k-means clustering | R | Default library R |
| k-means [*cluster*] | Visualization k-means | R | (Maechler et al., 2018) |
| Recursive feature elimination [*caret*] | Feature selection | R | (Kuhn, 2008) |
| Random forest [*randomForest*] | Random Forest/Feature selection | R | (Wiener and Liaw, 2002) |
| PCA [*stats*] | PCA | R | Default library R |
| SVM [e1071] | Support vector machines | R | (Meyer et al., 2019) |
| ANN [*neuralnet*] | Artificial Neural nets | R | (Guenther and Fritsch, 2010) |
| Multiple imputation [*mice*] | Multiple Imputation | R | (Buuren and Groothuis-Oudshoorn, 2011) |
| Standarization [*base*] | Scaling | R | Default library R |
| DBSCAN clustering [*dbscan*] | DBSCAN clustering | R | (Hahsler et al., 2019) |
| k-means [*scikit-learn*] | k-means clustering | Python | (Pedregosa et al., 2011) |
| Min-Max scaling [*scikit-learn*] | Scaling | Python | (Pedregosa et al., 2011) |
| Gradient Boosting [*scikit-learn*] | Ensemble learning | Python | (Pedregosa et al., 2011) |
| ANN [*Tensorflow/Keras*] | ANN | Python | (Abadi et al., 2016; Chollet, 2015) |
| 1D-CNN [*Tensorflow/Keras*] | 1D-CNN | Python | (Abadi et al., 2016; Chollet, 2015) |
| SVM and $\nu$-SVM [*scikit-learn*] | SVM | Python | (Pedregosa et al., 2011) |
| Decision trees [*scikit-learn*] | Decision trees | Python | (Pedregosa et al., 2011) |
| Visualization [*matplot-lib*] | Visualization | Python | (Hunter, 2007) |
| Array handling [*NumPy*] | Data handling | Python | (van der Walt et al., 2011) |
| Data frame handling [*Pandas*] | Data handling | Python | (McKinney, 2010) |

# 4 Data analysis on the build-up of nitrite in a WWTP

## 4.1 Introduction

Around one hundred years ago, the experiments held by Ardern and Lockett, (1914) on sewage aeration introduced the conventional activated sludge process, which led to the current concept for biological wastewater treatment. Regardless of the environmental benefits of WWTP, the biological processes involved are not devoid of operational problems, such as seasonality effects (temperature variations between winter and summer seasons). Nitrite ($NO_2^-$) is an intermediate product of N-DN (Chapter 2). During warm seasons (late spring and summer), nitrite concentrations in the effluent of WWTP are generally low. However, increased concentrations of nitrite are observed in cold seasons and can be featured to different factors: disturbance of the microbiological processes, insufficient aeration capacity and unfavorable conditions for NOB, caused by lower temperatures (Alawi et al., 2009; Burrell et al., 1999; Randall and Buth, 1984). In this chapter, the vast amount of information generated in a WWTP was used as an opportunity to tackle the problem of nitrite accumulation during winter season approaching it from a different perspective. The large heterogeneous data sources gathered in this WWTP were:

- online from sensors: the measurement interval of sensors ranged from seconds to hours
- laboratory data: parameters were measured in laboratories such as; organic carbon (as chemical oxygen demand or COD), phosphorus species (such as phosphate), ammonium ($NH_4$-N), among others, are often measured several times per week or month to monitor the water quality and validate sensor information (if available).

The resulting database comprises both online and laboratory parameters which are produced at different sample rates and conditions. Figure 4.1 illustrates the amount of observations (data-points) for around 130 parameters monitored over 13 years of operation in the municipal WWTP studied in this work.

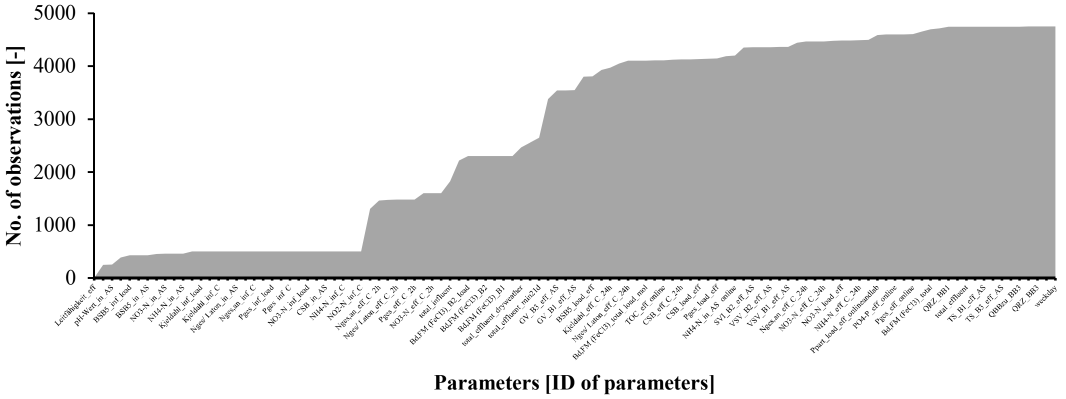


Figure 4. 1 Amount of recorded observations for common parameters in 13 years of operation.

Data-driven methods are especially suitable whenever the rate of data acquisition surpasses the ability to analyze the data, which is particularly true for WWTP operations (Dürrenmatt and Gujer, 2012). To address the problem of nitrite accumulation through data analysis in the heterogeneous dataset analyzed, a methodology based on data-driven methods to enhance knowledge extraction was developed. The main hypotheses of this chapter establish that through data segmentation or partitioning into subgroups i) the resulting analysis would lead to a better understanding of the system ii) the influence of both; amount of data points and number of variables, on further data analysis tasks, would be elucidated and, ii) insights on the nature and parameters influencing nitrite accumulation in cold seasons can be extracted through data-driven methods.

## 4.2 Methodology

Data from a WWTP in south Hesse, Germany gathered over 13 years of operation (2006-2018) was analyzed. The WWTP treats the wastewater of 74,000 person-equivalents. The installed technologies at the WWTP include mechanical, chemical and biological treatment. The biological treatment is composed of three conventional activated sludge lines (L1-L3) for nitrification-denitrification, including secondary clarification. The collected data reported values for 129 parameters. The direct application of data-driven methods to the dataset illustrated in Figure 4.1, requires *data completeness* (the same number of observations for each parameter). However, the amount of missing values in this dataset was high (Rubin, 1976). The process of knowledge extraction from this heterogeneous dataset are explained below.

**Step 1: Visualization of missing values in the raw dataset**

A glance of the missing values in the dataset is illustrated in Figure 4.2, the missing values in our dataset appear in black and cover 34% of the total amount of observations.

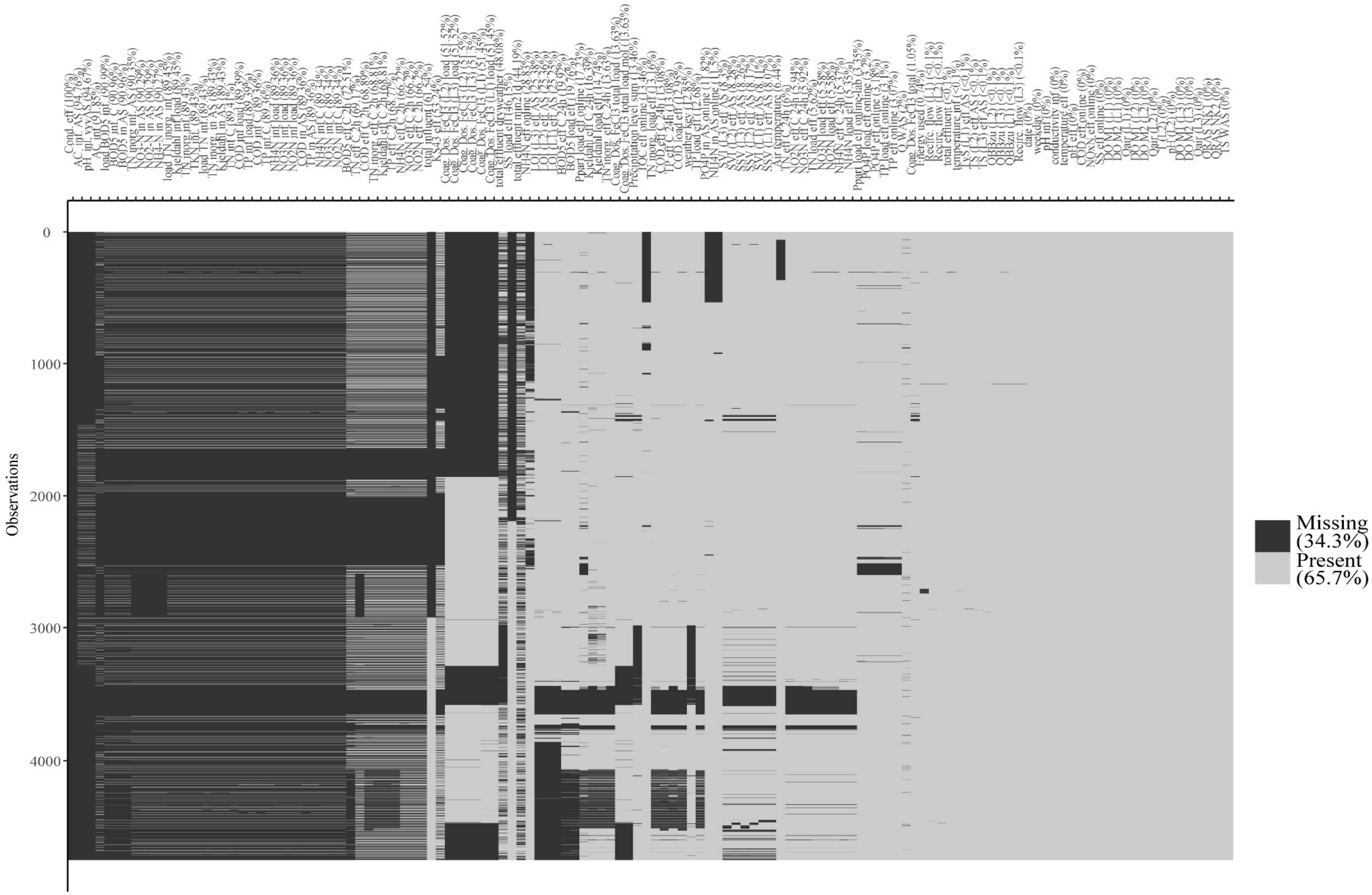


Figure 4. 2 Visualization of the missing values in the complete dataset studied. Maximum number of observations per feature are 4747.

### Step 2: Data segmentation

In this work, the dataset was segmented into subsets due to its heterogeneity. These subsets were built considering two aspects from the total raw dataset; the number of observations and the number of parameters. In an effort to account for as much information as possible, the heterogeneous dataset was partitioned into 5 subsets: S1, S2, S3, S4 and S5. Figure 4.3 and Table 4.1 show a detailed description of these subgroups (S1-S5).

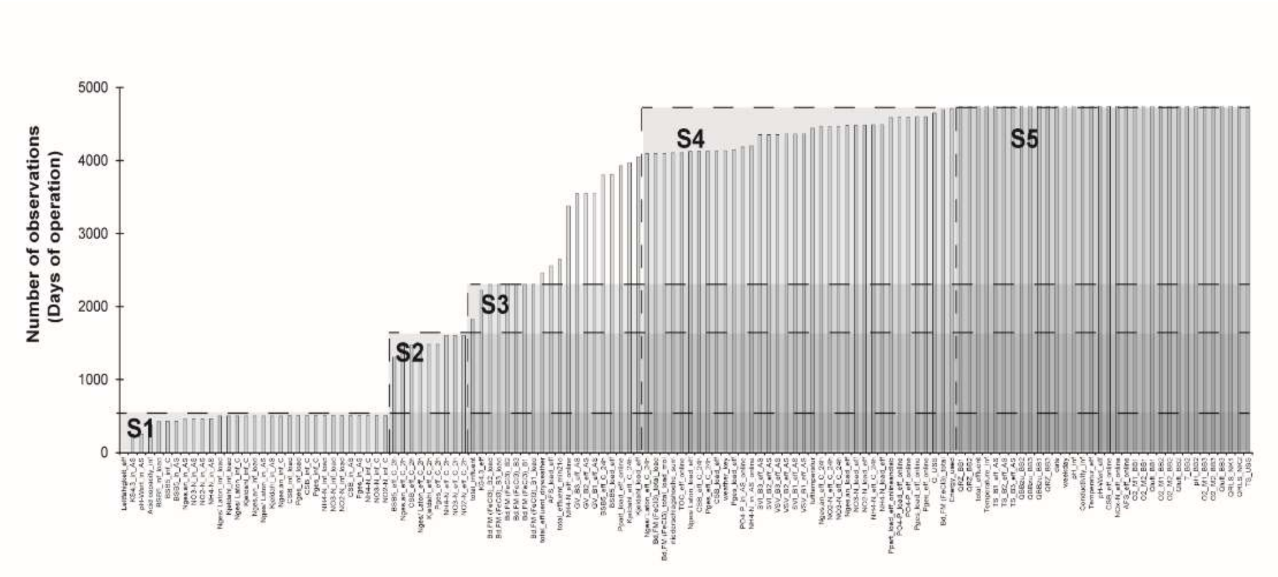


Figure 4. 3 Partitioning of data into subsets S1 to S5.

S1 covers the largest amount of parameters and the lowest amount of observations. From S2 to S5, the number of parameters decreases while the amount of observations increases (see Table 4.1).

Table 4. 1 Structure and partitioning of data for analysis of the 13 years of operation of a municipal WWTP

| Subgroup | No. observations | No. parameters |
|---|---|---|
| **S1** | 506 | 129 |
| **S2** | 1605 | 100 |
| **S3** | 2305 | 91 |
| **S4** | 4747 | 70 |
| **S5** | 4747 | 35 |

### Step 3: pre-processing

In each subset the missing values were identified, and multiple imputation was selected as imputation method. The amount of missing values in each subset was lower than 10% of the total amount of observations and was only necessary in subsets S1 to S4. After the segmentation of the raw dataset, due to the nature of the missing values in each subset (MAR mechanism), the multiple imputation would lead to unbiased estimates. Multiple imputation for multivariate missing data was implemented, each incomplete feature (in each subset) was imputed by a separate model, in this work the model selected for imputation was CART, this method generated estimates which resulted in low bias compared to the original dataset.

The subsets were then scaled with the minimum and maximum of each feature according to Equation 4.1. The min-max scaling method is implemented in the *scikit-learn* library in Python.

$$Var_{i,j_{sc}} = \frac{Var_{i,j} - \min(Var_{i,j})}{\max(Var_{i,j}) - \min(Var_{i,j})} \qquad \text{Equation 4. 1}$$

**Step 4: Data analysis**

After imputation, the subsets were partitioned into winter and summer seasons. The cold season or winter season was considered from the first day of October until the first day of May from the coming year, according to regulations this period is identified as cold season when nitrite emission in the effluent of wastewater is flexible. Whereas the summer season was considered from the 21st of June until the end of September each year. The partition into winter and summer seasons was done in all subsets. Afterwards, each subgroup was scaled and analyzed according to the workflow illustrated in Figure 4.4. In total, 15 subsets were analyzed; 5 for winter, 5 for summer and 5 for both seasons.

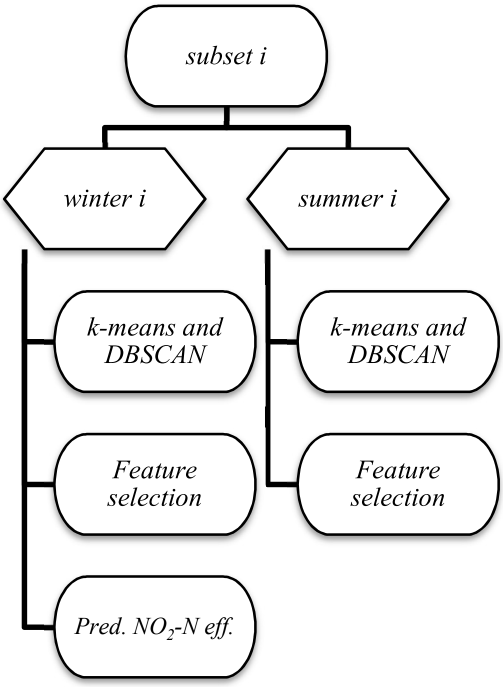


Figure 4. 4 Analysis workflow for each subset

Two hard-clustering methods were evaluated for the analysis of the subsets; k-means and DBSCAN. Since it is a multidimensional system, the shape of the clusters was unknown, thus both methods were applied. As previously described in Chapter 3, section 3.5, k-means clustering is the simplest implementation of the principle of maximum internal cohesion and maximum separation. Additionally, DBSCAN, a density based clustering method was also tested due to its flexibility when dealing with an arbitrary shape of clusters and distinguish noise. The results of both methods are compared and analyzed. To further study the clustering results, correlation matrixes were built to extract strong correlations, in this work correlations higher than $|0.9|$ (R=$|0.9|$) were extracted. A correlation matrix is a symmetric matrix, with the diagonal representing the correlation of each parameter with each other, therefore the diagonal has a value of 1.

After clustering analysis, feature selection was applied to study and extract strong interactions between nitrite related parameters and influent and process parameters in the WWTP. The method used for feature selection was recursive feature elimination (RFE), in this method, given an external estimator that assigns weights to features (e.g., importance ranking in random forest), RFE refers to the process of selection of feature by recursively considering smaller and smaller set of features. The estimator is trained on the initial set of features and the importance of each feature is obtained through feature importance. Then, the least important features are pruned from current set of features. The procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. The results of both winter and summer seasons are discussed. Finally, based on the feature selection process, the accumulation of nitrite in the effluent of the WWTP in winter seasons was predicted throughout the years. Subsets S4 was used for the prediction tasks, this subset contains complete observations for daily values along the period studied. Different methods were evaluated for the prediction of nitrite during the winter seasons; SVM, gradient boosting ensemble, deep neural networks and random forest, from which random forest delivered the best results. The random forest was composed of 500 trees and to prevent overfitting, the amount of leaf nodes was limited to 30.

## 4.3 Results and discussion

This work aims at evaluating nitrite accumulation during winter seasons from different configurations of parameters and observations with different data-driven methods. To find patterns, two clustering methods were evaluated; k-means (a classic hard clustering approach) and DBSCAN (a density based clustering approach). The optimal amount of clusters in k-means was found through the silhouette score (Rousseeuw, 1987), whereas in DBSCAN, the *eps* parameter was found through the K-nearest neighbors method (Altman, 1992) (see Appendix, Figures A1-A6). Figures 4.5-4.7 show the results obtained from the clustering analysis, each figure contains the results for all subsets and seasons studied. Since each subgroup contains multiple features (multidimensional datasets), the clusters are illustrated using PCA, in terms of the first two principal components; PC1 (x-axes) and PC2 (y-axes).

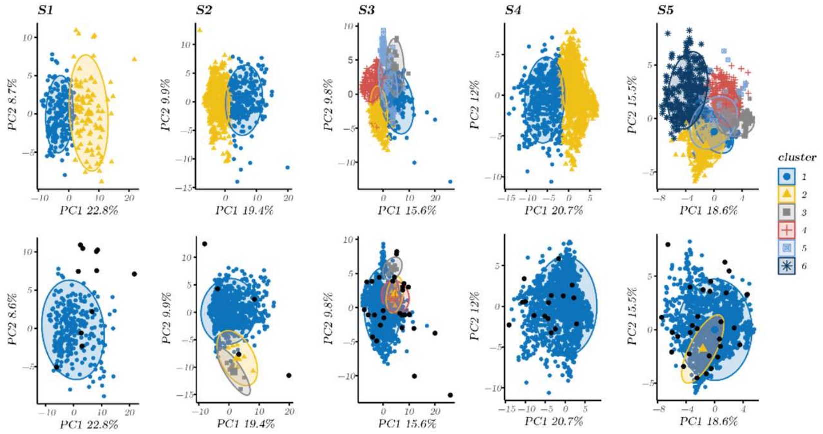


Figure 4. 5 Clustering analysis results from k-means (top) and DBSCAN (bottom) during winter for all subgroups

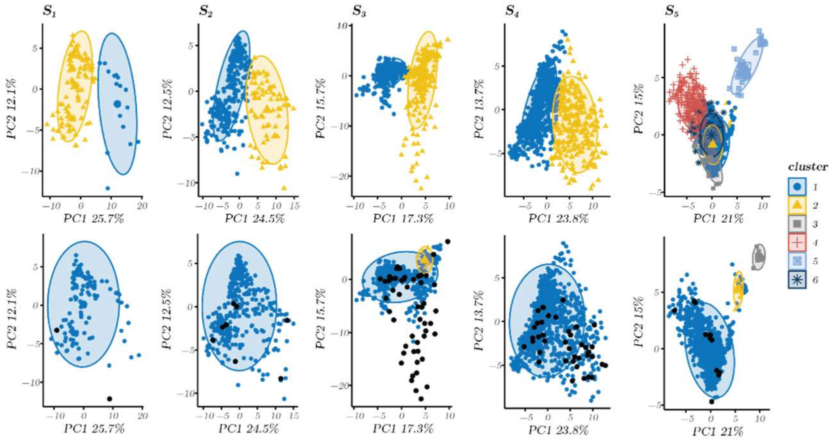


Figure 4. 6 Clustering analysis results from k-means (top) and DBSCAN (bottom) during summer for all subsets

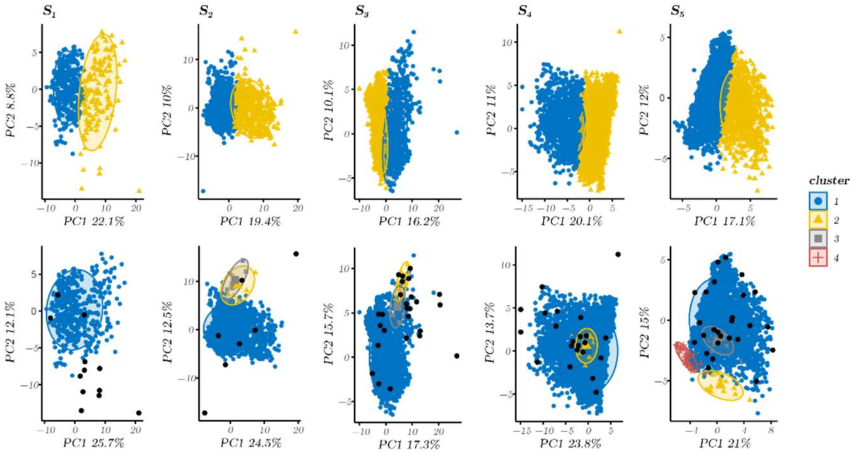


Figure 4. 7 Clustering analysis results from k-means (top) and DBSCAN (bottom) during all year for all subsets

From this first set of figures, the results show that the amount of clusters found with each clustering method differs from season to season (winter, summer and all seasons) and from subset to subset (S1 to S5). With DBSCAN, where the number of clusters is determined by the *eps* value, the number of clusters was mostly only one. The results show that, no relevant patterns can be extracted with DBSCAN (only one cluster found in most of the subsets), thus, for this particular study, a density based clustering method, such as DBSCAN, is not suitable for finding patterns.

On the other hand, in k-means clustering, the optimal number of clusters in the majority of the subsets and seasons were two; in Figure 4.7, in subsets S1 to S4 in Figure 4.6 and subsets S1,S2 and S4 in Figure 4.5. With an increasing amount of observations, the two regions identified showed high cohesion and separation in most of the subsets and seasons.

Clearly, k-means clustering outperformed DBSCAN in finding patterns. To extract further associations, the results from k-means clustering were analyzed in winter seasons (Figure 4.5). For this purpose, highly correlated parameters were extracted from the clusters in Figure 4.5 (k-means). The main motive behind this *filtering* process was to optimize the visualization of the results and their understanding. Recalling the initial setup and goal of this chapter, the analysis in this work used 15 datasets simultaneously; with 5 datasets for each season of the 3 seasonal datasets. Additionally, a minimum of two datasets out of these 15 datasets were produced through clustering (each cluster), which resulted in more than 30 multidimensional datasets.

Particular to the winter season, a total of 17 multidimensional subsets were analyzed; one for each cluster in each subset (S1-S5). Figure 4.8 summarizes the highly correlated features extracted from k-means clustering analysis in winter seasons.

As previously stated in the motivation of this chapter, the main reason for the segmentation was to evaluate the impact of the number of observations and parameters on the quality of the results. Each list in Figure 4.8, reports the highly correlated parameters extracted from each cluster among subsets S1 to S5.

Few highly correlated parameters found in the subsets were common in all clusters. In S1, parameters related to phosphorus concentrations in the effluent, nitrogen concentrations in the influent, coagulant dosage, sludge volume index (SVI), suspended sludge volume (SSV) and lost on ignition fraction (LOI) in the effluent stream were common in both clusters. In S2, less parameters were highly correlated and common in both clusters, however the trend was similar to S1; phosphorus effluent concentrations, SSV in the effluent, and nitrogen concentration in the effluent appeared in both clusters. In S3, only total phosphorus load in the effluent was common in all clusters, whereas in S4, results were similar to S1 and S2; parameters related to SSV and nitrogen concentration in the effluent were common in both clusters in S4. In S5, completely different results were obtained, the only parameter common in all clusters was the influent flow.

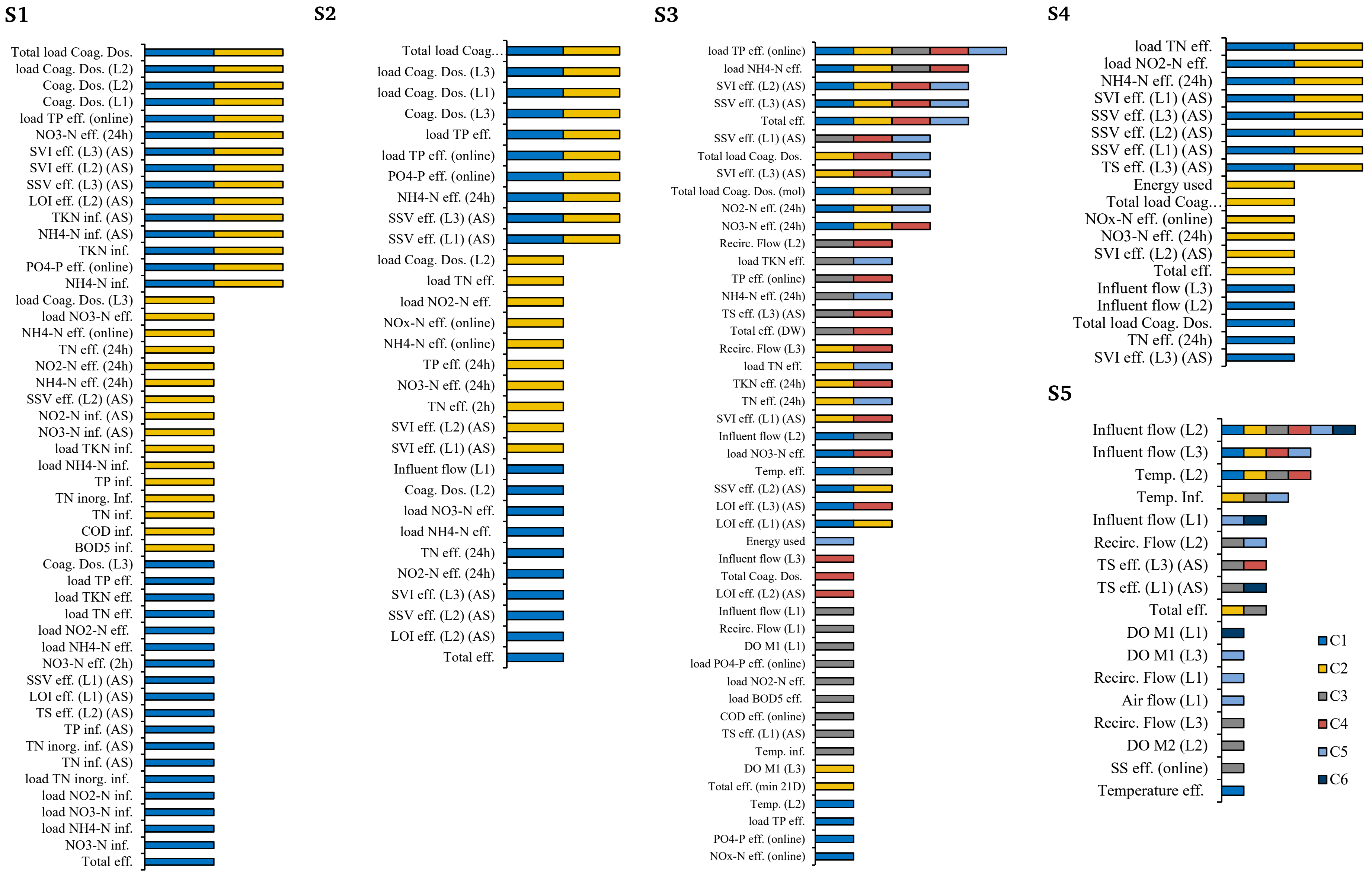


Figure 4. 8 Highly correlated features extracted from the clustering results through k-means in winter seasons. Ci=Cluster i

Figures 4.9-4.11 illustrate the correlation matrixes of the highly correlated parameters in all clusters and subsets for winter seasons.

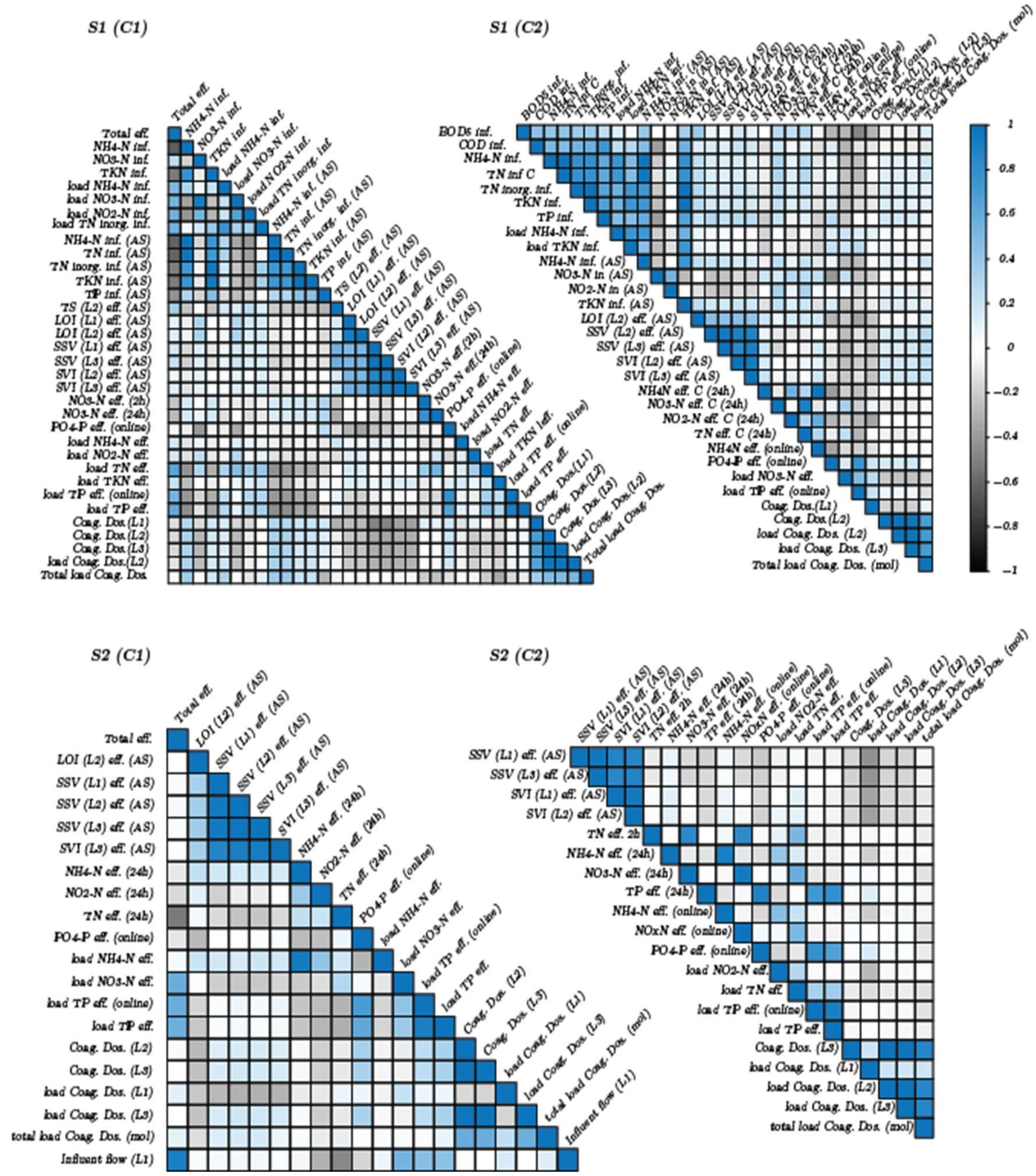


Figure 4. 9 Highly correlated feature in the clusters found through k-means for S1 and S2 for the winter seasons.

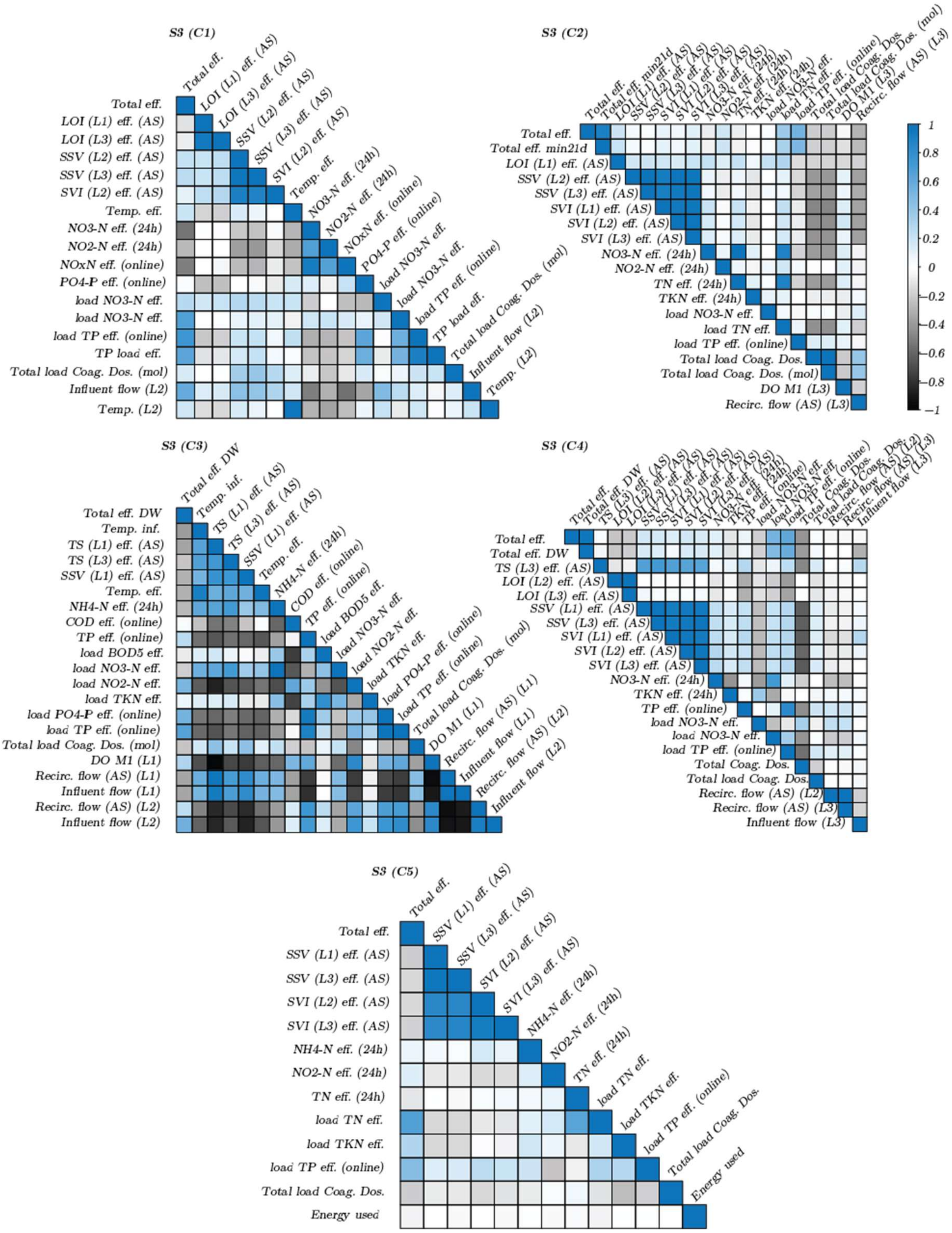


Figure 4. 10 Highly correlated feature in the clusters found through k-means for S3 for the winter seasons.

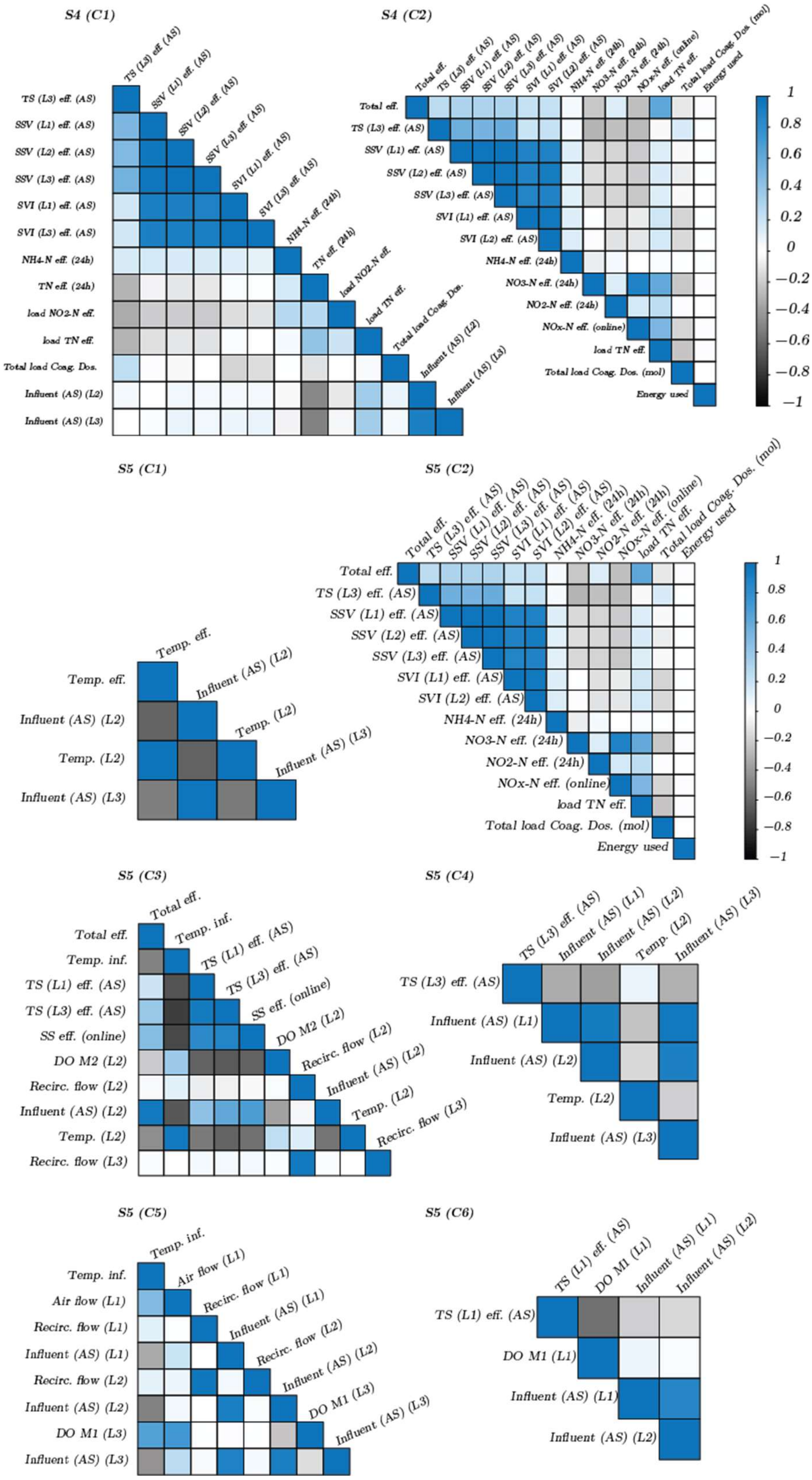


Figure 4. 11 Highly correlated feature in the clusters found through k-means for S4 and S5 for the winter seasons.

After analyzing the correlation matrixes, some interesting patterns were found in the winter season among the parameters and subsets. In S1-C1, the effluent flow rate (Total eff.) was highly opposite correlated to the nitrogen species ($NH_4$-N inf. and TKN inf.) and the total phosphorus concentration in the influent. Parameters related to the coagulant dosage were opposite correlated to the SSV, SVI and LOI, and highly correlated within each other. In S1-C2, high correlations were found related to $BOD_5$ and COD. These parameters were opposite correlated to total phosphorus influent concentrations. Coagulant dosage parameters were highly correlated within each other in both clusters in S1. In S2-C1, opposite correlation between total nitrogen and effluent and influent flow remained. In cluster 2, high correlations within SSV, SVI and LOI remained and opposite correlations to coagulant dosage were also found. Similar results were found in S3, in C1 once again opposite correlations between effluent flow and nitrogen species were identified. In S3, the most relevant correlations were found in cluster 3, in this cluster the load of nitrite in the effluent and the concentrations of TS were highly opposite correlated. High opposite correlations within DO, TS and recirculation flow were found in this cluster. SSV was highly correlated to the recirculation flow and phosphorus concentration was opposite correlated to the recirculation and influent flow. Temperature was opposite correlated to the recirculation flow and highly correlated to the NH4-N concentration in the effluent. Parameters related to the coagulant dosage were still opposite correlated to SSV, SVI and TS. In S5, the pattern with temperature and DO opposite correlation to effluent remained. Figures A.9-A.13 in the Appendix, illustrate the distribution of the aforementioned parameters in all subsets. Boxplots are used to illustrate the variability of the data from cluster to cluster, which is useful when comparing distributions between clusters. Some of the features such as total effluent flow and total nitrogen concentration in the effluent, presented the most variability within the clusters in all subsets. Both parameters are illustrated below (see Figure 4.12). The results show clearly that for most subsets, the total nitrogen load is lower for lower effluent flow values, whereas the opposite occurs in the other cluster.

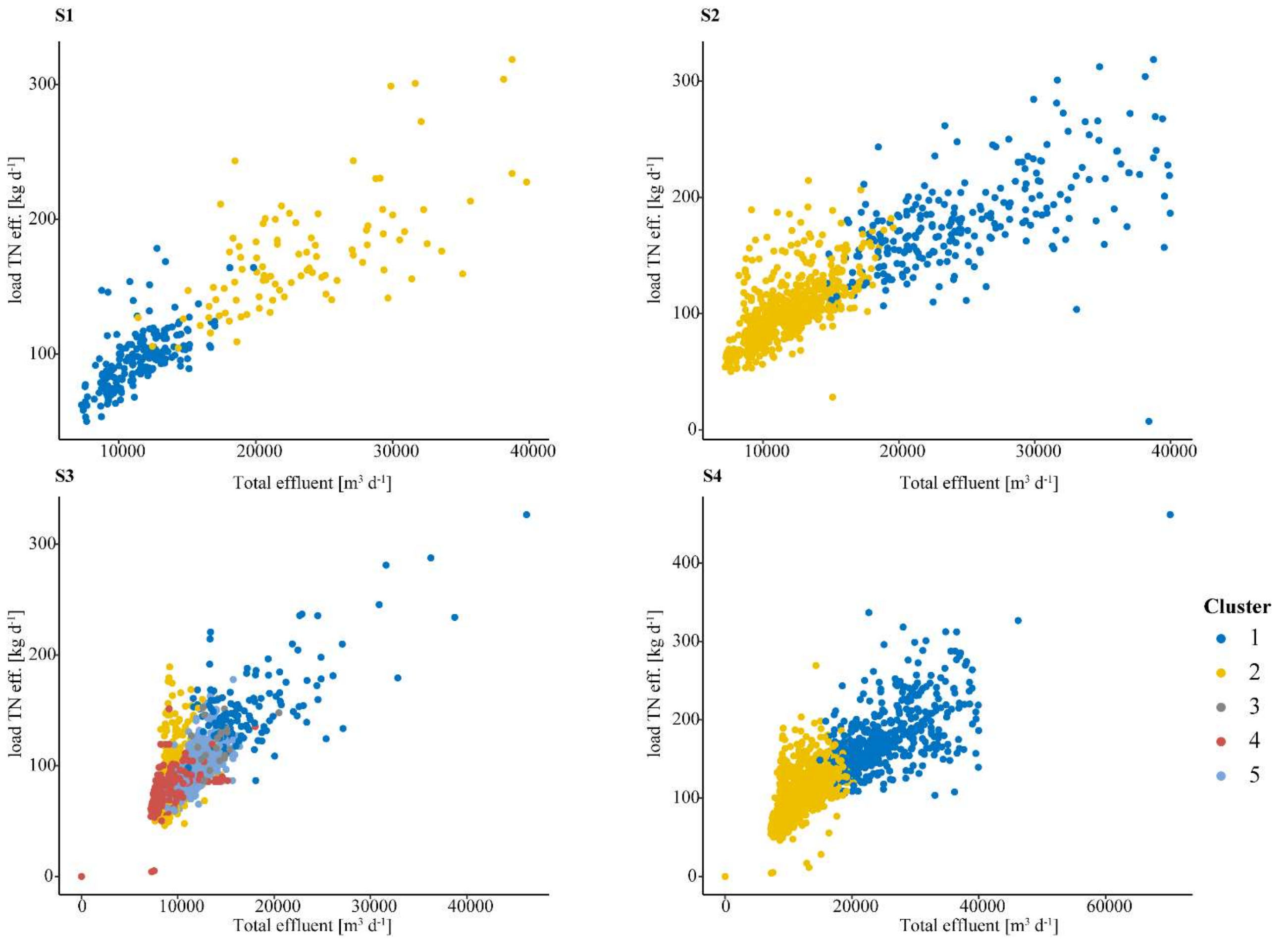


Figure 4. 12 Effluent TN load relation to effluent flow for subsets; S1 to S4 in winter season.

Clustering analysis allowed to extract patterns related to the winter seasons, highly correlated parameters were extracted and analyzed. The results showed clear associations between the effluent flow, sludge concentrations, temperature and nitrogen species, which were identified to be in general opposite correlated. Reasonably, higher influent/effluent flow resulted in lower concentration of pollutants such as nitrogen species and solids (high dilution), whereas the opposite with lower influent/effluent flow. Clearly, the outcome in each subset is very different, thus, showing the relevance of selecting a significant subset and its influence on the results interpretation. When we deal with heterogeneous datasets as the one studied in this chapter, the arbitrary selection of a dataset from the total raw data, would lead to a limited knowledge extraction, the results obtained from clustering analysis clearly support this statement.

After clustering analysis, feature selection was performed to extract strong interactions between influent and process parameters and nitrite concentration in the effluent. The main advantage of feature selection compared to correlation matrixes for extraction of knowledge, is that associations not necessarily highly correlated are extracted. Figure 4.13 shows the results obtained from feature selection for each subset described in Table 4.1. RFE was applied to study the interactions between influent/process parameters and nitrite concentration in the effluent. The feature selection process was applied to all subsets and the most relevant parameters were extracted. The results are summarized in Figure 4.13.

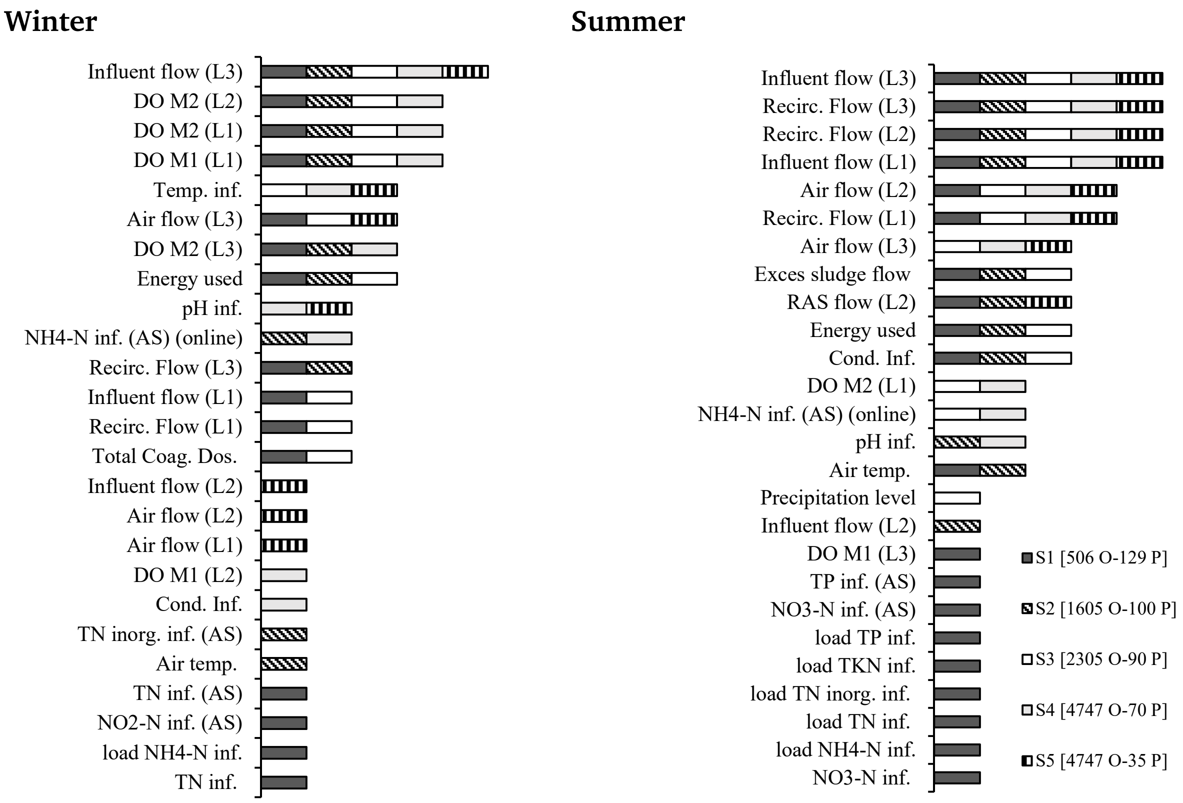


Figure 4. 13 Relevant parameters extracted from recursive feature elimination. Parameters which share strong interactions with nitrite concentration in the effluent are shown for both, summer and winter seasons. (AS: Activated sludge; Li: Line of AS). S1 to S5 are the subsets extracted from the segmentation of the heterogeneous datasets O: Observations and P:Parameters.

Figure 4.13 (left) shows that regardless of the amount of observations; influent flow, DO, temperature and air flow in L3, prevail in at least 3 out of the 5 subsets from the winter season data, from which, overall, 26 parameters were identified as relevant parameters. Nonetheless, the results from RFE highly differ from subset to subset, which is evident as well in the results obtained from the summer season data. In winter season for example, nitrite, ammonium and total nitrogen concentrations in the influent appear only in S1, whereas air temperature and total inorganic nitrogen concentration in the influent appear only in S2. DO concentration and conductivity of the influent appear only in S4 and air flow from L1 and L2 appear only in S5. In activated sludge processes, all the mentioned parameters are relevant in the production of nitrite in the N-DN processes. Any expert in the field, would find rare not to find these parameters within the analysis. Similar events occur in summer season (see Figure 4.13), at least 9 parameters appear only in S1 (bottom of the list in Figure 4.13), influent flow appears inly in S2 and precipitation level only in S3.

Similar to clustering analysis, feature selection is sensitive to the amount of data. Omitting any subset in the analysis, would lead to markedly different results. Clearly, no subset covers all the features extracted from RFE in winter or summer season.

The results from feature selection, clearly demonstrate that the data analysis of the initial raw heterogeneous dataset requires the right tools to extract information into actionable insights, in this chapter, the approach was the partition into subsets and further analysis.

Figure 4.14 illustrates all selected features in all subsets classified by season. When all seasons are compared, certain features are relevant during all seasons; air flow, influent flow, nitrate concentration in the effluent and ammonia concentrations in the influent. However, DO concentrations, pH, coagulant dosage, conductivity and temperature prevail as relevant parameters only during winter season while the return activated sludge and recirculation flow, excess sludge and precipitation level are more relevant in summer season. The results from feature selection clearly prove the influence of dataset selection to study the build-up of nitrite concentrations in different seasons.

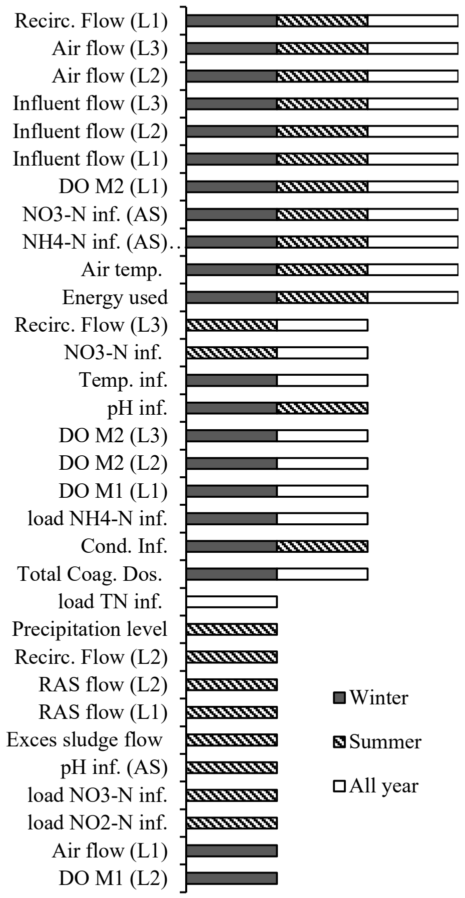


Figure 4. 14 Comparison of feature selection results throughout all seasons

The literature has not yet addressed the problem of combining different sources of data in wastewater treatment. However, the study of unbalanced datasets in different fields of engineering has been addressed by few studies (Bissonette, 1999; Kitchenham, 1998; Vannucci and Colla, 2018). Yet, methods to select or optimize the amount of data and features for analysis in heterogeneous datasets is unknown, or not yet explored in the water sector.

Subset S4 was applied for building predictive models to forecast nitrite accumulation during the winter seasons. The selected features from winter S4 in Figure 4.13 were used as predictors; influent flow (L3), DO concentrations in L1-L3 (DO M2 and M1), temperature, conductivity and pH in the influent flow, ammonia influent concentration from online sensors and conductivity in the influent. The winter S4 dataset was partitioned according to the scheme in Figure 4.15, k-fold cross-validation was applied for the dataset partition, a k value of 5 was used.

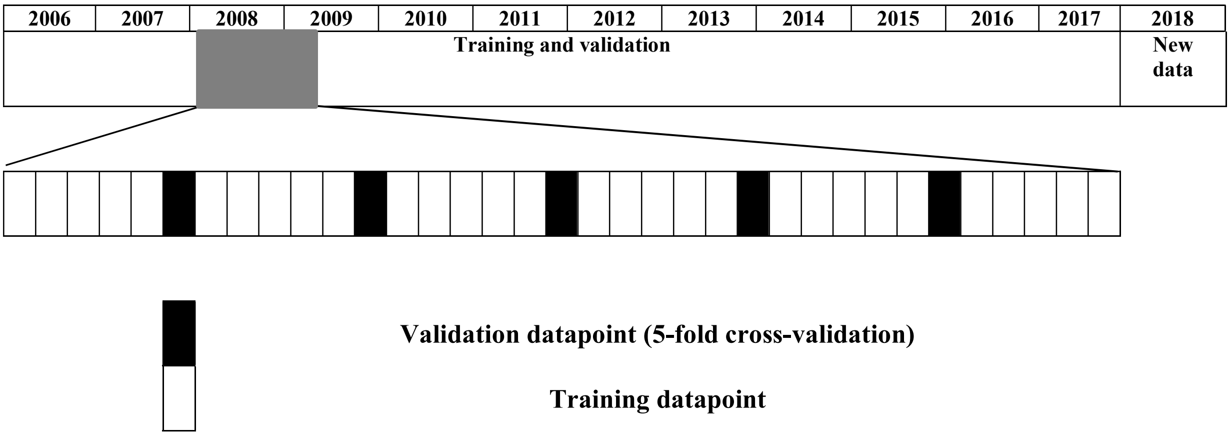


Figure 4. 15 Partition of winter season dataset. Training, validation and evaluation in new unseen data for nitrite accumulation in winter seasons.

The method which achieved the highest performance for the prediction of nitrite during the winter seasons was random forest. Figure 4.16 shows the results from the prediction of nitrite in winter seasons. The model was trained and validated with around 12 years of winter data, afterwards, the trained model was evaluated based on unseen data for the 13th year of operation. The results show that the model can achieve high performance in prediction. Table 4.2 shows the squared correlation coefficient between experimental and model data for the different datasets; training, validation and unseen new data, achieving values near to 1 for both nitrite concentration and load in the effluent.

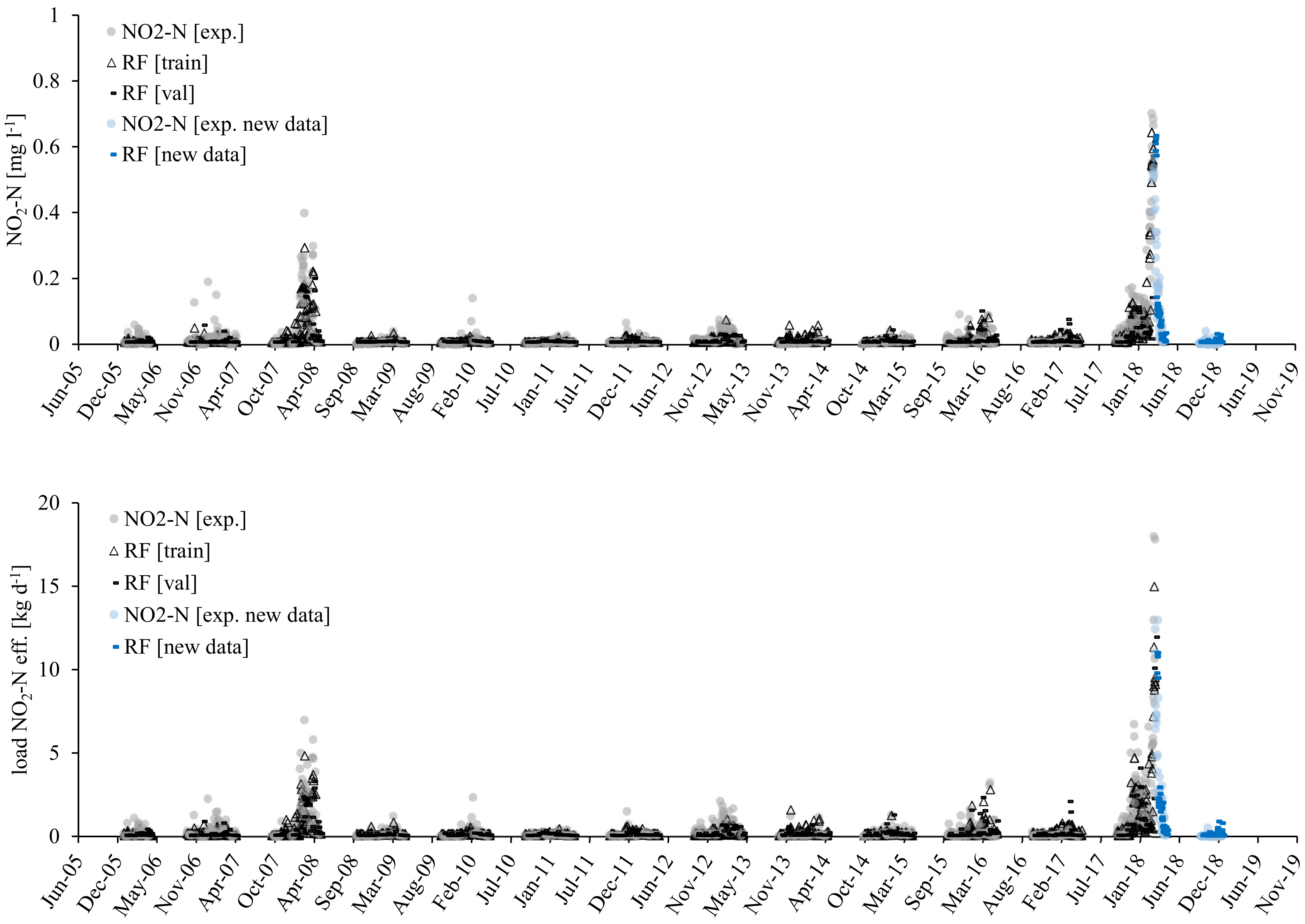


Figure 4. 16 Prediction of nitrite accumulation for the winter seasons over a decade. Top: nitrite concentration in the effluent. Bottom: nitrite load in the effluent.

| Squared correlation coefficient ($R^2$) | $NO_2$-N eff. | load $NO_2$-N eff. |
|---|---|---|
| Training | 0.96 | 0.96 |
| Validation | 0.81 | 0.83 |
| New unseen data | 0.82 | 0.81 |

Table 4. 2 Correlation coefficient between experimental and model results for nitrite accumulation in winter season

The results obtained after prediction demonstrates the opportunity of data-driven methods to selectively predict nitrite accumulation during winter seasons. The selection of relevant parameters through feature selection allowed to reduce the dimensionality of the dataset into a small subset of parameters to accurately predict the effluent concentration and load of nitrite. Two main benefits are identified compared to traditional ASM models; i) a lower amount of input data was required to build a highly accurate model (input parameters are limited to 10; (Figure 4.17) ii) calibration of the model depended mainly on the data availability (for training and validation) rather than kinetics. Although the amount of time to train the random forest model was lower than few minutes, the time invested in the pre-processing, feature selection, calibration, evaluation of different methods and knowledge extractions can certainly be comparable to an ASM approach.

## 4.4 Chapter conclusions

Quality of data is a main issue in the field of data science. The quality of data for data analysis purposes in wastewater treatment is not studied yet. In this work, an attempt to show the influence of data quality was made by analyzing a heterogeneous dataset of a WWTP in Germany. The main goal was to extract factors that influenced nitrite accumulation during winter seasons, to find patterns related to the build-up of nitrite in winter seasons and at the same time, show the impact of dataset selection when analyzing this phenomenon. Clustering analysis results showed that influent and effluent flow were clustered in two different clusters. High values of flow during winter seasons highly influenced the concentrations of the nitrogen species in the effluent. The clustering results were complemented with the feature selection results, which demonstrated once again the influence of data selection in the outcome of the analysis. Relevant features were extracted in feature selection; influent flow, DO concentrations, temperature, pH, conductivity and ammonia concentration in the influent. These parameters were at the same time used as predictors for a random forest model which was trained and validated with 12 years of operational data to predict nitrite accumulation in winter seasons. The accumulation during the 13th year of operation was successfully predicted with the validated model. Based on the results obtained in this chapter, and showing how the amount of observations influence the outcome of different data-driven tasks, a clear question arises from this chapter; how to select a significant dataset? The data analysis developed in this chapter was time consuming, thus more efficient ways to select a significant dataset out of the raw initial dataset should be found. This question is addressed in the next chapter, where two full-scale reactors were analyzed.

# 5 Feature selection in heterogeneous datasets and intelligent prediction methods

## 5.1 Introduction

The coupled process of PN-A is a well-studied process for biological nitrogen removal from wastewaters with high concentrations of ammonia ($NH_4^+$) and low concentrations of organic matter (COD) (Lackner et al., 2014). This process combines aerobic and anaerobic ammonia oxidizing bacteria (AOB and Anammox, respectively), both having an autotrophic lifestyle. The low oxygen and carbon requirements of PN-A processes favor them compared to N-DN (saving energy by up to 63% reduced oxygen demand and 100% in carbon supplement) (Fux and Siegrist, 2003; Lackner et al., 2015). Full-scale PN-A processes have been implemented in more than 100 full-scale installations worldwide already by 2014, most of these systems being SBRs and operated as single-stage. Modeling PN-A processes as well as other biological processes is an inherent part of the design and operation of wastewater treatment systems; it allows to anticipate the representation of the response to variations in wastewater quality and quantity, and process parameters in a feasible and cost-effective way (Wu et al., 2016). The ASM1 and ASM3 with the addition of Anammox kinetics and mass balances are commonly used for modeling PN-A processes, also studied in Chapter 2. However, ASM type models require in depth knowledge of the microbial processes and their kinetics, which leads to complex process models. This makes the optimization procedure highly time-consuming and dependent on the information available for model calibration (Hreiz et al., 2015). As a result, some studies have turned to simplified models, which are, however, probably less accurate than the ASM models (Kim et al., 2001).

As previously seen in Chapters 3 and 4, any WWTP generates large amounts of data on a daily base from monitoring the quality of the water and removal of pollutants in the wastewater. These sources belong to online from sensors, off-line from laboratories, among others. The results obtained in Chapter 4 showed that heterogeneous datasets obtained from biological processes require the right tools for efficient data analysis, and in order to extract representative information, arduous analysis is needed, which is time-consuming and not efficient. Which leads to a key unsolved problem in Chapter 4; the selection of a significant dataset from the total raw data. Accordingly, in this chapter, a procedure to select a significant dataset out of a heterogeneous initial raw data is proposed; to find the best configuration of features and amount of data (or observations). To evaluate the selection procedure, the resulting dataset was applied to create data-driven models and predict the effluent composition of two full-scale PN-A systems. Additionally, different case-scenarios were evaluated to elucidate the importance on selecting an optimal configuration of features and observations.

## 5.2 Methodology

### 5.2.1 Description of processes and datasets

Two full-scale SBRs ($SBR_A$ and $SBR_B$) were operated at a municipal WWTP (size 275.000 population equivalents) to treat reject water after anaerobic digestion. Both SBRs had a volume of 550 $m^3$, the ammonium influent concentrations were 960 ± 110 mg-N $L^{-1}$, soluble COD concentrations 320 ± 50 mg-$O_2$ $L^{-1}$. The reactors were operated with four six-hour cycles per day. A detailed description of the cycle and operation of the two SBRs can be found in Lackner et al., (2015).

The two full-scale SBRs were newly started up as PN-A reactors and were operated and monitored for approx. 600 days. The nitrogen turnover in the first 200 days was low with values below 0.1 kg-N $m^{-3}$ $d^{-1}$. With the increase in biomass concentration, establishing a sludge wastage regime, and an optimized aeration regime the nitrogen turnover reached values above 0.2 kg-N $m^{-3}$ $d^{-1}$. Nitrogen turnover and effluent concentrations were highly dynamic due to the many operational adjustments and optimization efforts. $SBR_A$ had effluent ammonium concentrations ranging from as low as 60 mg-N $L^{-1}$ up to as high as 300 mg-N $L^{-1}$, including an ammonium peak of 600 mg-N $L^{-1}$ around day 215. The ammonium concentrations in $SBR_B$ were in a similar range, without any drastic peak (maximum value of 400 – 450 mg-N $L^{-1}$ within the first 200 days of operation). Nitrate concentrations started to increase on day 110 in $SBR_A$ and reached concentrations of 250 mg-N $L^{-1}$ at the maximum. Nitrate levels in $SBR_B$ were generally higher esp. in the first 100 days, but lay in similar ranges afterwards.

The aeration pattern had a major influence on the performance of the SBRs and aeration intervals of 6 min with at least a break of 9 min seemed to be optimal for the recovery and stabilization of the reactor operation (Lackner et al., 2015).

Both systems were monitored by online sensors for ammonium, nitrate, oxygen, conductivity, pH, oxidation reduction potential (ORP) and temperature. Lab analyses were performed for additional parameters (COD, ammonium, nitrate, nitrite, solids) in the influent and effluent from one to a couple of times per week. Table A.3 summarizes basic statistical information from the variables considered in this study with their respective amount of observations. Figure 5.1 illustrates the amount of present and missing values for the parameters studied in both reactors; $SBR_A$ and $SBR_B$.

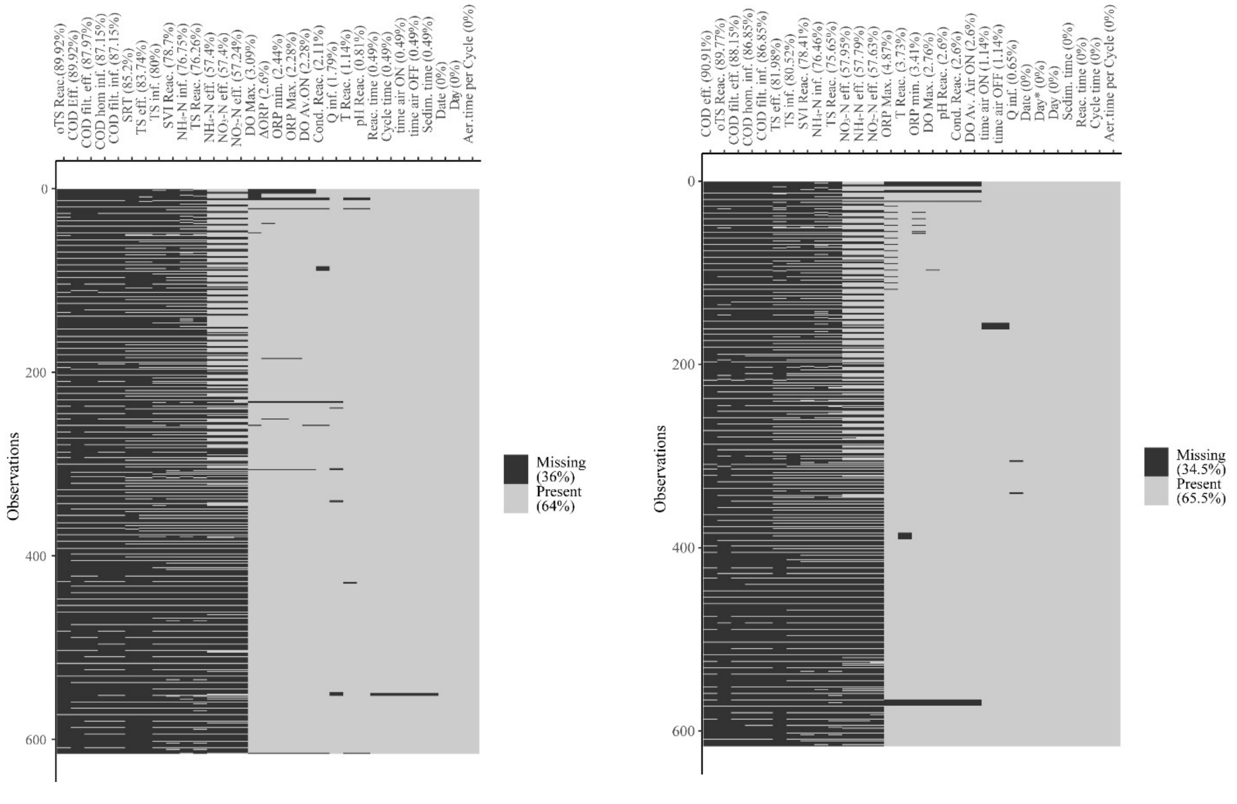


Figure 5. 1 Visualization of present and missing data for both reactors SBRA (left) and SBRB (right). The target features are in bold; $NH_4$-N eff. and $NO_3$-N eff.

The process variables (or features) domain was different from feature to feature, belonging to different orders of magnitude: some of them ranged from 0.4-7.6 mg-N $L^{-1}$ (*$NO_2$-N Eff.*), whereas others ranged from 47-3600 mg-$O_2$ $L^{-1}$ (*COD Inf.*). Therefore, before any further analysis, the raw data was scaled. All the features analyzed in the SBR's datasets were scaled with the statistical mean and standard deviation according to Eq. 3.1 in Chapter 3.

### 5.2.2 Feature selection mapping features and number of observations

A key issue with the SBR's datasets is their heterogeneity; high number of features, each feature containing different amount of observations i.e. the number of observations (i.e., tuples) varied from feature to feature (Figure 5.1). Similar to Chapter 4, the main issue was finding a representative dataset that would best describe the process. To find the optimal configuration of features and observations, the datasets of both SBRs were first partitioned into subsets based on the amount of missing values. Table 4.3 summarizes the subsets and number of observations in each subset. Similar to Chapter 4, S1 contains the largest amount of features and lowest amount of observations and S7 contains the lowest amount of features and the highest amount of observations.

Table 5. 1 Structure and partitioning of data for analysis of both SBRs

| | **SBR$_A$** | | **SBR$_B$** | |
|---|---|---|---|---|
| **Subset** | **No. Features** | **No. Observations** | **No. Features** | **No. Observations** |
| **S1** | 27 | 41 | 27 | 33 |
| **S2** | 26 | 55 | 26 | 44 |
| **S3** | 25 | 59 | 25 | 60 |
| **S4** | 24 | 91 | 24 | 98 |
| **S5** | 23 | 124 | 23 | 130 |
| **S6** | 22 | 136 | 22 | 142 |
| **S7** | 21 | 261 | 21 | 258 |

Feature selection was performed in each subset through recursive feature elimination (RFE). The selected features in each subgroup were pondered with a *score function* which evaluated two conditions; a reasonable amount of observations to train, and the existence of a feature in at least 70% of the subsets after feature selection. At the end, the optimized dataset would contain the best configuration of features and observations. Figure 5.1 summarizes the steps followed for partitioning the datasets and the selection of features.

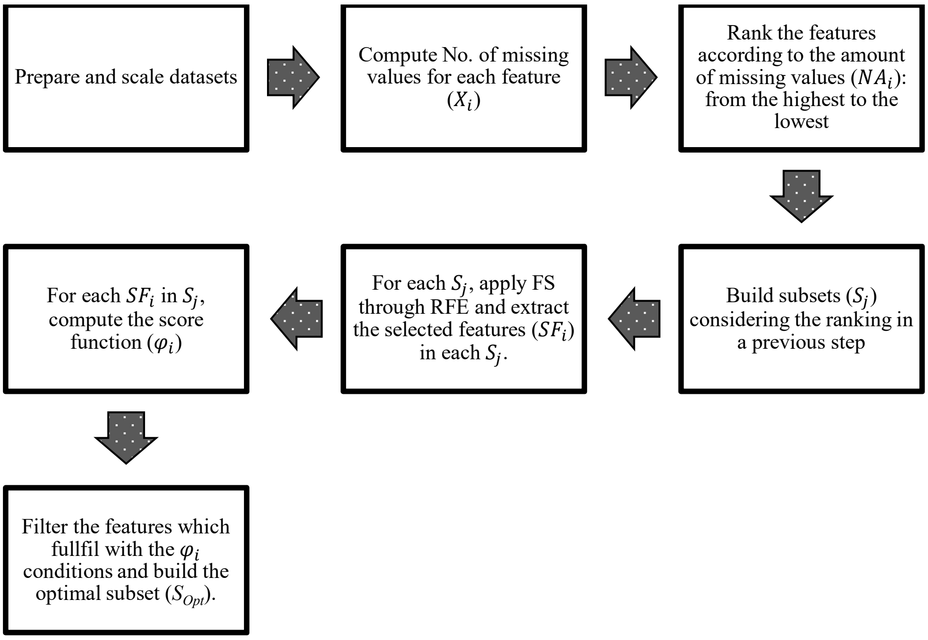

Prepare and scale datasets
Compute No. of missing values for each feature ($X_i$)
Rank the features according to the amount of missing values ($NA_i$): from the highest to the lowest
Build subsets ($S_j$) considering the ranking in a previous step
For each $S_j$, apply FS through RFE and extract the selected features ($SF_i$) in each $S_j$.
For each $SF_i$ in $S_j$, compute the score function ($\varphi_i$)
Filter the features which fullfil with the $\varphi_i$ conditions and build the optimal subset ($S_{Opt}$).

Figure 5. 2 Steps for finding the best dataset configuration.

A a score ($\alpha_j$) was defined for each subset $S_j$, and a score function for each selected feature $SF_i$ ($\varphi_i$), with $i = \{1, \ldots, m\}$ and $j = \{1, \ldots, n\}$, $n$ and $m$ are the maximum number of subsets ($S_j$) and selected features ($SF_i$), respectively. The value of $\alpha_j$ increases with increasing amount of observations. The score function ($\varphi_i$) and its components are defined in Equation 5.1.

$$\varphi_i = \sum_{j=1}^{n} \alpha_j \, \tilde{\varphi}_i^j$$

where:

$$\alpha_j = f\left(\theta_{S_j}\right)$$

$$\tilde{\varphi}_i^j = \begin{cases} 1 & SF_i \in S_j \\ 0 & SF_i \notin S_j \end{cases}$$

Subject to: Equation 5. 1

$$\varphi_i\left(SF_i \in S_{opt}\right) > \min(\varphi_i)$$

where the $\min(\varphi_i)$:

$$\min \varphi_i = \sum_{j=1}^{k} \alpha_j \, \tilde{\varphi}_i^j$$

where $k = 0.7n$ such that:

$$SF_i \in S_{opt} \leftrightarrow SF_i \in 0.7n \quad \wedge \quad \varphi_i > min(\varphi_i \,|SF_i \in 0.7n)$$

The last statement establishes that $SF_i \in S_{opt}$, only if $SF_i$ exist in at least 70% of all subsets $(0.7n)$ and its score will be higher or equal to a threshold, $\min \varphi_i$. The threshold computes the hypothetical score of a $SF_i$ given that it exists in at least 70% of the subsets, being these subsets the ones with lowest scores. The main goal of the score function is to guarantee that the selected features would be significant of the dataset and at the same time, hold enough observations for further prediction tasks.

### 5.2.3 Building predictive models

This work contributes with an alternative approach to predict the effluent of these systems without the use of kinetic based mechanistic models. Due to the dynamic nature of the processes and the limited amount of data available for training, state of the art ML methods were evaluated to predict the effluent composition of both SBRs. The selection of the methods was based on demonstrated performance of these methods in prediction tasks in other complex bWWTP (Huang et al., 2009; Kusiak and Wei, 2014; Mjalli et al., 2007; Xie et al., 2017). Moreover, a previous work has demonstrated the feasibility of building accurate models with limited amount of data for training, where SVM outperformed ANN (Alejo et al., 2018). Accordingly; support vector machines (SVM), $\nu$ support vector machines ($\nu$-SVM), gradient boosting ensemble of trees (GB-Trees) and random forest, were evaluated for the prediction of the effluent composition of both SBRs. All models were built in Python 3.5 in the Anaconda environment, and the library *scikit-learn* was used. Training and validation datasets were built from $SBR_A$ and $SBR_B$. Furthermore, cross-validation via *k-fold* was used to select both training and test datasets for evaluating the predictive models. A $k$ value of 4 was selected; this partition allowed to build a model with low bias and moderate variance between the predicted and experimental values (Alejo et al., 2018; Fushiki, 2011). In this work, the number of the training and validation datasets highly depended on the feature selection process, therefore this will be addressed in the results section. The performance of the ML methods was evaluated in the training dataset. To prevent overfitting of the models, the hyperparameters in each method were adjusted. The gradient boosting ensemble consisted of 80 decision trees for both reactors, the depth of the trees was limited to two, to prevent overfitting. Two types of SVM were used, the regular SVM and $\nu$-SVM. Different kernel functions were evaluated, the kernel function that best fitted the data of $SBR_A$ was a radial basis function, in both SVM and $\nu$-SVM models, whereas a polynomial kernel function of degree 3 best fitted the data for $SBR_B$. The cost hyperparameter ($C$) was kept to values close to 1 for all SVM models to prevent overfitting. The random forest model was composed by 220 and 300 trees for $SBR_A$ and $SBR_B$, respectively. To prevent overfitting, the maximum node leafs was limited to 10.

### 5.2.4 Model evaluation

The accuracy of the models was evaluated through the Pearson correlation coefficient ($R^2$), mean squared error (MSE), root MSE (RMSE), mean absolute percentage error (MAPE) and mean average deviation (MAD) (James et al., 2013).

## 5.3 Results and discussion

As previously described (Figure 5.1), our approach aimed at finding the best setup for features and observations to build a significant dataset and use it to predict the effluent composition of both SBRs systems; ammonia ad nitrate.

The subsets were sorted according to the number of missing values they contained. The number of observations in each subset was different and increased gradually (the trend was: decreasing amount of variables with increasing amount of observations). Then, RFE was applied to all subsets and the best *predictors* in each dataset were obtained. The results are summarized in Figure 5.3

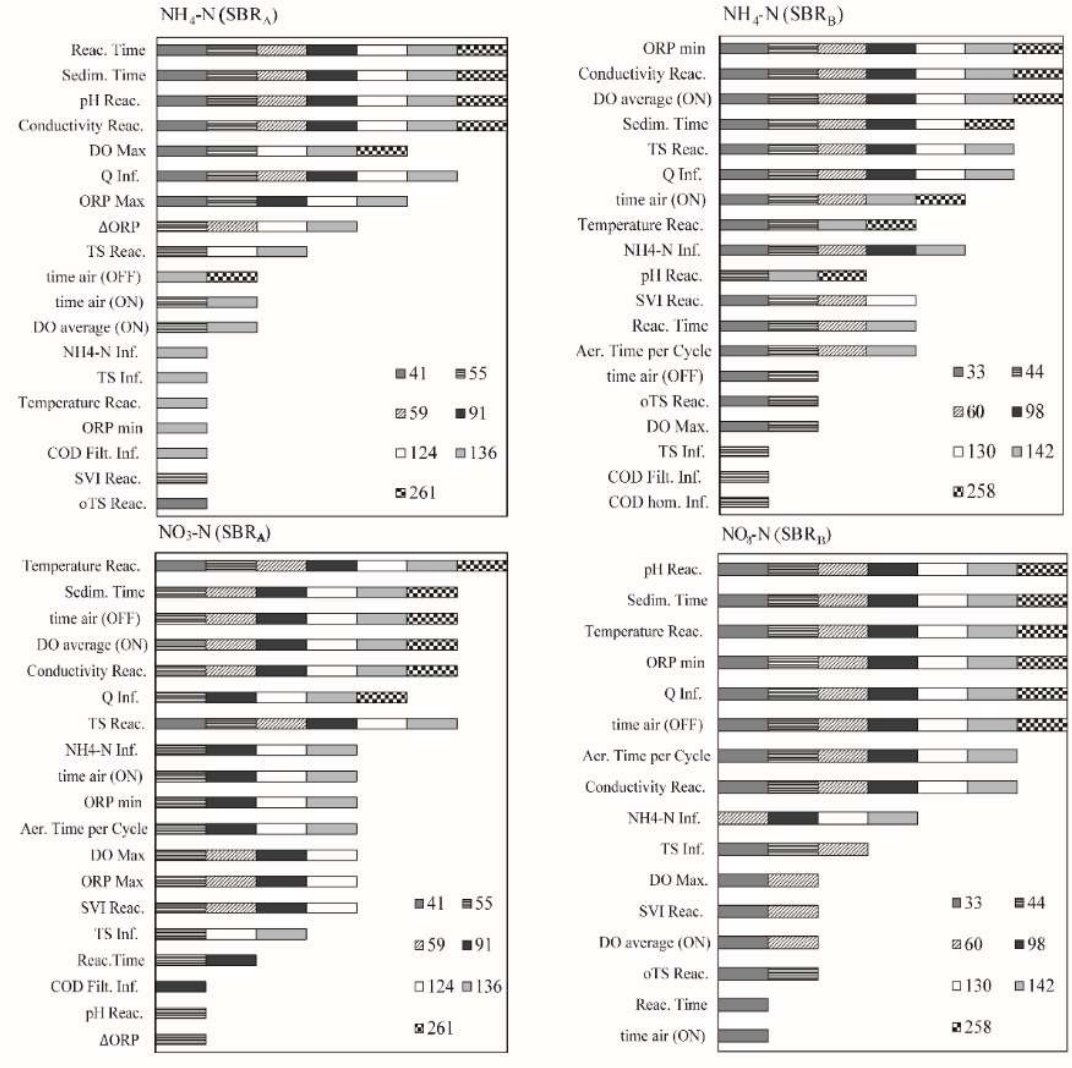


Figure 5. 3 RFE analysis results. For each system (SBRA and SBRB), seven subsets ($S_j$) were obtained (section 5.2.1). Subsets S1 to S7 contained; 41, 55, 59, 91, 124, 136 and 261 observations for SBRA, and: 33, 44, 60, 98, 130, 142 and 258 observations (or tuples) for SBRB, respectively.

In both systems (SBR$_A$ and SBR$_B$), the sedimentation time, conductivity, the volumetric influent (*Q inf.*), temperature and the time period when aeration was off (*time air OFF*), had great impact on the process performance (*$NH_4$-N Eff.* and *$NO_3$-N Eff.)*.

## 5.3.1 Best configuration in feature selection: mapping features and observations

The RFE delivered a list with the best (*prediction)* features for each subset. The optimal number of features for building predictive models and perform cluster analysis were then selected through the *score function* (*$\varphi_i$*) as seen in Figure 5.4.

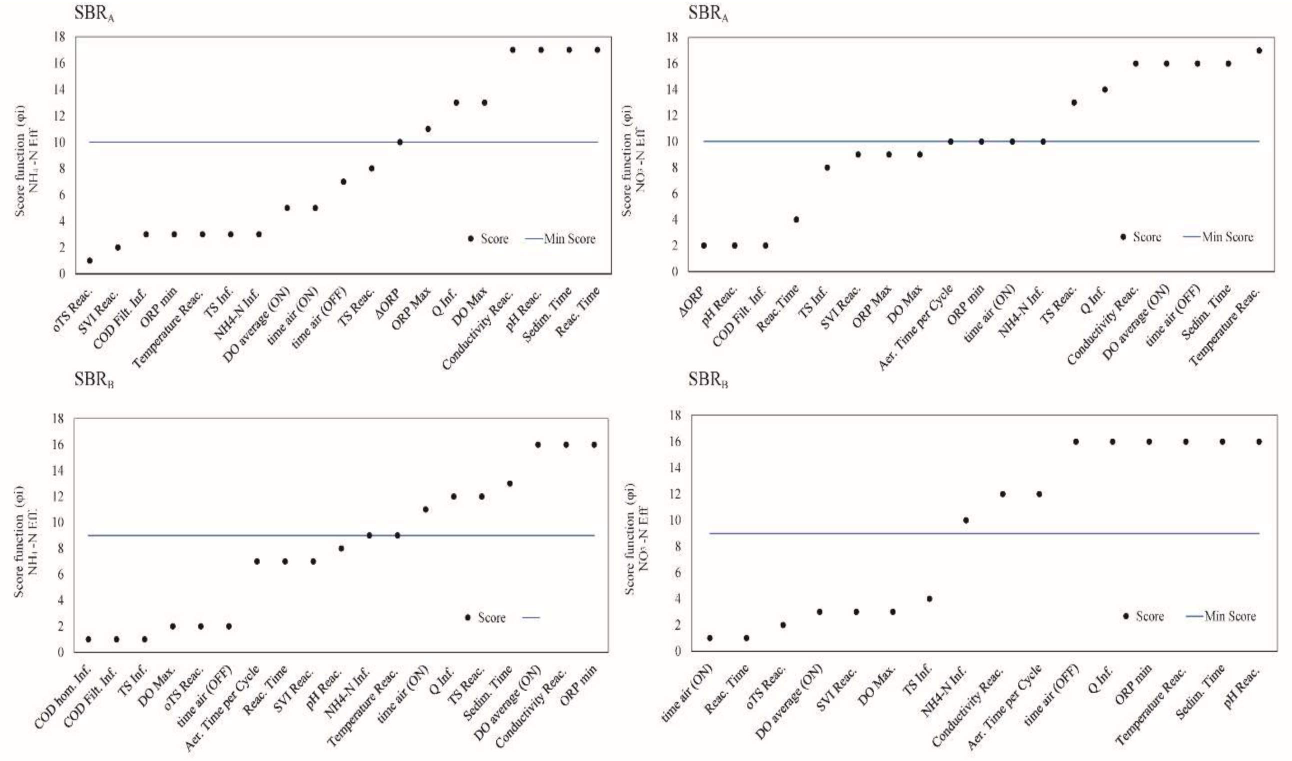


Figure 5. 4 Results obtained through the score function (φi) in the configuration of feature selection. Here, the score of each variable (●) and the minimum score (—) for each system (SBRA or SBRB) are plotted. Only the variables above the minimum score are considered as selected variables for building the predictive models.

The minimum score for all features was computed, and the subset of best features or *predictors* (SFi) were selected as those having a value above the minimum score or *threshold* (Figure 5.4). Table 5.2 summarizes the finally selected features as best *predictors* for further analysis.

Table 5. 2 Best configurations of features and number of observations for building predictive models. The X indicates if the feature is a predictor.

| Predictors | *$SBR_A$* | | | *$SBR_B$* | | |
|---|---|---|---|---|---|---|
| | **$NH_4$-N Eff.** | **$NO_3$-N Eff.** | **Predictor subset** | **$NH_4$-N Eff.** | **$NO_3$-N Eff.** | **Predictor subset** |
| Q Inf. | X | X | X | X | X | X |
| Sedim. Time | X | X | X | X | X | X |
| Conductivity Reac. | X | X | X | X | X | X |
| $NH_4$-N Inf.* | X | X | X | X | X | X |
| DO average (ON) | | X | X | X | | X |
| time air (OFF) | | X | X | | X | X |
| Temperature Reac. | | X | X | | X | X |
| pH Reac. | X | | X | | X | X |
| ORP min | | | | X | X | X |
| TS Reac. | | X | X | X | | X |
| ORP Max | X | | X | | | |
| DO Max | X | | X | | | |
| Reac. Time | X | | X | | | |
| time air (ON) | | | | X | | X |
| Aer. Time per Cycle | | | | | X | X |
| **Total Number of features** | **8** | **8** | **12** | **8** | **9** | **12** |

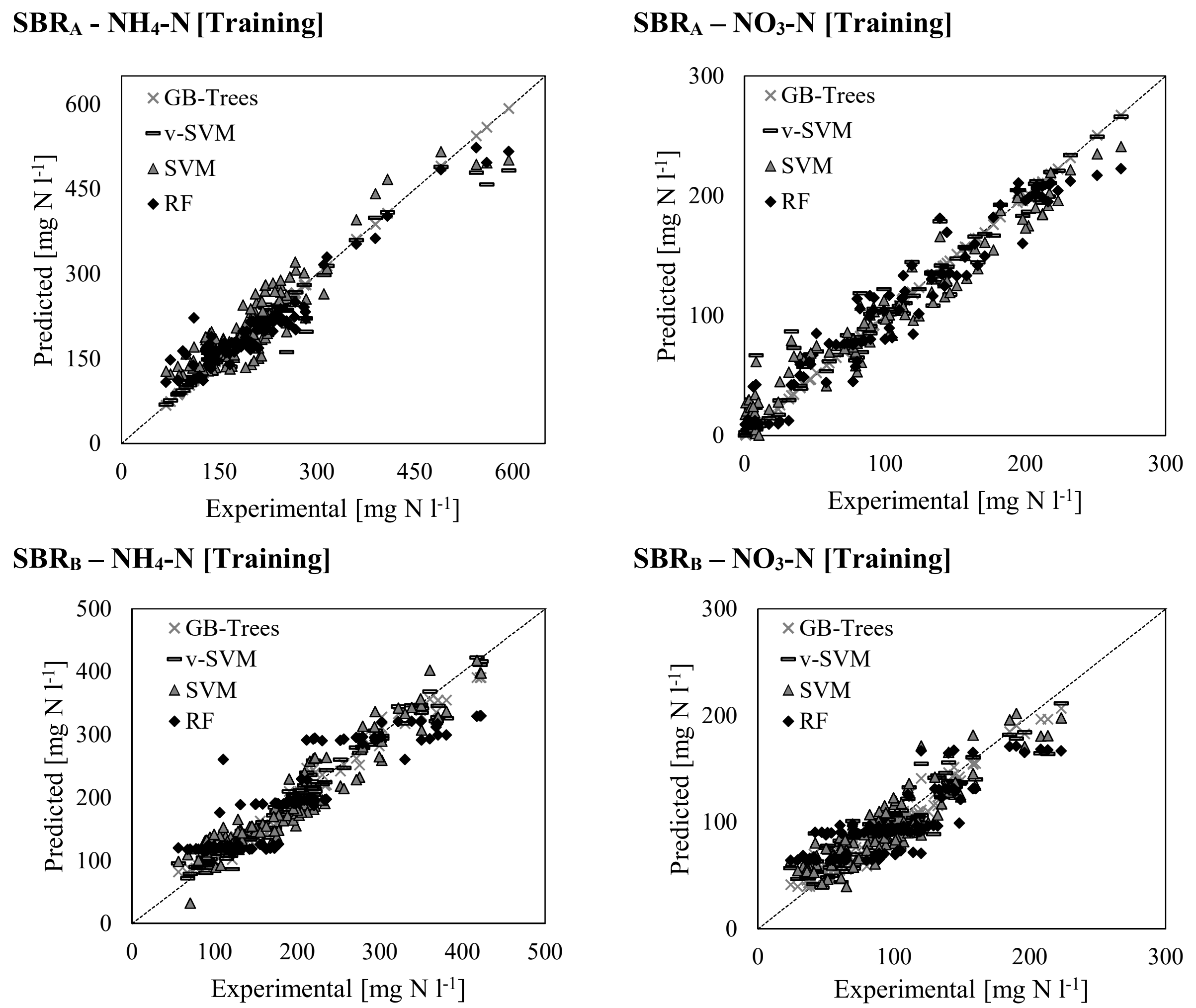
*Only this feature was added intentionally due to its *relevance* from an operation perspective; nitrogen concentration in the influent.

The features in Table 5.2 were combined to create the dataset to build the prediction models. The size of these two new subsets were 12 variables with 136 observations for the $SBR_A$ and 12 variables with 142 observations for $SBR_B$. The approximate resolution for the datasets was 4.5 days.

### 5.3.2 ML-based prediction of PN-A effluent

State-of-the-art supervised ML prediction methods were evaluated for the effluent prediction of two full-scale PN-A systems ($SBR_A$ and $SBR_B$). The techniques included; GB-Trees, SVM, $\upsilon$-SVM and random forest. The training and validation datasets were partitioned according to a *4-fold cross validation.* The size of these datasets were 34 and 36 observations for validation of $SBR_A$ and $SBR_B$, respectively. The remaining amount of data was applied for training purposes.
Figures 5.5 to 5.7 show the results of the prediction of the effluent for the training and validation datasets, respectively.

Figures 5.5 and 5.6, allows the comparison of the performance of the models evaluated for the prediction of $NH_4$-N and $NO_3$-N concentration (in mg N $l^{-1}$). Experimental concentration values are illustrated against the predicted values.



SBR$_A$ - NH$_4$-N [Training]

SBR$_A$ – NO$_3$-N [Training]

SBR$_B$ – NH$_4$-N [Training]

SBR$_B$ – NO$_3$-N [Training]

Figure 5. 5 Comparison of the performance of the models evaluated for the prediction of NH4-N and NO3-N concentration (in mg N l-1) in the training phase. Experimental values are illustrated against the predicted values.

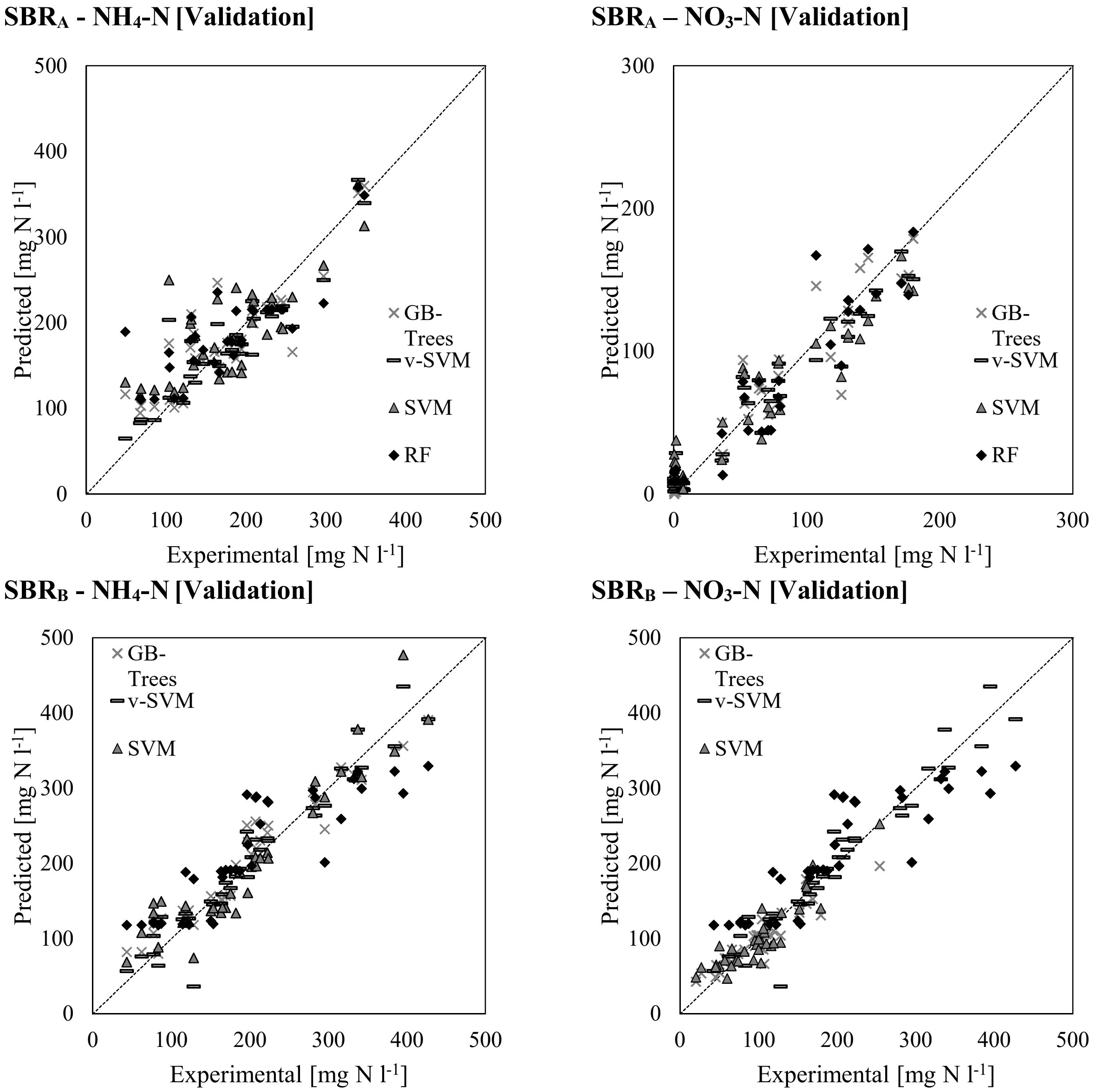


Figure 5. 6 Comparison of the performance of the models evaluated for the prediction of $NH_4$-N and $NO_3$-N concentration (in mg N $l^{-1}$) in the validation phase. Experimental values are illustrated against the predicted values.

Table 5.3 and 5.4 summarize the performance measurements for all models in the validation step for $SBR_A$ and $SBR_B$ reactors, respectively. From the evaluated models, *ν*-SVM and GB-Trees ensemble achieved the highest performance (highest correlation coefficient in both training and validation). However, the MAD, MSE, RMSE and MAPE for *ν*-SVM was considerably smaller than GB-Trees for both reactors.

Table 5. 3 $SBR_A$ Accuracy evaluation for both models through: mean average deviation (MAD), mean squared error (MSE), root MSE, mean absolute percentage error (MAPE) and squared correlation coefficient ($R^2$).

| Method | NH$_4$-N | | | | | |
|---|---|---|---|---|---|---|
| | MAD* [val] | MSE* [val] | RMSE* [val] | MAPE* [val] | $R^2$ [val] | $R^2$ [train] |
| **GB-Trees** | 27 | 1,328 | 36 | 21 | 0.87 | 0.99 |
| **nu-SVM** | 21 | 849 | 29 | 14 | 0.91 | 0.97 |
| **SVM** | 38 | 2,217 | 47 | 30 | 0.77 | 0.92 |
| **RF** | 31 | 1,821 | 43 | 28 | 0.82 | 0.96 |
| **Method** | **NO$_3$-N** | | | | | |
| | **MAD* [val]** | **MSE* [val]** | **RMSE* [val]** | **MAPE* [val]** | **$R^2$ [val]** | **$R^2$ [train]** |
| **GB-Trees** | 12 | 306 | 17 | 112 | 0.96 | 0.99 |
| **nu-SVM** | 12 | 236 | 15 | 345 | 0.97 | 0.98 |
| **SVM** | 19 | 469 | 22 | 956 | 0.95 | 0.97 |
| **RF** | 16 | 392 | 20 | 528 | 0.94 | 0.98 |

*in mg N $L^{-1}$

Table 5. 4 $SBR_B$ Accuracy evaluation for both models through: mean average deviation (MAD), mean squared error (MSE), root MSE, mean absolute percentage error (MAPE) and squared correlation coefficient ($R^2$)

| Method | NH$_4$-N | | | | | |
|---|---|---|---|---|---|---|
| | MAD* [val] | MSE* [val] | RMSE* [val] | MAPE* [val] | $R^2$ [val] | $R^2$ [train] |
| **GB-Trees** | 22 | 700 | 26 | 16 | 0.97 | 0.99 |
| **nu-SVM** | 18 | 627 | 25 | 12 | 0.97 | 0.99 |
| **SVM** | 26 | 1,100 | 33 | 20 | 0.94 | 0.95 |
| **RF** | 43 | 2,679 | 52 | 29 | 0.87 | 0.90 |
| **Method** | **NO$_3$-N** | | | | | |
| | **MAD* [val]** | **MSE* [val]** | **RMSE* [val]** | **MAPE* [val]** | **$R^2$ [val]** | **$R^2$ [train]** |
| **GB-Trees** | 15 | 377 | 19 | 20 | 0.93 | 0.97 |
| **nu-SVM** | 14 | 334 | 18 | 21 | 0.92 | 0.92 |
| **SVM** | 16 | 396 | 20 | 24 | 0.90 | 0.89 |
| **RF** | 21 | 779 | 28 | 32 | 0.81 | 0.83 |

*in mg N $L^{-1}$

Figure 5.7 shows the performance of $\nu$-SVM over the operational time of the SBRs. The results for training and validation are illustrated for both reactors. The prediction results involve the complete operational period for both systems, including the start-up of the reactor. While the ammonium in $SBR_A$ experienced steep changes (i.e., accumulation of ammonium between days 184 and 254), the $\nu$-SVM model performed better. On the other hand, ammonia reduction gradually improved in $SBR_B$. In this system the $\nu$-SVM model also outperformed the other approaches. The results suggest that $\nu$-SVM, are a robust approach to predict PN-A SBRs of dynamic nature where limitation of data for training exists. While these findings cannot be extrapolated to other systems, these ML methods indeed provide with an alternative effective approach to complex predictions. Unlike ASM models, the systems were effectively modeled using complex relationships among variables, such as conductivity, aeration pattern information, ORP, among other variables (See Table 5.2).

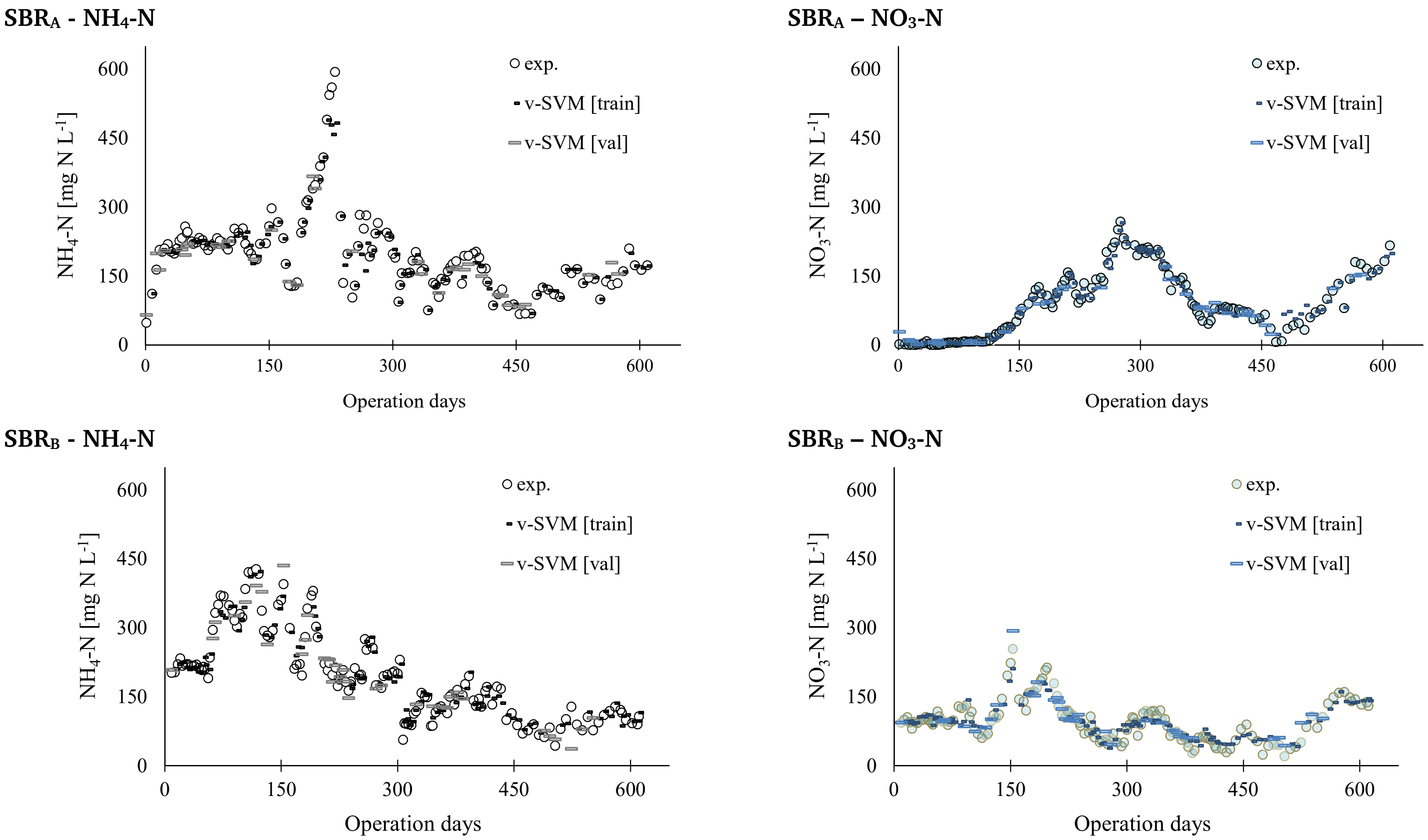


Figure 5. 7 Prediction of the effluent composition for SBR$_A$ and SBR$_B$ reactor through $\nu$-SVM along the operation of the reactor. The effluent concentration of ammonia (NH$_4$-N) and nitrate (NO$_3$-N) are predicted.

So far, the results show that a careful selection of a significant dataset (out of the initial raw data) led to highly accurate models to predict the effluent composition of both SBRs. As well, the most promising methods to predict the effluent of both SBRs were gradient boosting tree ensemble (GB-Tree) and $\nu$-SVM.

The application of ML techniques allowed a deep and comprehensive analysis of two full-scale PN-A systems ($SBR_A$ and $SBR_B$). Thus, this work showed a novel approach that leads to a practical application of predictive models by using state-of-the-art supervised ML. Hence, there are three promising findings:

(1) Data were carefully pre-processed and scaled in order to allow further feature selection to select the best configuration for the features and observations. Feature subsets were built while progressively eliminating variables which contained the highest number of missing values or observations. As a result, seven subsets were obtained for each system.

(2) The best *predictors* from each feature subset were extracted by using RFE. However, since the number and variables were to some extent different in each subset, the best configuration of features and observations was found through a score function ($\varphi_i$), which was proposed in this work.

(3) Supervised ML techniques were applied for the prediction of both systems. Highly accurate models were obtained, through the predictive models (average correlation coefficient higher than $R^2 > 0.90$), where the $\nu$-SVM method outperformed GB-Trees, and RF, for the prediction of the effluent of both systems (average correlation coefficient above $R^2 > 0.95$).

To elucidate the real contribution of the optimization of feature selection and to evaluate the effectiveness of this procedure, different case-scenarios were evaluated. The first case-scenario evaluated the predictive models; GB-Tree and $\nu$-SVM in subset S6 without feature selection. The second case-scenario evaluates subset S7 for building predictive models, S7 contains the highest amount of observations. In case-scenario 2, the hypothesis is that the performance of the models would increase since more observations are provided to train. Finally, the third case-scenario evaluates the implementation of new engineered features with the application of feature engineering, to demonstrate the feasibility of using non-conventional input parameters in the models. In all these case-scenarios, only $SBR_A$ dataset is used.

*Case-scenario: 1*

In this case scenario, subset S6 (136 Obs.) was used for building predictive models to demonstrate the importance of feature selection. Thus, all features in this subset were used as input parameters instead of the ones previously determined. The total amount of input features in S6 is 19, compared to 12 after optimization of feature selection (previous results). Figure 5.9 and Table 5.5 show the results obtained when 19 input features are considered to predict $NH_4$-N and $NO_3$-N.

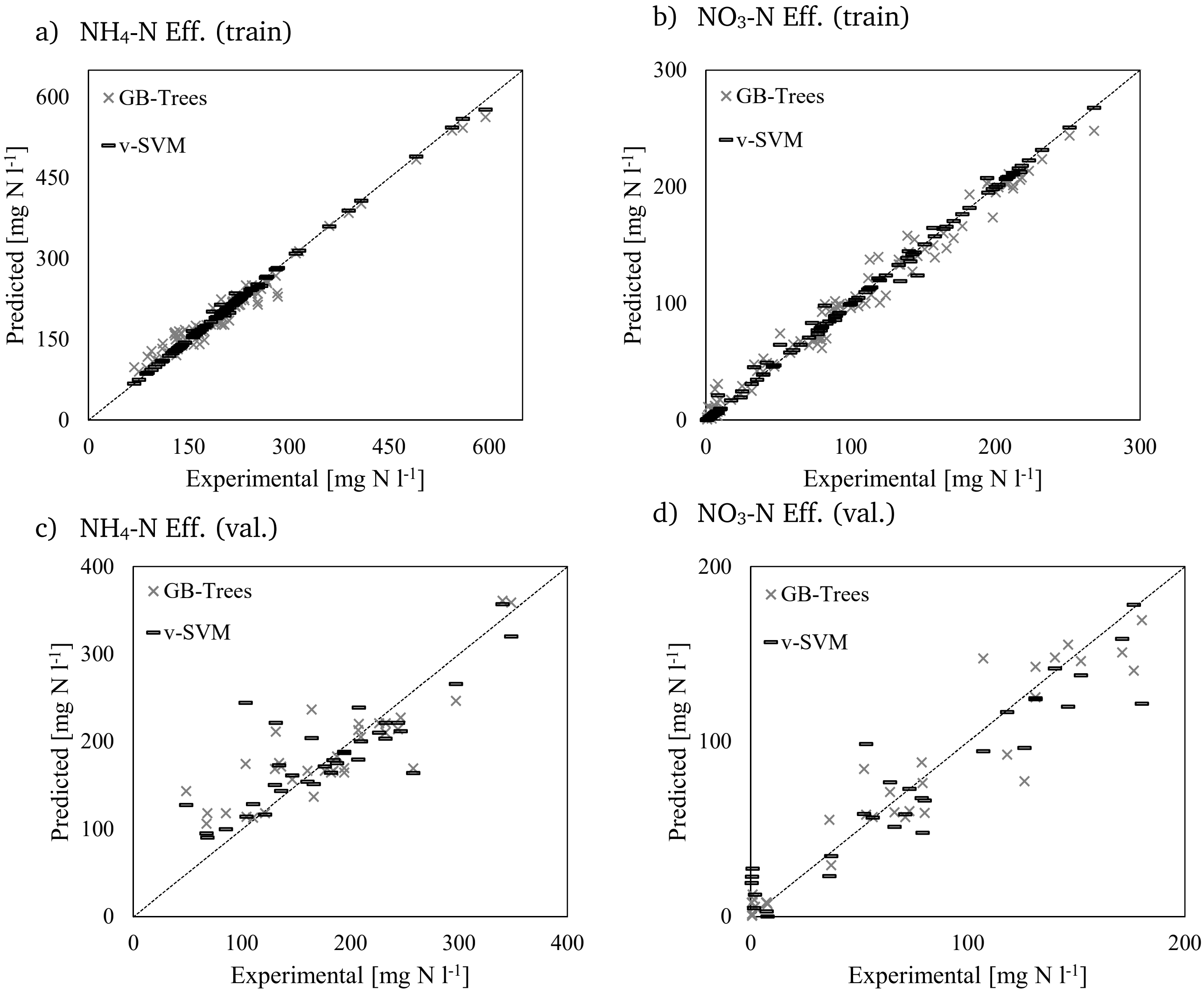


Figure 5. 8 Training and validation results for the prediction of SBRA effluent with S6

Table 5.5 shows the squared correlation coefficients obtained in the training and validation phases when 19 input parameters were considered to predict the effluent of SBR$_A$ (no feature selection). When compared with the models previously built (optimization of feature selection, Table 5.2), the squared correlation coefficients on the validation are considerably better than this case-scenario.

Table 5. 5 Squared correlation coefficient results for training and validation

| **Method** | **Train** | | **Validation** | |
|---|---|---|---|---|
| | **$NH_4$-N** | **$NO_3$-N** | **$NH_4$-N** | **$NO_3$-N** |
| GB-Trees | 0.99 | 0.99 | 0.85 | 0.96 |
| $\nu$-SVM | 1.00 | 1.00 | 0.83 | 0.96 |

Computationally, the complexity of any ML predictive model mainly depends on the amount of data to train and the input features; the highest the amount of input features, the higher the computational complexity to build a model. In SVM, the complexity order also depends on the amount of support vectors and in GB-Trees depends on the number of trees. However, assuming that comparable amount of trees and support vectors where used in this case-scenario, then, the complexity order mainly depends on the amount of input features. Clearly, 19 input features is higher than 12 (previously built model). Additionally, the influence of irrelevant input features clearly reflects in the results in Table 5.5. The models without feature selection achieved inferior performance than when feature selection was performed.

*Case-scenario: 2*

In this case-scenario, the subset with highest amount of observations was used for building the predictive models, S7 (261 Obs.). The amount of input parameters in this subset is 18 features. For building the models in this case-scenario, 195 observations were considered for training and 66 for validation (*4-fold* cross validation). Figure 5.10 shows the results obtained.

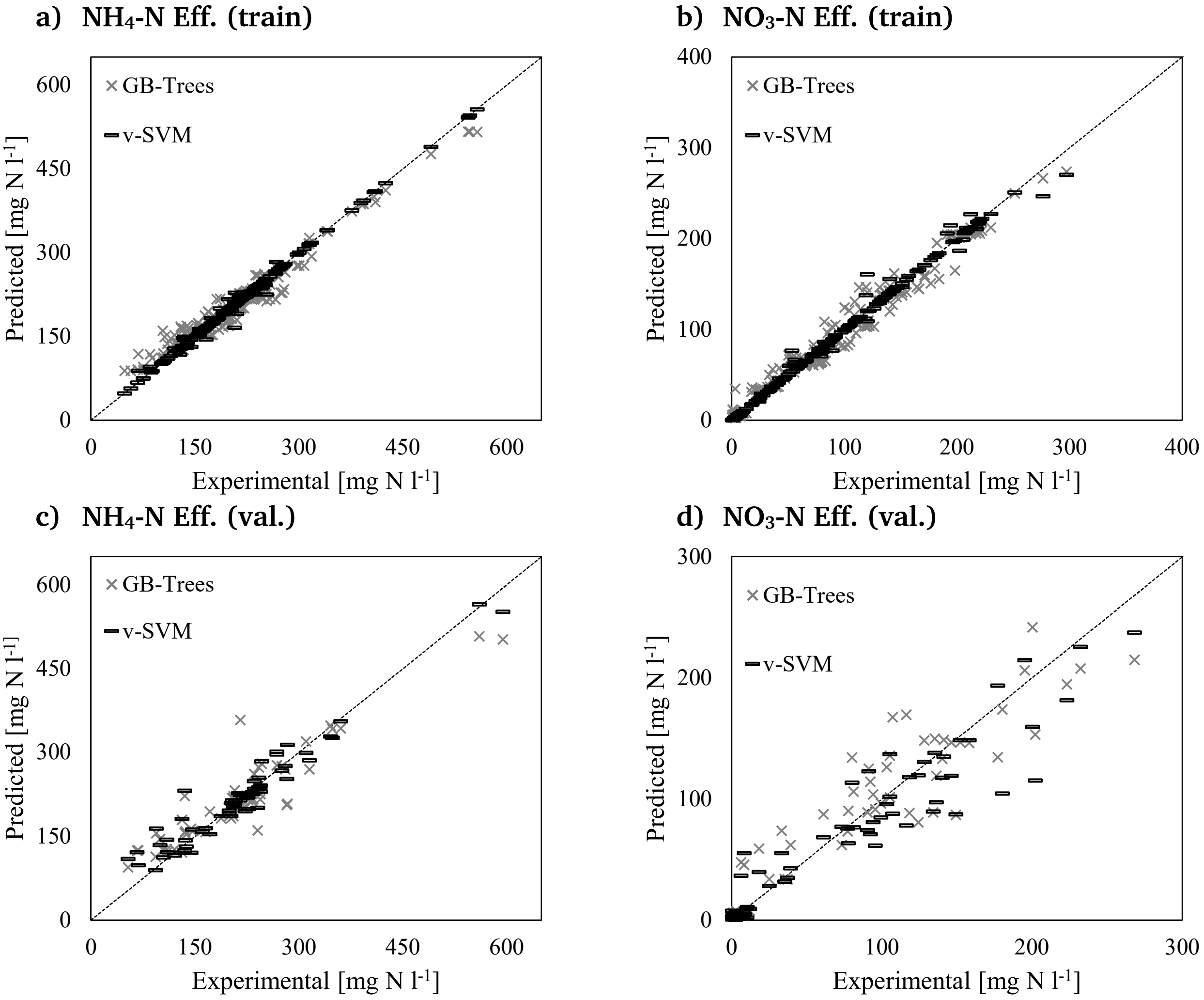


Figure 5. 9 Training and validation results for the prediction of SBRA effluent with S7

Table 5.6 summarizes the squared correlation coefficients obtained in the training and validation stage. Clearly, highly accurate results can be obtained with higher amount of observations. The main reason is that higher amount of observations, provide more examples for the models to train, and as a result, more accurate models can be obtained. However, in this case-scenario, still the complexity order is high (18 input features). In the next case-scenario, this dataset (S7) will be used and feature engineering will be also applied to obtain new predictors and reduce the dimensionality of the problem.

Table 5. 6 Squared correlation coefficients obtained from training and validation datasets in case-scenario 2.

| Method | Train | | Validation | |
|---|---|---|---|---|
| | $NH_4$-N | $NO_3$-N | $NH_4$-N | $NO_3$-N |
| GB-Trees | 0.98 | 0.99 | 0.93 | 0.94 |
| $\nu$-SVM | 1.00 | 1.00 | 0.97 | 0.95 |

*Case-scenario 3: Application of feature engineering*

Although more accurate results were obtained in case-scenario 2 (more observations provided to train), the order of complexity is still high in subset S7. An alternative is the development of new features and/or transformation of the dataset to create more accurate models i.e. feature engineering. Feature engineering allows the transformation of the datasets with the application of mathematical operations; natural logarithm, multiplication, division, etc. Other transformations such as Yeo-Johnson which belongs to the power family of functions and is defined in the Appendix, A2.
In this case-scenario, 13 additional non-conventional features were created randomly computing ratios of parameters from subset S7; pH:T, pH:Cond., pH:DO Av., pH:DO Max., T:Cond., T:DO Av., T:DO Max., Cond.:DO Av., Cond.:DO Max, Cond.:pH, Cond.:T, pH:Cond:T. Afterwards, Yeo Johnson transformation was applied and feature selection was performed to the modified dataset (Yeo and Johnson, 2000). Figure 5.11 shows the feature importance ranking of random forest for this modified dataset. Random forest was selected as feature selection method in this case-scenario.

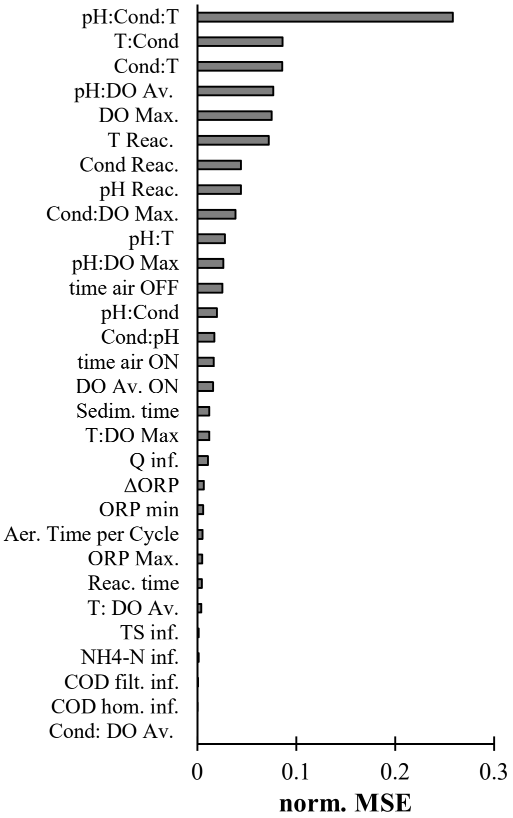


Figure 5. 10 Feature importance ranking obtained for the modified S7 subset. norm. MSE: normalized mean squared error.

The results clearly show the new features were placed at the top of the ranking, highlighting their relevance for the prediction of $NH_4$-N and $NO_3$-N. The results demonstrate the potential use of feature engineering for finding new features. The top 5 features in the ranking were selected as predictors; DO

Max., pH:DO Av., Cond:T, T:Cond and pH:Cond:T, this selection was based mainly to reduce the computational complexity, it is important to clarify that more features from the ranking may be considered to build the models. Figure 5.12 shows the results obtained after building the predictive models.

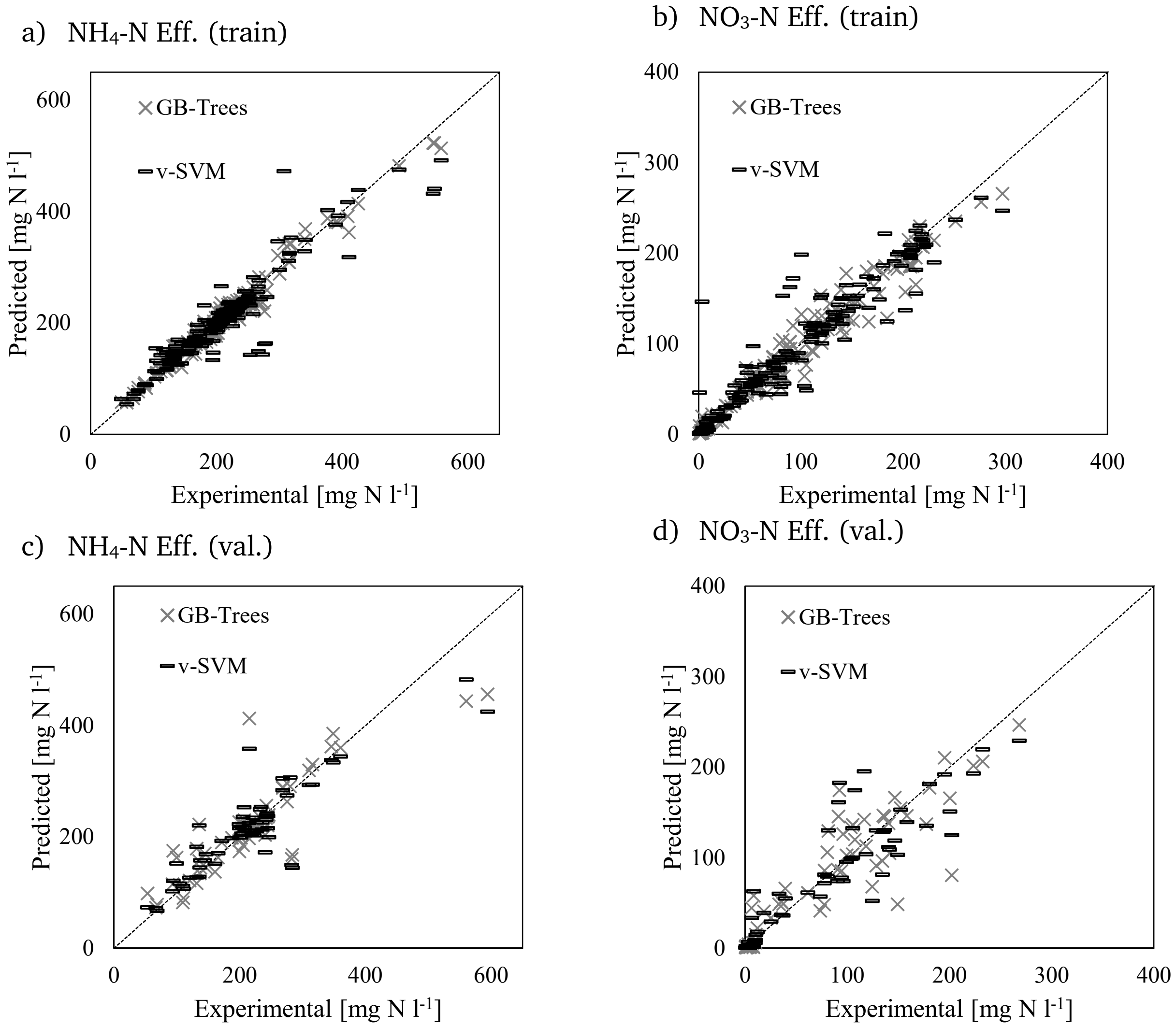


Figure 5. 11 Training and validation results for the prediction of SBR$_A$ effluent after feature selection in S7

Table 5.6 shows the results for the squared correlation coefficients obtained for training and validation. The results demonstrate the feasibility of obtaining highly accurate models with the non-conventional predictors obtained after feature engineering, compared to the results obtained in the initial set-up (when optimization of feature selection was performed), the squared correlation coefficients were higher than in this case-scenario: more than 0.95 for $NO_3$-N and around 0.9 for $NH_4$-N.

Table 5. 7 Correlation coefficient results for training and validation datasets.

| **Method** | **Train** | | **Validation** | |
|---|---|---|---|---|
| | **$NH_4$-N** | **$NO_3$-N** | **$NH_4$-N** | **$NO_3$-N** |
| GB-Trees | 0.99 | 0.98 | 0.88 | 0.90 |
| $\nu$-SVM | 0.93 | 0.96 | 0.89 | 0.91 |

The results are comparable to the obtained in this case-scenario and in addition, the computational complexity of the resulting model (5 input features) is lower than the initial configuration (12 features) and in previous case-scenarios (19 and 18 input features).

The practices implemented in this chapter, although powerful, are often overlooked in the current framework of ML in wastewater treatment. Furthermore, compared to other studies, in this approach, previously-optimized sets of engineered features were used.

## 5.4 Chapter conclusions

In this work, a comprehensive data analysis of two full-scale *Partial Nitritation-Anammox* (PN-A) reactors was performed. Due to the heterogeneity of the datasets, similar to Chapter 4, this chapter aimed at finding an optimal configuration of features and observations to build a predictive model to forecast the effluent composition of both SBRs. To select an optimal configuration for the number of features and observations a score function was defined. Incorporating a *score function* together with the feature selection (RFE) allowed to find the best configuration of features and number of observations. The hypothesis was that the resulting configuration would result in a significant subset of the total raw data; a balance between important features and amount of observations. The best configurations of features and observations for both reactors were; 12 and 136, respectively for $SBR_A$ and 12 feature and 142 observations for $SBR_B$.

After the optimization of the feature selection process, different supervised ML methods were evaluated; GB-Trees, $\nu$-SVM, SVM and RF. Overall, $\nu$-SVM and GB-Trees achieved the highest performance ($R^2>0.95$) and lowest error (MAPE<10%) in the training and validation stages. To evaluate the actual contribution of the feature selection optimization, different case-scenarios were evaluated. In the first case-scenario, the performance $\nu$-SVM and GB-Trees decreased since no feature selection was performed, achieving a squared correlation coefficient of 0.79 in comparison to values over 0.95 for the optimized configuration. Additionally, the computational complexity in this case-scenario was higher than the optimized configuration; 19 input features in S6 compared to 12 in the optimized configuration. In the second case-scenario; the subset with the highest amount of observations (S7) was applied for prediction tasks. More observations allowed more training examples, although the computational complexity was similar to the first case-scenario. In the third case-scenario, the input features of S7 were engineered to create better predictors; DO Max., pH:DO Av., Cond:T, T:Cond and pH:Cond:T. All these parameters can be obtained from the online dataset. The performance of the models with the new engineered features achieved squared correlation coefficients up to 0.9, demonstrating the high potential of applying both, unconventional and engineered features to model the effluent composition of full-scale PN-A systems.

# 6 A new approach to explore patterns in combined sources of data in PN-A systems

## 6.1 Introduction

In any bWWTP, monitoring certain process parameters is essential for the control and evaluation of the performance of the system. Thereby, some parameters are more relevant than others. However, due to i) the high density of data (from once per day down to even every second), ii) highly dimensional systems (multiple variables monitored) and iii) the dynamic nature of the processes involved, most of the data is commonly not studied or analysed in depth. The high dimensionality of the datasets (many parameters) and large amount of data makes it challenging to process information for better understanding or proper decision-making (Corominas et al., 2018). Accordingly, more efficient and robust statistical tools for data analysis are required to extract knowledge from these datasets. So far, heterogeneous datasets were analysed to extract patterns and to predict the effluent of bWWTP through ML. In this chapter, a new approach to explore patterns and combine different sources of data (online, lab and batch experiments) generated in PN-A systems is proposed. The first part of this chapter was focused on feature selection to find strong interactions and relevant patterns within online and offline data from different PN-A systems; two lab scale moving bed biofilm reactors (MBBR) and two full-scale SBRs (same reactors studied in Chapter 5). In the second part of this chapter, the selected features from the previous step were coupled to a (small) dataset obtained from batch activity tests to the biomass of the PN-A systems.

The main hypothesis of this chapter was that by using unsupervised ML (i.e., data clustering) with the aid of feature selection (supervised ML), unseen operational conditions that may benefit the performance of the reactors while studying both off-line, online (high amount of observations) and batch activity (low amount of observations) data, information can be discovered.

Different studies have applied unsupervised ML methods in wastewater treatment applications (Aguado et al., 2008; Di et al., 2019; López Garcı a and Machón González, 2004). So far, ML methods were not applied to study datasets which share different amounts of observations.

## 6.2 Methodology

### 6.2.1 Operation of the MBBRs

Two PN-A MBBRs containing different carrier materials (K3 and Biofilm Chip$^{TM}$ M, AnoxKaldnes, Sweden) were operated for around two years. The first year the systems were fed with synthetic feed ($NH_4$-N only) and exposed to seasonal temperature fluctuations back and forth from 20°C to 10°C, the second year, they ran with real municipal wastewater (organic matter) and the same seasonal temperature fluctuations. Both reactors were monitored with online sensors (online data) and also sampled daily or weekly for further influent and effluent monitoring; around 25 parameters were analysed in this work (off-line and online datasets). Additionally, along the operation of both reactors, biomass was taken regularly from the laboratory-scale reactors to measure the specific rates for nitritation, nitratation, anammox and heterotrophic activity in *ex situ* batch experiments (batch activity datasets); ammonium oxidizing bacteria (AOB), nitrite oxidizing bacteria (NOB) and Anammox bacteria (AnAOB). Further information on the reactor operation and batch tests is available in Agrawal, (2018) and Gilbert et al., (2014).

This work was partly published in Alejo, L., Atkinson, J., Lackner, S., 2020. Looking deeper – exploring hidden patterns in reactor data of N-removal systems through clustering analysis. Water Science and Technology 81, 1569–1577. https://doi.org/10.2166/wst.2020.029

### 6.2.2 $SBR_A$ and $SBR_B$ batch activity datasets

Batch activity tests were conducted to follow the specific activities of AOB, NOB and AnAOB in the two SBRs studied in chapter 5 ($SBR_A$ and $SBR_B$), the operational conditions of these reactors can be found accordingly, in chapter 5.

From comparing batch activity tests with reactor performance it was evident that the specific biomass activity was well suited to represent overall reactor performance. Batch activity losses even preceded the drop in reactor performance. Aerobic and anaerobic batch activity tests were regularly conducted for both reactors ($SBR_A$ and $SBR_B$), to monitor the activity of the biomass (Table A.4). These tests were only performed every other week. The density of this kind of data was small and therefore not considered in the feature selection process.

### 6.2.3 Datasets analysis

A key issue with the datasets studied in this chapter was the high number of features (i.e., variables) involved in the lab and full scale systems, which may significantly affect the results and quality of any data analysis task such as clustering. In order to address this, first, feature selection methods were used to reduce data dimensions and to make the clustering task more efficient. With feature selection, strong interactions between influent and effluent parameters were found. In this chapter, feature selection was used to study the influence of process parameters in key variables for PN-A performance; removal of ammonium ($NH_4^+$), nitrate ($NO_3^-$) and organic matter (COD) for both full and lab scale systems. Additionally, nitrite ($NO_2^-$) was studied also in the lab-scale reactors.

The size of the batch activity datasets is small compared to the off-line and online datasets, these datasets were not used in the feature selection process, which bring us to the second part of the study. Once feature selection was performed (with the lab and online datasets), the selected features subset was coupled to the batch activity datasets (Figure 6.1), and from thereon, it is referred to as *Clustering dataset*. The feature selection was performed with the *randomForest* library in R (Kuhn, 2008; R Foundation for Statistical Computing, 2016).

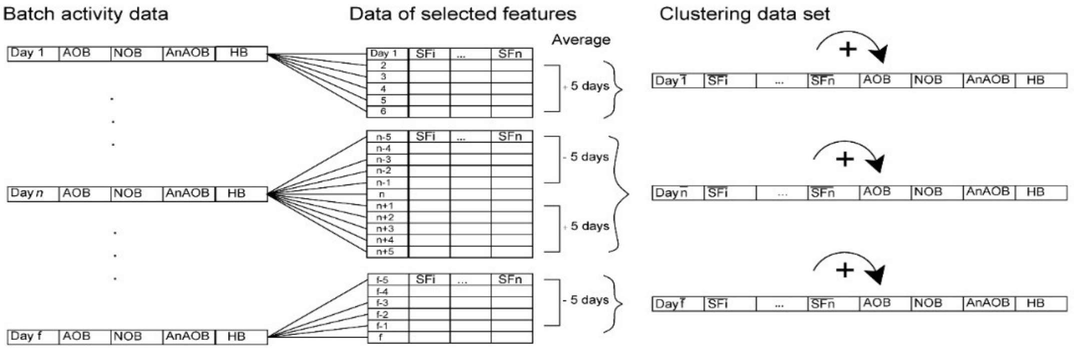

Batch activity data
Data of selected features
Clustering data set
Average
Day 1 | AOB | NOB | AnAOB | HB
Day n | AOB | NOB | AnAOB | HB
Day f | AOB | NOB | AnAOB | HB
Day 1, 2, 3, 4, 5, 6 | SFi | ... | SFn
n-5, n-4, n-3, n-2, n-1, n, n+1, n+2, n+3, n+4, n+5 | SFi | ... | SFn
f-5, f-4, f-3, f-2, f-1, f | SFi | ... | SFn
+ 5 days
- 5 days
+ 5 days
- 5 days
+
Day1 | SFi | ... | SFn | AOB | NOB | AnAOB | HB
Dayn | SFi | ... | SFn | AOB | NOB | AnAOB | HB
Dayf | SFi | ... | SFn | AOB | NOB | AnAOB | HB

Figure 6. 1 Scheme for building the Clustering dataset: combination of batch activity and the subset of features after dimensionality reduction in the lab and online datasets. Representative data from the lab and online data information is computed as the average of a range between [-5 - +5] days for each selected feature (SFi).

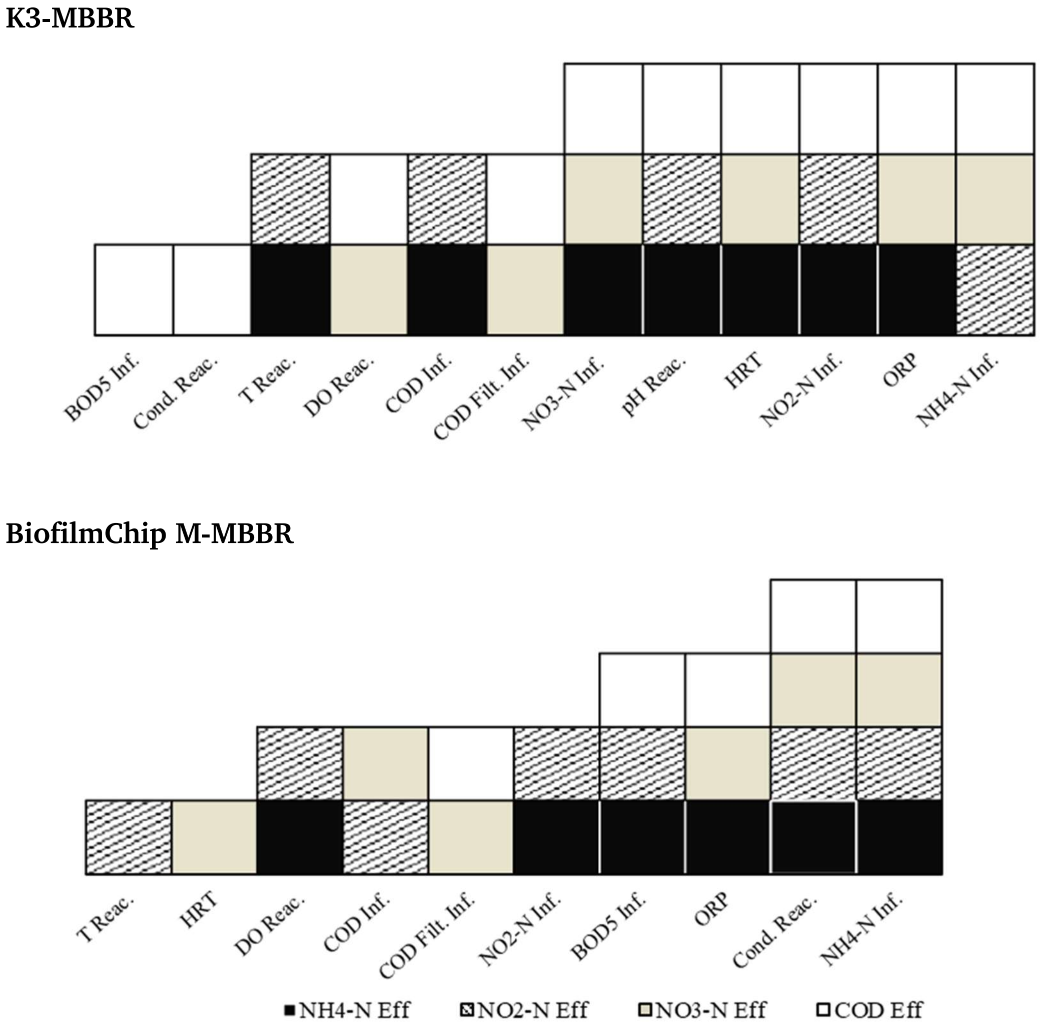
Each data point (i.e. tuple) in the *Clustering dataset* (Figure 6.1) was built so that the dates when the batch activity test were performed matched an average of ±5 days of the lab-scale dataset (after FS). After building the *Clustering dataset*, the k-means clustering algorithm was applied to build related clusters. K-means estimates the unknown cluster centers (aka. *centroids*) for each of K clusters by minimizing each data to the closest centroid, at the same time, each cluster's centroid is defined when the distance between a data point in the dataset is far from a previous centroid. Clustering analysis with k-means was implemented with the *stats* (a default library in R) and *cluster* libraries in the R system (Maechler et al., 2018).

The optimal number of clusters were determined by using the *elbow-criterion*; the number of clusters should be selected such that by adding other cluster, no new information is gained i.e. from the information theory perspective, it does not reduce *entropy*. The *within-cluster sum of squares* (Wk), is a parameter from k-means, which was plotted against an increasing number of clusters. Graphically, Wk decreases monotonically as the number of clusters increases, but from a certain number of clusters, the decrease flattened markedly, resulting in the optimal number of clusters (Tibshirani et al., 2001). For this purpose, a Wk was computed for a different number of clusters for both datasets (See Figure A.18, Figure A.19 in the Appendix). Since datasets contain highly-dimensional data, clusters cannot properly be illustrated. Instead, generated clusters were represented using data's Principal Component Analysis (PCA) representation. In our experiments, only the two most important principal components were used as score vectors.

## 6.3 Results and discussion

### 6.3.1 MBBRs

This work aimed at uncovering hidden patterns through clustering analysis for two MBBRs. For this purpose, pre-processing and feature selection tasks were first performed as seen in Figure 6.2. The feature selection method selected the most relevant variables within the computed ranking through the random forest algorithm while minimizing the root mean squared error from the importance score. The *relevance* of a feature (i.e variable) is defined as strong interactions between the parameters within a dataset, for example, in a prediction task, the relevant features will share strong interactions to the output features (Blum and Langley, 1997). Here, the set of features in Figure 6.2 were the most influencing for the effluent features analyzed in this work; organic matter concentration (COD Eff.), nitrate concentration ($NO_3$-N Eff.), nitrite concentration ($NO_2$-N Eff.), and ammonium concentration ($NH_4$-N Eff.). Data dimension was reduced up to 50% of the original (i.e., from ~25 down to around 12 features). Even though the biofilm characteristics between both reactors were different (2mm vs. 10 mm thickness of the carrier), both reactors shared almost the same relevant features; temperature (Temp. Reac.), ORP, dissolved oxygen concentration (DO), conductivity (Cond. Reac.) (more relevant in BiofilmChip M-MBBR than K3-MBBR), hydraulic retention time (HRT) and ammonium ($NH_4$-N Inf.), nitrite (NO2-N Inf.), biological oxygen demand ($BOD_5$ Inf.) and chemical oxygen demand (COD Inf.) concentrations in the influent. The influence of the concentrations of the nitrogen species is undisputable in deammonification processes since they are target pollutants to be biologically removed. The influence that organic matter has on deammonification systems is also expected; the literature supports that COD/N ratios greater than 1 have a marked influence on the efficiency of ammonium removal (Hulle et al., 2010). The impact of temperature on the nitrite concentration in the effluent also agrees with related research (Gilbert et al., 2014). The ORP is a parameter used in wastewater treatment to monitor the occurrence of specific biological reactions. The presence of an oxidizing agent such as oxygen, increases the ORP value, while the presence of a reducing agent such as substrate, decreases the ORP value (Wouters-Wasiak et al., 1994; Zipper et al., 1998). Accordingly, the values of ORP correlated well with the oxygen concentration just before both reactors were exposed to different feed compositions (i.e., real wastewater, containing organic matter). The ORP value then dropped to values below and around 50 mV in both reactors; the same results were found in the clustering analysis.


K3-MBBR

BiofilmChip M-MBBR

Figure 6. 2 Feature selection through random forest for both MBBR lab and online datasets.

In PN-A systems, the conductivity signal is an important parameter for controlling the ammonium concentration. Joss et al., (2009) found, that both signals correlated linearly. Accordingly, in this work, the correlation coefficients between the conductivity and the ammonium concentration were above 0.7. In general, DO plays a key role in the out-competition of NOB while maintaining high Anammox activity in any PN-A system (Hao et al., 2002; Mattei et al., 2015; Vangsgaard et al., 2012; Wyffels et al., 2004). Mattei et al., (2015) proposed different scenarios where both the DO and shear stress influenced the microbial distribution of a biofilm PN-A systems; prolonged exposure of the biofilm to DO concentrations around 5 mg DO $l^{-1}$ resulted in the inhibition of AnAOB. Our results suggest (also further seen in the clustering analysis) that an average value of 0.3 mg DO $l^{-1}$ would benefit the activity of AOB and at the same time guaranteed no further inhibition of AnAOB.

After feature selection, clustering analysis was conducted by using the k-means method. Following the *elbow-criterion,* three was the optimal number of clusters for both MBBRs datasets (see Figure A.18). Figure 6.3 shows the identified clusters. Cluster 1 in both reactors grouped the data where the highest efficiency of ammonium removal and highest AOB and AnAOB activity was found (Figure 6.4 and Table A.6), whereas Cluster 3 (K3) and Cluster 2 (BiofilmChip M) enclosed the lowest activity of the biomass and lowest efficiencies. The remaining clusters grouped the data where the systems were exposed to real municipal wastewater (i.e., presence of organic matter).

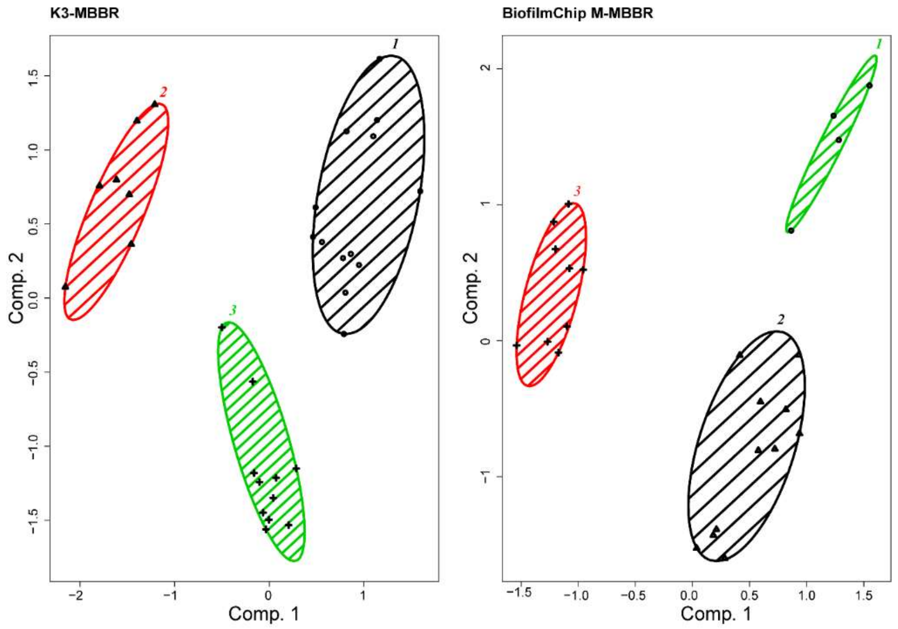


Figure 6. 3 Clusters obtained after using the k-means algorithm: K3-MBBR (left) and BiofilmChip M-MBBR (right).

The results are consistent from an operational perspective. Higher relative activity of the AOB resulted in accumulation of nitrite in the effluent (Figure 6.4). This accumulation was due to the temperature drop in both systems from 20 to 10°C (Gilbert et al., 2014). Moreover, in the K3-MBBR reactor, the ORP value dropped from around 300 mV in cluster 1 to an average value of 200 mV in cluster 3, whereas for the BiofilmChip M-MBBR, the ORP decreased progressively (Figure 6.5).

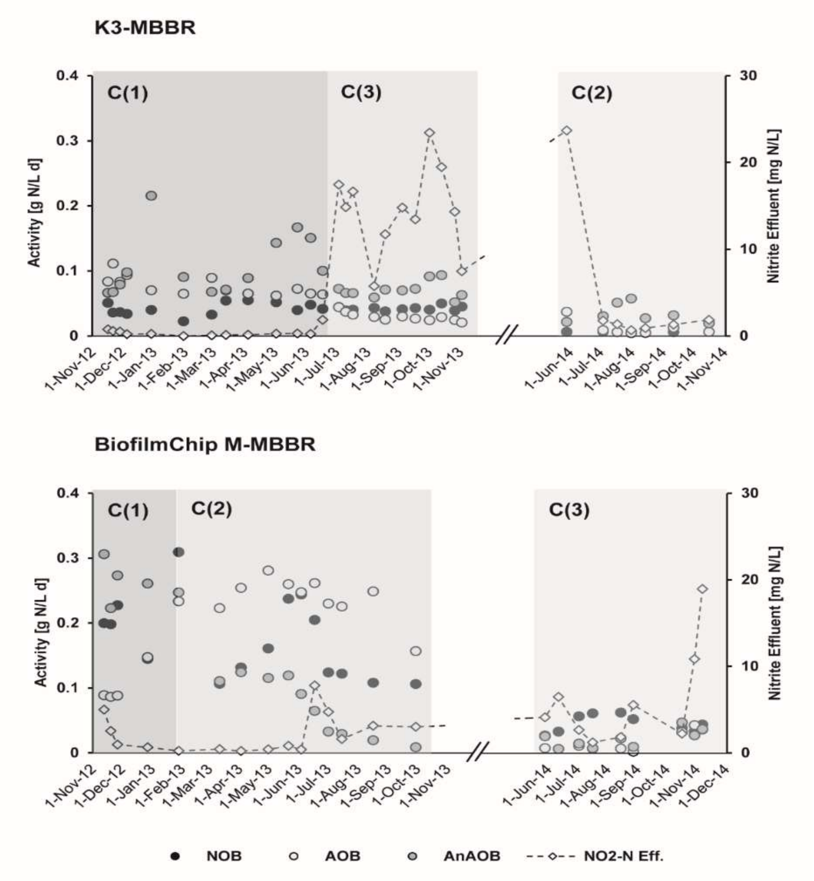


Figure 6. 4 Clusters analysis results: Activity of biomass and nitrite accumulation. C(i): Cluster i, were i: [1,2,3].

The AOB to nitrite oxidizing bacteria (NOB) ratios (AOB/NOB) were analyzed in Figure 6.5. For cluster 1 and 3 in the K3-MBBR, the AOB/NOB ratio changed uniformly while a clear difference was observed in the BiofilmChip M-MBBR. The AOB/NOB ratio was characterized by values above 1 for the BiofilmChip M-MBBR in cluster 2, whereas for the other two clusters, the AOB/NOB was in average below 1, the latter leads to nitrite accumulation in cluster 2. Cluster 2 also revealed a drop in AnAOB activity suggesting inhibition by nitrite.

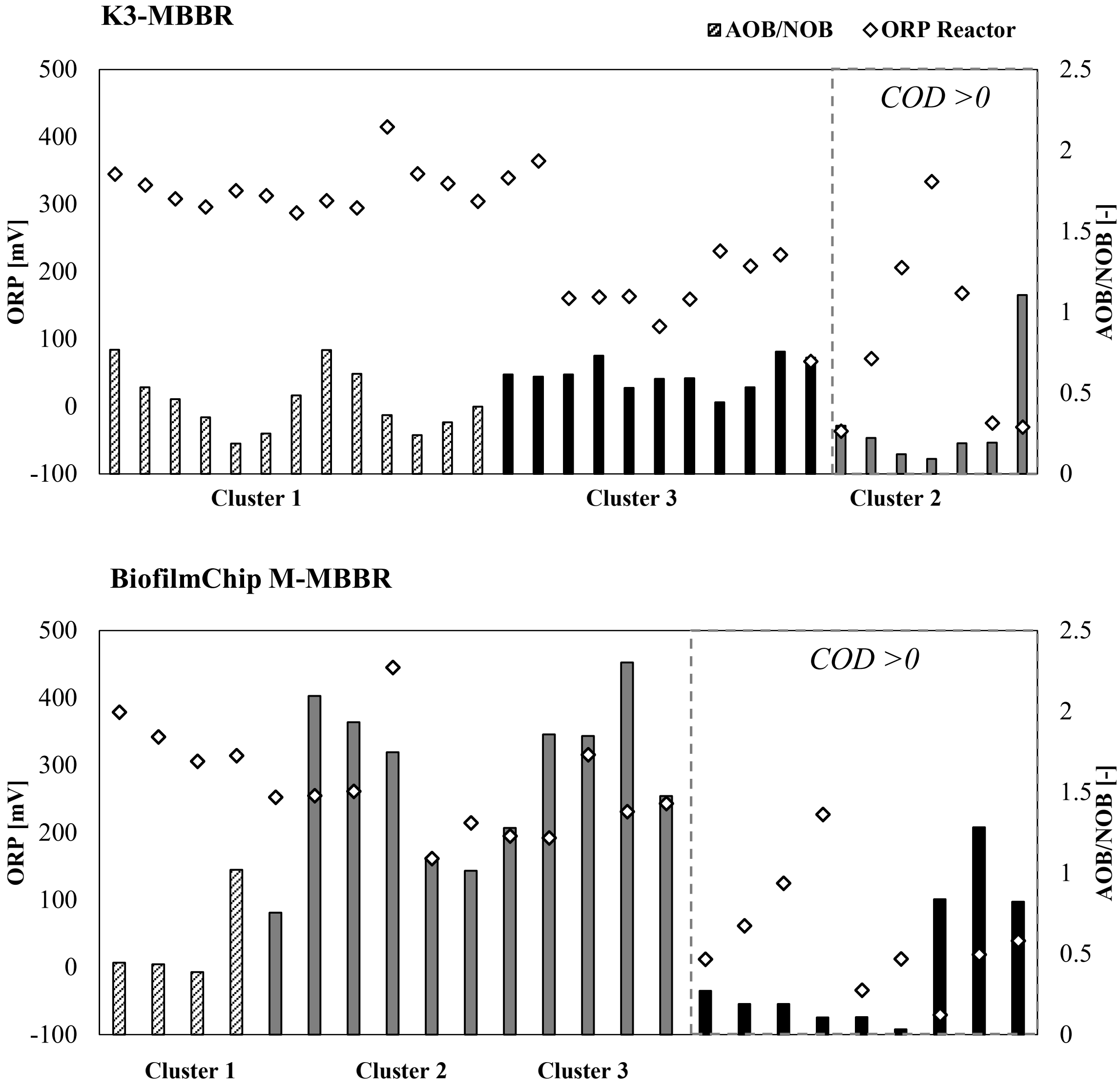


Figure 6. 5 Clustering results correlated to the activity of the biomass (AOB to NOB ratio) and ORP.

Table A.5 summarizes the centroids of each cluster in both datasets (a centroid being the average value of the data points enclosed in each cluster). By analyzing the centroids of the clusters (Table A.5), further patterns were detected; during the operational period with synthetic feed, both MBBRs showed higher values of ORP (around 200 mV), whereas in presence of organic matter, the ORP dropped to values on average of lower than 100 mV. Moreover, while the MBBRs were fed with synthetic wastewater, higher values of ORP (~ 200 mV) resulted in higher activity of AnAOB, this trend was clearer for the BiofilmChipM-MBBR than for the K3-MBBR.

On the other hand, the DO concentration correlated well with the conductivity just before the system was exposed to organic matter (real wastewater). The average HRT in the K3-MBBR fluctuated around 1.9 days in cluster 1 and 3 whereas in cluster 2 (operation with real wastewater, i.e. organic matter), this value dropped to 0.95 days. Finally, the HRT in the BiofilmChip M-MBBR was very dynamic, jumping from values of 1.3 to 2.5 days and 1.1 days in cluster 1 to 3, respectively.

### 6.3.2 SBRs

Figure 6.6 shows the results from k-means clustering in $SBR_A$ and $SBR_B$. The optimal number of clusters was also three in these datasets (Figure A.19, Appendix). The implication of feature selection in the datasets of the SBRs was thoroughly reviewed in Chapter 5, which highly depends on the amount of observations in the dataset.

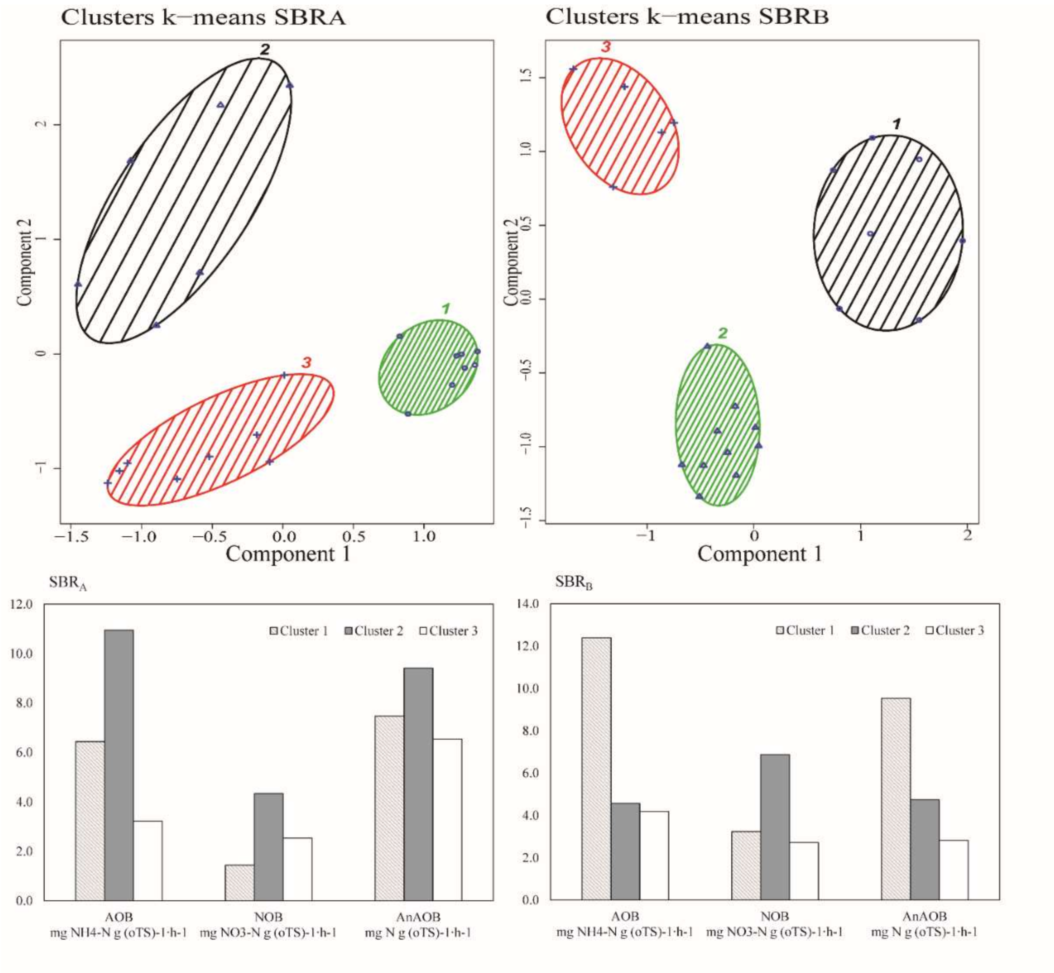


Figure 6. 6 Presented as Principal Component Analysis: Clusters obtained from the clustering analysis through k-means. Bars: Centroids of the clusters obtained from the clustering analysis.

Cluster 1 ($SBR_A$) and cluster 2 ($SBR_B$) were characterized by high activity of the biomass, whereas cluster 3 in both systems showed lower activity for both AOB and AnAOB. In $SBR_A$, NOB suppression was achieved in the operational period enclosed by cluster 2 which also showed the highest AOB and AnAOB activities. The results revealed that reaction times above 320 minutes, DO average concentrations of 0.3 mg $L^{-1}$, maximum DO concentrations (in the aeration periods) of 0.8 mg $L^{-1}$, and lower sedimentation times (9.1 minutes) favored the conditions for high ammonia removal (Table A.6). The results show also that DO concentrations highly influenced the activity of the aerobic biomass; cluster 3 enclosed the operational period where the highest DO concentrations and highest sedimentation times were registered and as a result, clear inhibition of AnAOB occurred. The pH did not play a key role, since the value in the three clusters was the same. In $SBR_B$, cluster 1 encloses the period where the highest ammonia removal was registered, this cluster was highly dependent on aeration pattern variables; *Aer. Time per cycle*, *Time Air ON* and *OFF*. Moreover, in this cluster, the sedimentation time and temperature, achieved the lowest and highest values, respectively. The TS concentration was similar to $SBR_A$, however, high concentrations of nitrate in the effluent were still found (around 94.2 mg N $L^{-1}$). Cluster 2 in $SBR_B$ showed the highest values for aeration pattern variables and the influence on the NOB activity was evident since the highest concentration of NOB was registered in this cluster.

Overall, the results agree with the literature; aeration and dissolved oxygen concentrations play a key role in the efficiency of PN-A systems. In this work, operational conditions, which favored NOB suppression, were discovered through clustering for both reactors; lower DO concentrations and lower sedimentation times favored ammonium removal and high activity of the biomass. Previous research (Laureni et al., 2019;Brockmann and Morgenroth, 2010) also studied NOB suppression through experimental and modeling, respectively. Furthermore, some research also showed that out competition of NOB was feasible under oxygen limiting conditions and elevated temperatures, resulting in higher AOB growth rates compared to NOB (Brockmann and Morgenroth, 2010).

### 6.3.3 A new approach to extract patters out of bWWTP data

Up to date, no other work in the literature has combined different data sources such as the ones presented in this chapter to extract knowledge. Aguado et al., (2008) applied self-organizing maps (SOM) and principal component analysis (PCA) for the analysis and interpretation of enhanced biological phosphorus removal processes. The size of their dataset comprised 11 parameters and 328 observations. The data was gathered over three months of operation which comprised also the start-up of the process. The main results demonstrated the feasibility of the application of unsupervised graphical methods to analyze multidimensional systems. Similarly, Garcia and Gonzalez, (2004) applied SOM and k-means to estimate and monitor the diverse states of waste treatment in an chromic acid WWTP. In this process 7 parameters were analyzed. The methodology was similar to the work of Aguado et al., (2008). Just recently, Di et al., (2019) explored two clustering methods to analyze more than 15 WWTPs in China and further explored patterns within the data. They employed partitioning-around medoids (PAM) and expectation–maximization (EM) (soft-clustering) analyzing 4 parameters in 18 monitoring sections in 7 administrative regions in the Yangtze River Basin (China) from 2016-17. Their results indicate that unsupervised ML can be applied for the identification of heavily polluted wastewater discharges and heavily polluted surface water. In most of these studies, a uniform dataset was used for analysis, which means that the resolution of the parameters analyzed was similar.

This work introduces a set of steps to combine three types of datasets (Figure 6.7): online, off-line and batch experiment data. These three main categories of datasets (online, laboratory and batch experiment data) are common in wastewater treatment.

- The online datasets include parameters that are monitored online by sensors and analyzers with the highest amount of observations (readings) in all categories (readings every second to every hour).
- Parameters measured in the laboratory constitute the lab-datasets. These parameters are measured daily, weekly or even only monthly; in our case, this dataset contained the second highest amount of readings and observations with almost daily measurements.
- Finally, in this study, batch experiments were performed also to monitor the activity of the biomass. These tests were measured every other week, thus, they show the lowest number of readings.

These different categories of datasets mentioned were studied through a new approach proposed. First, average daily values of the online datasets were considered and coupled to the lab dataset. Afterwards, feature selection was applied with the aim of; i) extract the parameters that have the highest impact in the process performance (i.e. identify strong interactions between influent and effluent parameters of interest) ii) reduce the dimensionality of the dataset so only key parameters would be analyzed with the batch activity data. In the second part of the approach, the batch activity and the selected features from online and lab datasets were merged according to the procedure described in Figure 6.7.

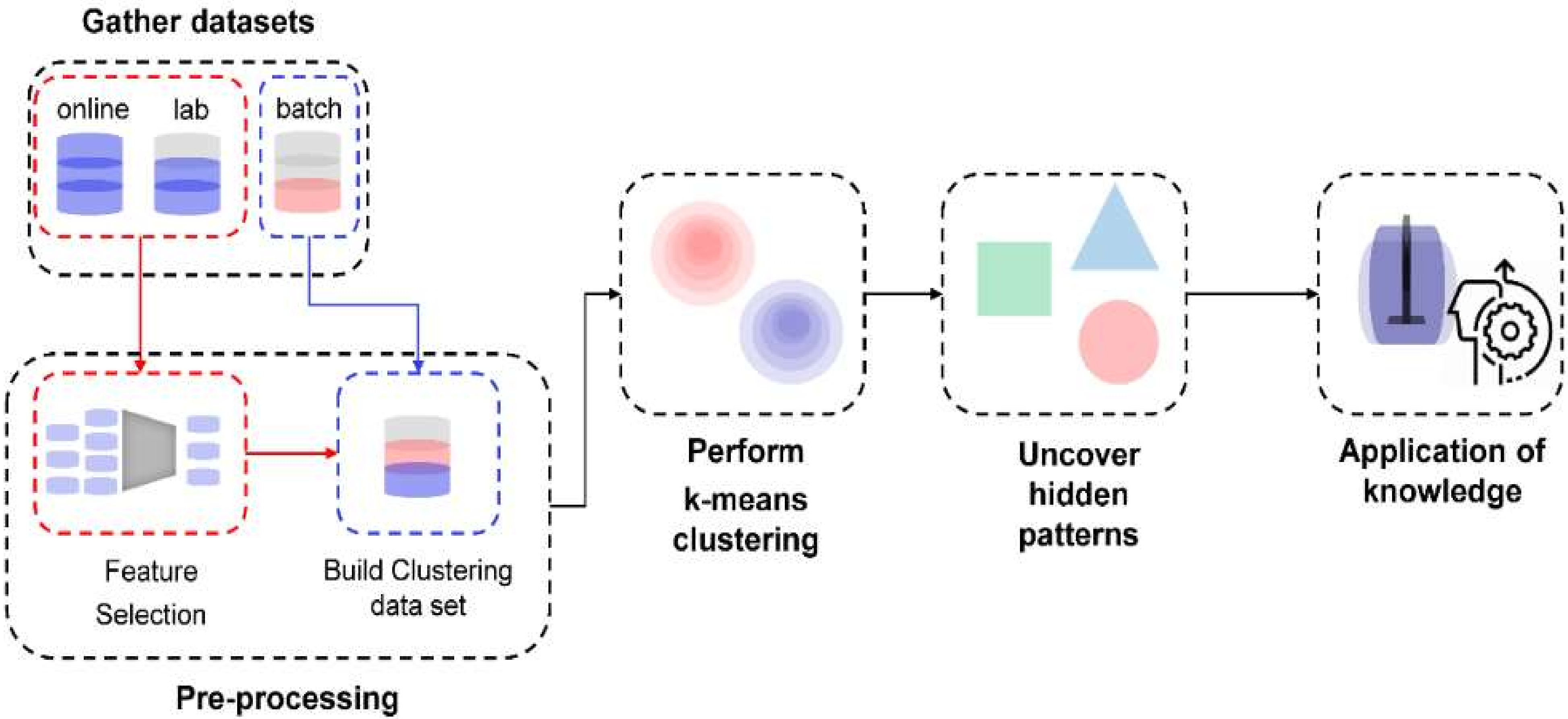


Figure 6. 7 Methodology for the extraction of patterns conducted in this work.

The results found not only agree with the literature but in addition, new hints to improve reactor performance were discovered. Therefore, the methodology illustrated here arises as an opportunity for the application to study similar dynamic and complex biological systems. Moreover, other methods from unsupervised ML can be applied instead of k-means clustering for discovery of patterns, such as association rules (Clark and Ma'ayan, 2011; James et al., 2013). The selection of methods for exploratory analysis (application of unsupervised ML), depends on the goal of the study. Although high amounts of clustering algorithms have been created after k-means, this method is still widely applied. The main advantages of k-means are: i) easy to implement and interpret the results ii) allows the organization of data into sensible groupings iii) adaptable to new data iv) the data to be analyzed does not require labels that tag the examples with prior identifiers. However, some disadvantages of k-means and other similar clustering algorithms are: i) the final results depend on the initial centroids (when random initialization is not applied) and ii) it is a method that is sensible to scaling. Particular to our work, the advantages are the interpretability of the results and the goal matched the method, the aim was to find groups in a high dimensional dataset and was accomplished. Methods such as PCA analysis are directed more towards a graphical representation and reduction of dimensions. In k-means clustering, on the other hand, we look for groups within the database that can elucidate relationships between sets of points. Another popular unsupervised ML method is SOM, which is a method build towards unsupervised neural networks. However, one of the main disadvantage of this method and its applicability in this chapter, was the amount of data. SOM requires sufficient data in order to develop meaningful clusters. The weight vectors must be based on data that can successfully group and distinguish inputs. Lack of data or irrelevant data in the weight vectors will add randomness to the groupings. Finding the correct data involves determining which factors are relevant and can be a difficult or even impossible task for several problems. The ability to determine a good dataset is a deciding factor in determining whether to use a SOM or not.

## 6.4 Chapter conclusions

Important knowledge was efficiently extracted from lab and full scale PN-A systems. The new approach proposed allowed to explore correlations and relations among parameters from online, lab and batch experiment datasets while coupling the information from all three groups to conduct clustering analysis. First, the identification of strong interactions between influent and effluent variables and at the same time dimensionality reduction was possible with feature selection. K-means clustering was applied in the *Clustering dataset* which comprised the parameters extracted from the feature selection process and coupled them to the batch activity dataset. In both, MBBR and SBR, clustering of the operational periods with high and low activity occurred. In the MBBR systems, DO and ORP, were discovered as relevant parameters for monitoring AOB and AnAOB activity. For the SBRs, reaction times above 320 minutes, average DO concentrations of 0.3 mg $L^{-1}$, maximum DO concentrations (in the aeration periods) of 0.8 mg $L^{-1}$, and lower sedimentation times (9.1 minutes) favored high ammonia removal.
Furthermore, this chapter contributes to promote the usage of clustering analysis for high dimensional, small and large datasets to discover and explore hidden patterns.

# 7 Application of ensemble learning in advanced wastewater treatment

## 7.1 Introduction

Very low levels of total phosphorus (TP), ~ 50 $\mu g$ $L^{-1}$ after secondary clarification (SC) can be achieved by means of in-line coagulation (chemical process) and the subsequent application of filtration, e.g. with membranes (physical process) (Gnirss and Dittrich, 2000; Zheng et al., 2012)·. Chemical phosphorus removal is a complex process to understand (Bratby, 2016). The general modeling approach in wastewater treatment for enhanced phosphorus removal is the family of ASM2 models (Chapter 2) (Gujer et al., 1995). Although these models consider chemical precipitation kinetics, the current framework is focused on; i) the removal of higher phosphorus concentrations than in aWWTP ii) the calibration of a model in extreme low concentrations of phosphorus would be unfeasible since the kinetic parameters under these conditions are yet unknown. Hauduc et al., (2015) developed a novel approach to model accurately the phosphorus removal through chemical precipitation considering different pathways of phosphorus removal and focusing on iron dosing. The pathways for phosphorus removal from bulk solution possible pathways are; i) adsorption of phosphates onto hydrous ferric oxide (HFO) by sharing an oxygen atom with iron; ii) co-precipitation of phosphate species into the HFO structure; iii) precipitation of ferric phosphate and iv) precipitation of mixed cation phosphates (Hauduc et al., 2015; Smith et al., 2008). Although the model developed by Hauduc et al., (2015) is robust and to correctly predict the initial fast removal of phosphorus, the model poorly describes further slow removal. Although this new approach is helpful to understand the mechanisms behind the chemical phosphorus removal, a key issue for the optimization of phosphorus removal is still the metal salt dosing, whether to reduce the metal salt dosing (economic revenue) or control of the chemical sludge production (Paul et al., 2001). In this context, data-driven methods based on ML appear as an opportunity to build a model that considers the metal dosing for phosphorus chemical removal without the kinetic complexity of a mechanistic models. Over the last decades, the application of data-driven methods, has been studied to study different bWWTP and has gained great popularity due to the high adaptability and low computational demand in comparison to deterministic models such as the activated sludge models (ASM) (Corominas et al., 2018). The generation of data in WWTP, that result from monitoring the quality of the effluent to meet the environmental regulations (Rieger et al., 2010), has led to a source of databases which can benefit from ML to extract novel and valuable knowledge. In supervised ML, a model is provided with examples of data for training. This past experience is used to fit the model that relates the *predictors* to the response (input to output variables) (Mitchell, 1997). Two factors will highly influence the performance of most supervised ML methods: 1) the size of the training dataset and 2) the problem and the method complexity (hyperparameters to adjust). In this chapter, both problems were approached. The data gathered comprised daily to monthly values of around 15 parameters from the aWWTP, the amount of information was limited to 344 datapoints (i.e. tuples) for two years of operation. Due to the dynamic nature of the process and the limited amount of information for analysis, advanced supervised ML such as *deep neural networks*, would require higher amount of data for training. Therefore, ensemble learning was explored to analyze the phosphorus removal in aWWTP. In Chapter 5, ensemble learning through gradient boosting trees demonstrated to be a suitable method where data limitation problem exists, such as in this chapter. The main premise of ensemble learning is that by combining multiple models, the errors of a single predictor will likely be compensated by others, and as a result, the overall prediction performance of the ensemble would be better than that of a single model (Chapter 3, section 3.4.5). Random forest is a type of bagging ensemble, where the output of the algorithm will be the average output from hundreds to thousands of decision trees (Breiman, 2001). On the other hand, boosting ensembles combine several weak learners or *predictors* sequentially, each trying to correct the predecessor (Friedman, 2001). These methods are mainly applied for prediction and classification tasks in ML. Some applications of gradient boosting ensemble in environmental sciences involve regression tasks to; forecast the organic fraction of municipal solid waste based on waste production and socio-demographic historical data (Adeogba et al., 2019), other study applied gradient boosting ensemble to predict the net ecosystem carbon exchange based on historical data involving water vapor, energy between terrestrial, ecosystems and the atmosphere, sum of global radiation, average air temperature and precipitation (Cai et al., 2020). In both studies, gradient boosting ensemble was compared to other state-of the art supervised ML methods such as support vector machines (SVM) and random forests, where the performance of the gradient boosting ensemble was better, thus validating the potential use of ensemble learning.

Yet another quality can be extracted from the ensembles. In specific, *feature importance* is a measurement extracted from random forest (Chapter 3, section 3.4). In this chapter, *feature importance* and a novel gradient boosting ensemble were applied to i) to understand the major factors affecting phosphorus removal in aWWTP processes, and ii) to study the potentialities of a novel $\nu$-SVM ensemble to forecast extremely low levels of phosphorus based on parameters like metal salt dosing. The last goal was also based on the good performance of $\nu$-SVM in prediction tasks with data limitation, then, an ensemble built from a sequence of $\nu$-SVM is proposed in this chapter.

## 7.2 Methodology

### 7.2.1 Process description

Phosphorus removal employing aWWTP after a secondary effluent (SE) from a WWTP in the south of Hesse (Germany) was explored for almost two years of operation. The evaluated filtration techniques were a cloth filter (CF) from Mecana Umwelttechnik GmbH and a micro-filtration/ultra-filtration (MF/UF) unit from Pall Corporation, which were operated simultaneously with the addition of two types of metal salts; aluminium-(III)-chloride ($AlCl_3$) and ferric-(III)-chloride ($FeCl_3$), see Figure 7.1a (Fundneider et al., 2019).

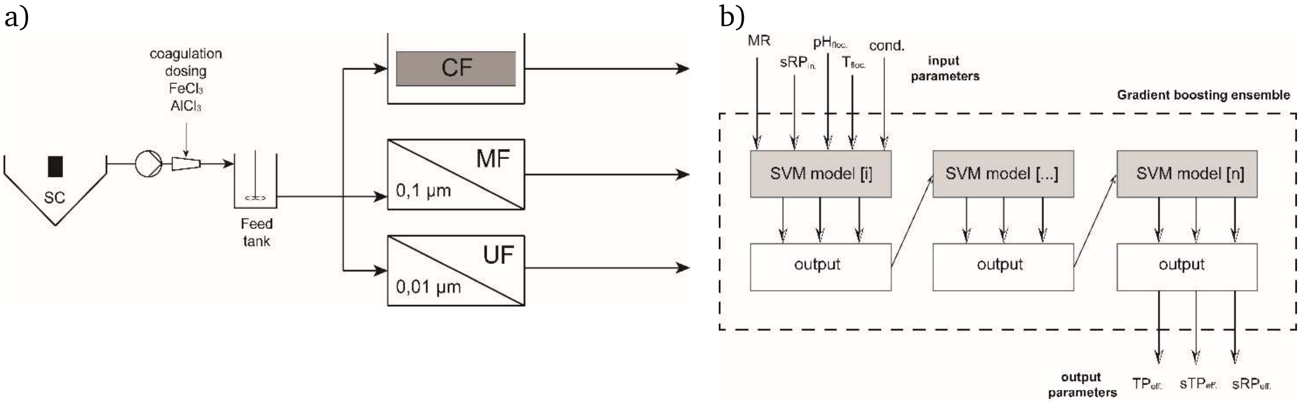


Figure 7. 1 a) Process diagram of the phosphorus removal in advanced wastewater treatment after the secondary clarifier (SC) and b) ensemble model architecture based on $\nu$-SVM – to forecast; sRP, TP and sTP in the effluent. Input parameters from the feature selection were: moral ratio of coagulants (MR), sRP, pH and temperature of flocculation and conductivity (cond.), these parameters were the best *predictors*.

Due to the high complexity of the process (i.e. alternating application of different metal salts and filtration technologies), data-driven methods were evaluated to analyze and model the removal of three phosphorus parameters: total phosphorus (TP), soluble TP (sTP) and soluble reactive phosphorus (sRP). Based on the operational information gathered, a dataset with over two years of operation was used for analysis. The feature or operational parameters were; moles of metal salt dosing per moles of sRP removed (MR), pH and temperature of flocculation (pH floc., T floc.), conductivity (cond), acid capacity (AC), turbidity (TB), soluble organic matter (sCOD), sRP, sTP and TP in both the influent (in.) and effluent (eff.) of the aWWTP. The pH in the flocculation stage was on average 6.94 ± 0.23 ($n_{obs.}$ = 344 d), and the influent concentrations of TP and sRP were 0.62 ± 0.15 mg $l^{-1}$ ($n_{obs.}$ = 344 d) and 0.41 ± 0.1 mg $l^{-1}$ ($n_{obs.}$ = 344 d), respectively.

### 7.2.2 Data-driven methods for extraction of knowledge

In supervised ML, feature selection allows dimensionality reduction in a dataset while selecting a subset of the most relevant features. The relevance of a feature is understood as the strong interactions between this feature (input) and output feature within a dataset. In a prediction task, the relevant parameters will share strong interactions between input and output features. The main benefits of feature selection is the reduction of the bias while reducing the noise that irrelevant feature have on supervised ML models, as seen in Chapter 5 (Blum and Langley, 1997). There are three main approaches to feature selection in supervised ML; *wrapper, filter* and *embedded methods* (Chandrashekar and Sahin, 2014) (Chapter 3, section 3.6). In this work, we applied *embedded methods* which combine both *wrapper* and *filter* characteristics. Random forest is a bagging ensemble method for classification and regression (Breiman, 2001). At the same time, the interactions between input and output features are identified through random forest (*embedded method*), which allowed to identify the strong interactions between input parameters and the concentration of phosphorus in the effluent of the aWWTP. The *feature importance* is a measurement of frequency and position of each predicting feature (input) in the decision trees that compose the random forest. The higher the value of *feature importance* of a feature, the better the capabilities of this feature to predict the desired output (sRP, TP and sTP, in this chapter). The size of the dataset involved 15 features (from the influent, process and effluent) with a total amount of 344 data points (Figure 7.1a). The interactions between these parameters and the phosphorus concentrations in the effluent; sRP, TP and sTP, were studied through feature selection and the *feature importance* measurement.

In the second part of this study, a novel ensemble model based on $\nu$-SVM was built to predict the phosphorus concentration in the effluent of the aWWTP; sRP, TP and sTP. In supervised ML a group of *weak learners* is called and ensemble; thus, this technique is called *ensemble learning*, and an *ensemble learning* algorithm is called an ensemble method. Gradient boosting is a type of ensemble method that combines several weak learners into a strong learner. The general idea of most boosting methods is to train *predictors* sequentially, and specifically in gradient boosting, the adding of *predictors* to the sequence tries to fit the new *predictor* to the residual errors made by the previous *predictor*. The gradient boosting ensemble for multiple outputs is not implemented in any library of Python or R, therefore, this gradient boosting ensemble was programmed from scratch. To evaluate the performance of the created ensemble, this method was compared to a single $\nu$-SVM model. In addition, to elucidate the importance of the training size in some ML methods, convolutional neural networks (1D-CNN) was also evaluated, this method was only applied to demonstrate this statement. For both the $\nu$-SVM ensemble and the single $\nu$-SVM, the resulting features from feature selection were the input to the model, while in 1D-CNN, no previous feature selection was applied. 1D-CNN have the characteristic of extracting the most relevant features in the initial layers, in subsequent layers the irrelevant parameters are dropped (LeCun et al., 2015). The training and validation datasets were split following a criteria of k-fold cross validation with k = 5 (Browne, 2000).

## 7.3 Results and discussion

In this work, the interactions of different parameters in aWWTP with the effluent concentration of phosphorus and the prediction of very low levels of phosphorus achieved in the process were studied and modeled through ensemble learning methods. The dataset used for this analysis was comprised by around 15 parameters and two years of operation, during this time two metal salts; $FeCl_3$ and $AlCl_3$, were dosed for chemical phosphorus removal. Figure 7.2a illustrates the parameters studied in this work; MR, AC, TB, sRP, sCOD, TP and sTP in both influent and effluent streams, T floc., pH floc. and cond. Before feature selection, pre-processing steps were conducted; multiple imputation (less than 10% of the data was imputed) and normalization of the data. Afterwards, the *feature importance* was extracted from a random forest model composed by 300 decision trees. The depth of the trees in random forest were limited to prevent overfitting. Figure 7.2b shows the *feature importance* ranking which was built to study the interactions between input feature (listed in the ranking) and output features; sRP, TP and sTP effluent concentrations. The features position in the ranking indicate the quality of interactions; the higher the position in the list, the stronger the interactions between input and output parameters. In particular, the *feature importance* measurement in random forest focus on two aspects; the frequency an input parameter is found in the forest and the position in each individual tree. To explore these interactions, the top 8 parameters in the ranking were studied further.

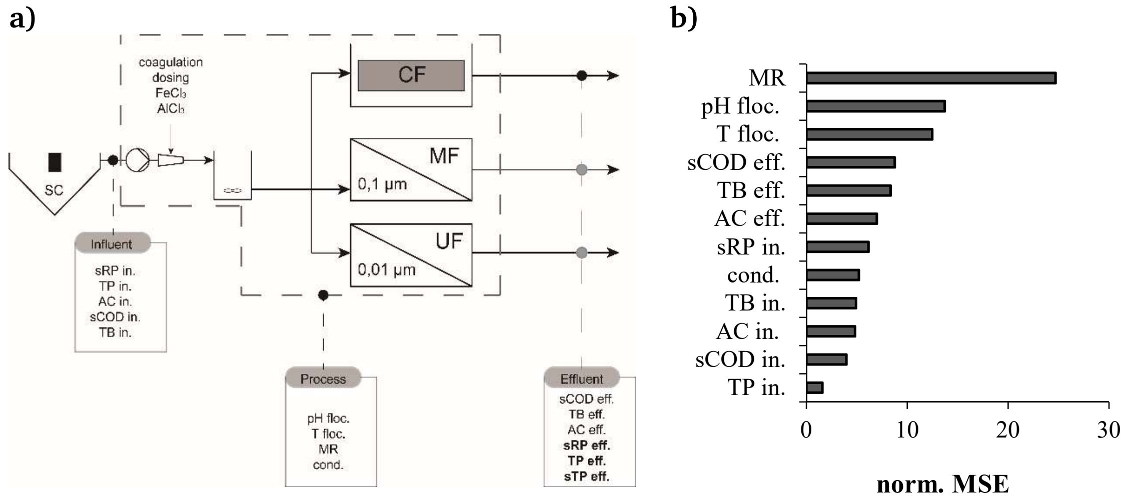


Figure 7. 2 Feature selection (FS) process: a) parameters considered in FS (influent, process and effluent), in bold the desired parameters to be modelled; effluent (eff.) sRP , TP and sTP b) Feature importance measurement extracted from random forest (RF), norm. MSE: normalized mean squared error. The best predictors are displayed at the top of the ranking.

The analysis through the *feature importance* measurement agrees with the literature. Szabó et al., (2008) studied design and operation factors in chemical phosphorus removal with different salts, including FeCl$_3$ and AlCl$_3$, the experimental results reported that phosphorus removal efficiency is greatly affected by pH, alkalinity, metal dose, metal type, initial and residual phosphate concentration, mixing, reaction time, age of flocs, and organic content of wastewater. Without the application of complex mechanistic methods, the *feature importance* ranking was able to find the same key parameters and to rank them. In this particular process, the degree of sRP removal was mainly controlled by: MR, pH, temperature and organic matter concentration (sCOD) (Bratby, 2016; Fundneider et al., 2019). Figure 7.3 shows the distribution of the top 8 parameters in the ranking, the results suggest that no significant difference exists in phosphorus removal with the application of one or the other metal salt, these results agree with Szabó et al., (2008), where very low phosphorus concentrations were achieved for both aluminium and iron salts with a broad pH range; 5 to 7.

As well, Yang et al., (2010) reported the advantages of aluminium and iron salts over calcium salts since they are not sensitive to pH.

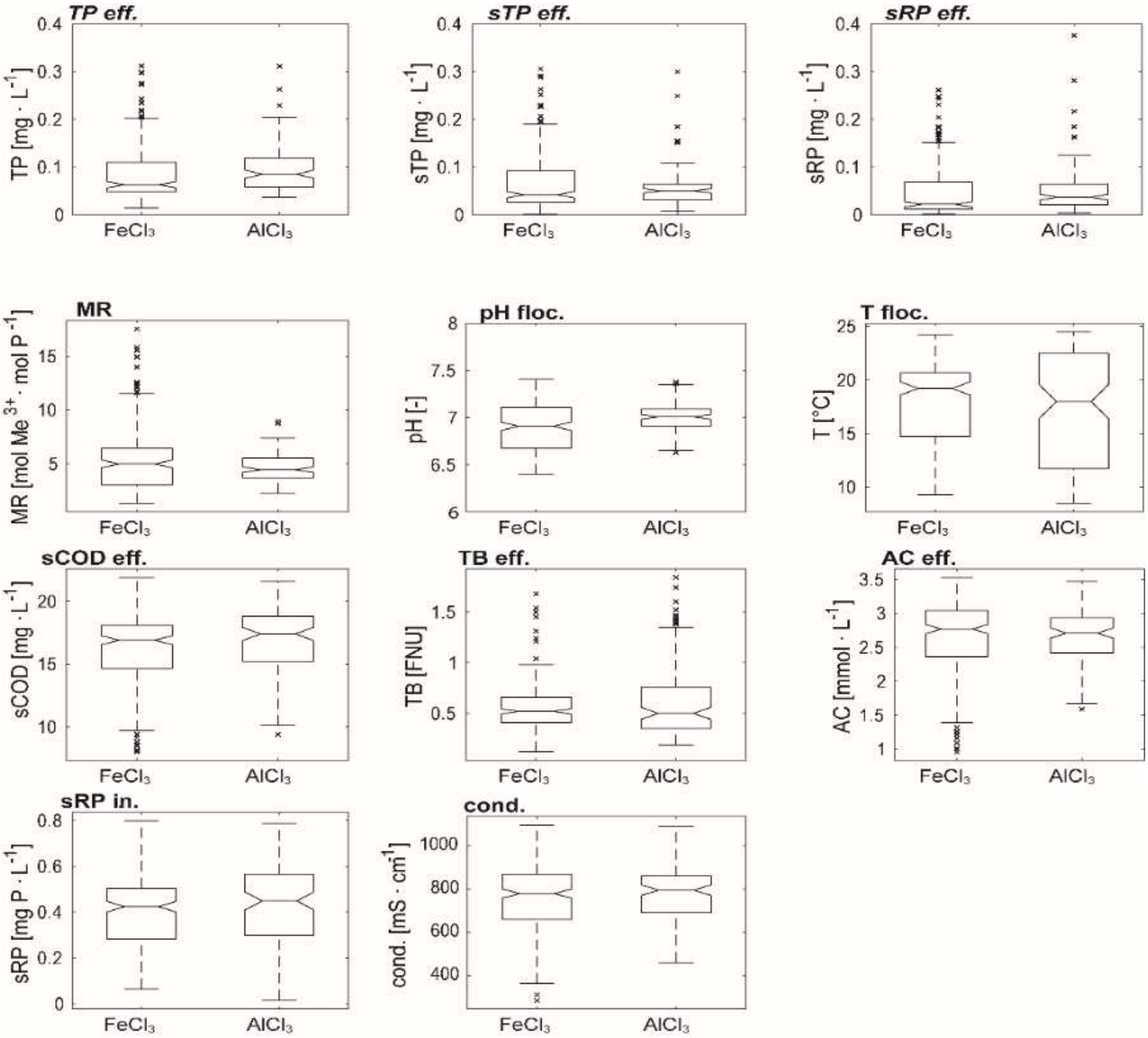
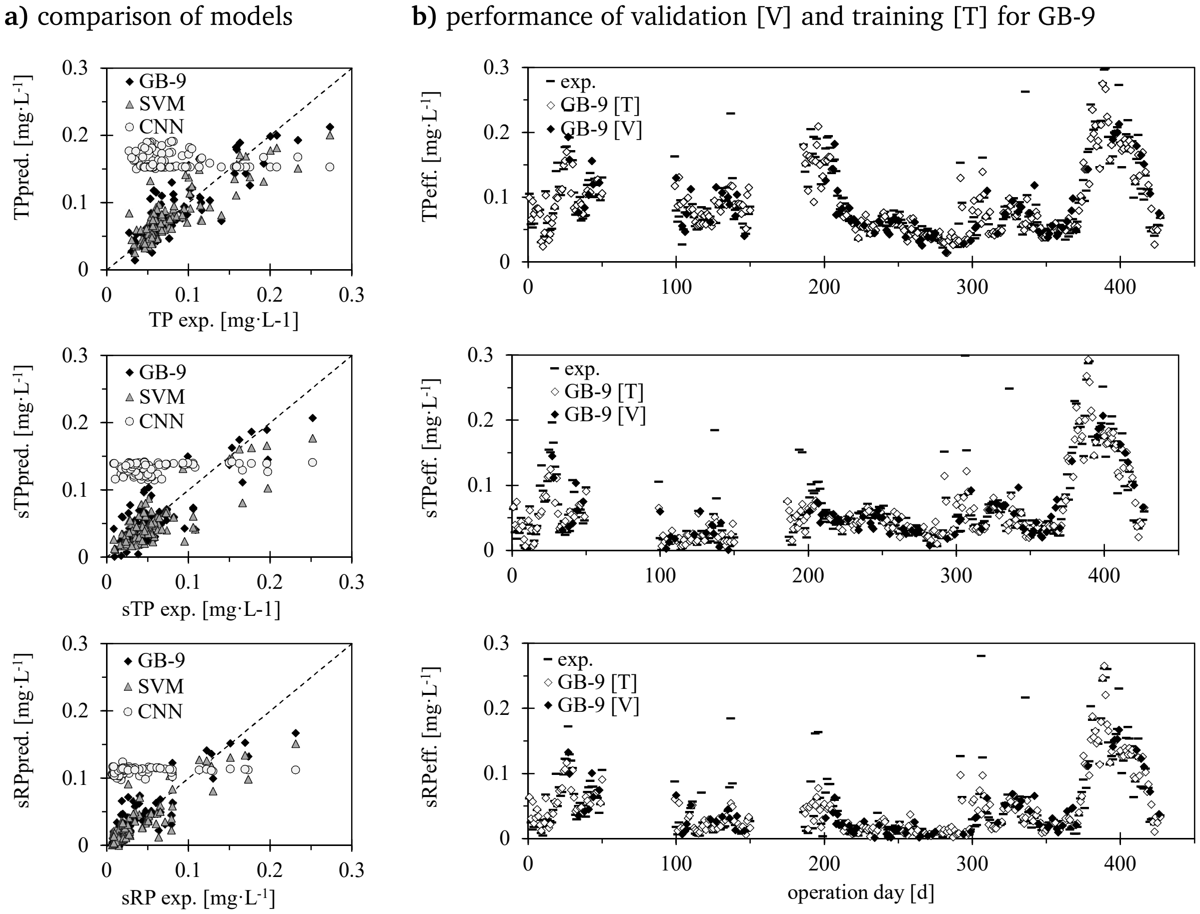


Figure 7. 3 Relevant parameters found in feature selection (FS). The boxplots show the distribution and outliers (x) of these parameters for both the coagulants FeCl3 (Fe) and AlCl3 (Al). The first row (in italic) shows the effluent concentrations studied; TP, sTP and sRP, and the following rows, the parameters found in FS.

Afterwards, ensemble methods, were evaluated for the prediction of the phosphorus species in the effluent (TP, sTP and sRP), the size of the training dataset was 258 out of 344 points (see Figure 7.4a), the remaining data points (86) were used for validation. For comparison, a single $\nu$-SVM was also studied. In addition, to elucidate the importance of the training size in more advanced ML methods such as 1D-CNN was also evaluated. From the *feature importance* ranking, the input features to the models were; MR, pH floc., T floc., sRP in. and conductivity (cond.). The effluent parameters (sCOD eff., TB eff., AC eff.,) were not considered as *predictors*, since these features would be unknown if the models are applied for further forecast tasks of phosphorus removal. The gradient boosting ensemble was built based on single $\nu$-SVM models added up, the number of $\nu$-SVM in the ensemble that resulted in the highest squared correlation coefficient was 9. In order to prevent overfitting, the $\nu$ value was restricted to a value of 0.5 ($\nu \in [0,1]$), the $\nu$ parameters is understood as an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors relative to the total number of training examples (Chang and Lin, 2002). Figure 7.4 illustrates the results obtained with the $\nu$-SVM gradient boosting ensemble (GB-9), compared to a single $\nu$-SVM and 1D-CNN.


**a)** comparison of models

**b)** performance of validation [V] and training [T] for GB-9

Figure 7. 4 Results obtained from the prediction of phosphorus species after advanced wastewater treatment. (a) performance of validation [V] and training [T] for GB-9 (b) comparison of models. SVM: single $\nu$-SVM method; GB-9: ensemble composed by 9 SVM models; CNN: convolutional neural networks

Both $\nu$-SVM gradient boosting ensemble and a single $\nu$-SVM models outperformed 1D-CNN in the prediction task. In particular, a single $\nu$-SVM achieved an average Pearson squared correlation coefficient ($R^2$) of 0.84 in the validation stage (Figure 7.4a) compared to the average $R^2$ of 0.056 for 1D-CNN, which demonstrates the importance of the training size to build accurate 1D-CNN. The amount of data for training and the variability of the data influences the accuracy of some methods in supervised ML, such as 1D-CNN. The amount of parameters to adjust these networks is higher, and therefore, more examples provide more training iterations and more weights can be updated. In this chapter, the simplest configuration of 1D-CNN consisted of 2,693 parameters. However, the size of the training dataset was 258 data points. As a result, the performance of the 1D-CNN was poorer than for the $\nu$-SVM, demonstrating the limitations of applying 1D-CNN for these type of problems where; i) there is data limitation and ii) the system is highly dynamic.

The $R^2$ obtained in the validation stage with a single $\nu$-SVM for the three phosphorus species; TP, sTP and sRP were 0.85, 0.81 and 0.87, respectively. Figure 7.4a shows the results from the performance of a single $\nu$-SVM model and an ensemble of 9 $\nu$-SVM models in gradient boosting. On average, the $R^2$ between the experiment and the model predictions demonstrated that the ensemble is more accurate than a single $\nu$-SVM achieving an average $R^2$ of 0.90 in the validation stage. The average improvement in the $R^2$ with the application of the ensemble was 5% higher than a single $\nu$-SVM. However, when considering only levels of TP lower than 0.1 mg $l^{-1}$, both $\nu$-SVM and $\nu$-SVM ensemble achieved $R^2$ of 0.6 and 0.7, respectively. Therefore, to further exploit the potentialities of the gradient boosting ensemble built in this work, a second gradient boosting ensemble was built and additionally (similar to Chapter 5), feature engineering was explored to evaluate the possibility to create better *predictors*. For this purpose, the original dataset was filtered so that the maximum TP concentration was 0.1 mg $l^{-1}$, the resulted dataset comprised 240 data points or tuples. Through feature engineering the new features created were; MR:T, MR:pH, MR:T:cond., sRP in.:T, sRP in.:MR and MR:sRP in. (mainly ratios within the previous selected *predictors*). Figure 7.5 summarizes the *feature importance* ranking with the new *predictors*. Similar to the results found in Chapter 5, the created features appear to be better *predictors* than the previously selected.

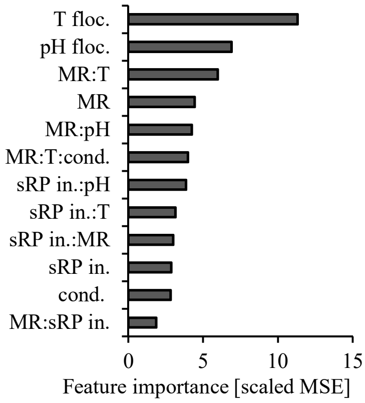


Figure 7. 5 Feature importance ranking extracted from a random forest for prediction of TP, sRP and sTP.

Figure 7.6 shows the results obtained from the $\nu$-SVM gradient boosting ensemble and a single $\nu$-SVM for the selective prediction of TP concentrations lower than 0.1 mg $l^{-1.}$.

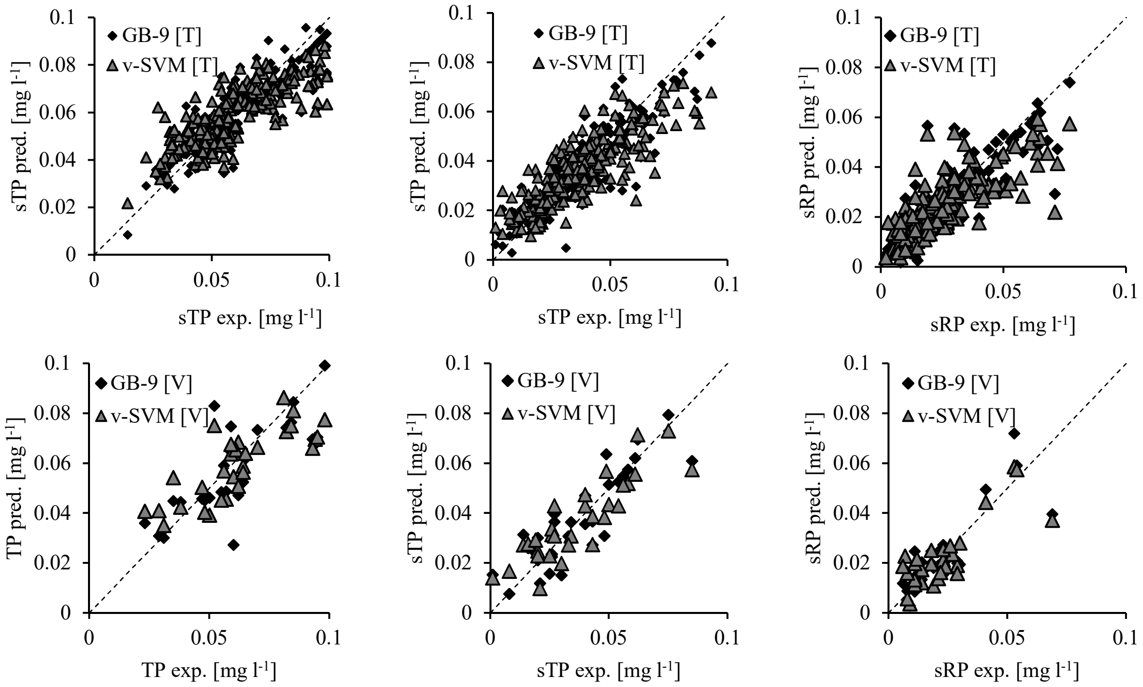


Figure 7. 6 Results obtained for selective models for a threshold concentrations lower than 0.1 mg l-1. Top: results from training (T), bottom: results obtained in validation (V).

The gradient boosting ensemble was highly accurate in the training stage, achieving a $R^2$=0.9 and only 0.8 for a single $\nu$-SVM. However, the performance in the validation set was comparable, since both methods achieved an average $R^2$ of 0.82, where the gradient boosting ensemble was better for the prediction of sRP in the effluent ($R^2$=0.83) whereas a single $\nu$-SVM achieved higher performance in the prediction of TP ($R^2$=0.80). The selective models were able to improve by more than 10% the performance of the previous results, thus, demonstrating the feasibility of ensemble learning and $\nu$-SVM to achieve the prediction of very low levels of phosphorus removal in aWWTP.

This work proves the potential use of ensemble learning to evaluate and model the quality standards for phosphorus removal in aWWTP, with both feature selection and GB-9 model. Current models for prediction of phosphorus removal with in-line coagulation are highly complex and are focused mainly on the application of iron (Fytianos et al., 1998; Hauduc et al., 2015; Smith et al., 2008). In particular, the model developed by Hauduc et al., (2015) showed the capability to accurately predict phosphate adsorption for initial phosphate concentration as low as 1 mg $l^{-1}$. However, the application of this model for lower initial phosphate concentrations would require an evaluation with additional experiments. In general, there are two important issues regarding the application of mechanistic models for the prediction chemical phosphorus removal;

- First, current models are capable to model the initial fast removal process (less than one minute) but poorly the further slow removal of phosphorus (Hauduc et al., 2015).
- Second, the models assume a homogeneous system with rapid mixing and fast initial consumption of the metal salts, which were achieved successfully in laboratory experiments, but are not representative of real WWTP. The mean velocity gradient at the dosage point in real WWTP are in the order of 20 to 100 s-1 compared to 300-1000 s-1 in experimental studies (Szabó et al., 2008).

From above, current laboratory results obtained from experiments cannot be directly applied to implement models for full-scale plant effluent phosphorus concentrations, as mixing and hydraulics will play a large role at these low concentrations. Rapid mixing is essential to disperse metal salts uniformly i.e. allow adequate contact between the coagulant and particles. However, a model that integrates mixing in the chemical removal of phosphorus is not yet explored. Several works on computational fluid dynamics (CFD) have studied mixing issues in activated sludge processes (Karpinska and Bridgeman, 2016). Chemical phosphorus removal could benefit from an integrated model of CFD-kinetics to better understand this complex process.

The approach presented in this chapter is helpful to relate the influent and effluent parameters in advanced wastewater treatment, and may be applied for further design and operation of such processes, however similar to mechanistic models, mixing was not integrated to this approach and thus the influence was not evaluated. Several wastewater treatment processes may benefit from the application of ML methods, however, as discussed already, the training dataset size is relevant for the performance of these methods. The application of ensemble methods suggests the feasibility to achieve higher accuracy when modeling highly dynamic and complex processes where data limitation exists. Currently, only one work in literature has evaluated the application of ensemble methods in wastewater treatment (Nourani et al., 2018). Moreover, an alternative tool to improve our understanding of the process was explored through feature selection, which allowed the extraction of the most relevant parameters for the removal of TP, sTP and sRP, i.e. MR, sRP concentration in the influent, cond, T and pH in the flocculation stage. The results suggest that our approach is an effective tool to understand and model the process, saving time and resources compared to deterministic models.

## 7.4 Chapter conclusions

Two ensemble learning methods were explored to understand in a comprehensive manner the major factors affecting phosphorus removal and model extreme low levels of phosphorus in aWWTP. In ML, ensemble methods are composed by a group of *predictors* which usually perform better than a single *predictor*, thus: *wisdom of the crowd*. Random forest, a bagging ensemble method, was first applied to identify strong interactions between; total phosphorus (TP), soluble TP (sTP) and soluble reactive phosphorus (sRP) effluent concentrations, and other process parameters. The results suggested that the molar ratio (MR) of coagulants, pH and T of flocculation were by far the most influencing parameters for phosphorus removal, results that agree with the literature. Afterwards, TP, sTP and sRP concentrations in the effluent were modelled. Gradient boosting ensemble, was evaluated. The gradient boosting algorithm was built from a sequence of nine $\nu$–SVM which was highly accurate to model the phosphorus species (average $R_{GB\text{-}SVM}^2 = 0.90$), this ensemble was compared to a single $\nu$-SVM. However, the $\nu$-SVM ensemble just improved the predictive capability of a single $\nu$-SVM by 5%, when the $\nu$-SVM ensemble was evaluated only in the prediction of levels of TP lower than 0.1 mg $l^{-1}$, the ensemble outperformed the singe $\nu$-SVM by 10 %, thus validating the advantages of ensemble learning with limited data.

# 8 General conclusions, contributions and perspectives

## 8.1 Conclusions

AI is changing our lives, is rapidly evolving, is influencing our way of living, from how we communicate to intelligent prosthesis and humanoids. ML methods are used at profiling users, improve diagnosis, preventing crime, mapping the social behavior, etc. The biological complexity of the processes in wastewater treatment together with the high amount of data generated has motivated researchers to apply and learn from these methods, search for solutions to the problem of obtaining useful insights, predictions and decisions from datasets.

However, when ML methods are applied, it is essential to characterize the sources of data in each field, also in wastewater treatment. In bWWTP, three main sources of data exist; on-line data from sensors, off-line data from laboratories and on/off data from equipment. These data sources have different time intervals of sampling which results in highly heterogeneous datasets; some with low and some with high density of information. Thus, the methodology has to be adapted to each particular case. Current research has focused mainly on predictive tasks, to forecast the effluent composition and performance of different bWWTP. However, none of these studies addresses the issue of heterogeneity of the datasets. Most of the studies in the literature related to data-driven methods in wastewater treatment do not explicitly report pre-processing steps or nature/sources of data. For proper data analysis, it is fundamental to know the data sources, selection and pre-processing workflows in order to evaluate the extents of the results and limitations of the approach. In-depth analysis of the current literature allowed to formulate research questions (RQ) that aimed at filling gaps in the field of ML applications in wastewater treatment. These RQs guided the development of this work and are answered below.

RQ1: *It is essential to distinguish and clarify the differences and goals of activated sludge models (ASM-type) models and ML-based tasks in the current framework of wastewater treatment. What is the state of the art towards the application of data-driven methods in the water sector compared to mechanistic approaches? Which are the limitations of both approaches in the water sector?*

The application of mathematical models has started in the beginning of the 1980's, and ever since, several modifications to the initial ASM1, ASM2 and ASM3 were proposed with the aim to overcome limitations identified in these models. From a biochemical perspective, these upgrades range from one to two step N-DN, the integration of $N_2O$ kinetics and Anammox in ASM1 and ASM3, to account for the kinetics of GAOs and denitrifying PAOs in EBPR, e.g. in ASM2 type models. However, the current framework on mechanistic modeling in wastewater treatment faces several challenges. The initial ASM-type models were developed to model a conventional activated sludge process. New technologies, such as membranes and biofilm systems, suggested the adaptation of the ASM models to these new technologies, adding complexity to the modeling process. The literature had vastly addressed key issues on the adaptation of ASM-type models in these systems and other configurations. In membrane systems, not only the biokinetics are different, but the mechanical processes involved; the filtration process, backwashing and fouling (Fenu et al., 2010; Naessens et al., 2012a, 2012b), the characterization of the biomass is crucial for building a realistic model when based on ASM type model (Cosenza et al., 2013; Jiang et al., 2009). On the other hand, in biofilm systems, a key drawback on the implementation of ASM-type models is the oversimplification of mass transfer. In specific, the oxygen diffusion is a process that is usually oversimplified in the ASM models, leading to a misinterpretation of "optimal" DO oxygen concentrations in biofilm systems where diffusion process is a limiting kinetic process. Recent studies suggest the need of more detailed models that consider floc/size distribution. Still today, a general problem of ASM type models is their difficulty to adjust a significant set of kinetic and stoichiometric parameters for a respective system. The settings of the kinetic parameters should be characteristic to each biochemical process and wastewaters (industrial or municipal). However, to achieve an accurate characteristic model, parallel experiments are necessary to determine kinetic and stoichiometric parameters, which is time consuming and expensive.

On the other hand, the application of data-driven methods based on machine learning in wastewater treatment has focused mainly on prediction of influent and effluent composition. The main reasons that moved the wastewater treatment community to apply these methods in predictive tasks are two-fold; i) is the availability of data gathered from monitoring different bWWTP and ii) the already mentioned complexity of biological processes. The high adaptability of data-driven methods based on ML to dynamic systems has driven the research community to a wide application of these methods. However, a key issue emerges from the reviewed studies that partially set the basis for the main hypothesis in this work. One of the issues is the methodology. The current studies related to data-driven methods in wastewater treatment do not explicitly describe the pre-processing of the data used for analysis; was there outlier removal? which was the frequency considered for the data selection? and the rationale behind the selection of the methods and the input parameters for prediction tasks are not clearly founded. Second, is the rationale behind the selection the data for analysis. The majority of the studies use similar input parameters to those used in ASM-type models, ignoring the potential use of other parameters which are monitored in any bWWTP and not necessarily implemented in the mechanistic models; ORP, conductivity, turbidity, and in general, data from sensors. Third, the diversity of data sources in wastewater treatment is clear. The potentialities of data-driven methods based on machine learning is the extraction of valuable knowledge from data, yet, not sufficiently explored. The combination of these data sources for extraction of knowledge is not yet studied in bWWTP.

RQ2: *The data sources in wastewater treatment have different natures; online from sensors, on/off data from equipment and off-line data from laboratories. For one particular period of operation, the amount of information gathered from these sources is significantly different resulting in a heterogeneous dataset; the amount of data points differs from parameter to parameter. How sensitive are the data-driven methods to the amount of data and parameters considered for analysis? How the results from a data-driven task will change with different sizes of data?*

This question was addressed in Chapter 4. In general, in this thesis, heterogeneous datasets from different biological wastewater treatment processes were studied. By heterogeneous we understand that these datasets were composed by online and offline monitored parameters or features. In particular, in Chapter 4 a heterogeneous dataset from a municipal WWTP was studied. The raw initial dataset was partitioned or segmented into subsets, each subset contained different amount of data points or observations and different amount of parameters. The motivation behind the segmentation was to have a wider understanding of the process and most important, to show the influence of data size; data points and features, on the outcome of different data-driven methods such as clustering and feature selection. The results clearly showed that an arbitrary selection of a subset out of the raw initial dataset (given that the initial raw data is heterogeneous) would lead to unreliable results, thus, data-driven methods are very sensitive to the amount of data considered for analysis. In k-means clustering, the results from the different subsets ended in different amount of clusters. In feature selection, the selected features were different from subset to subset. No clear evidence existed on which subset could be significant from the system. This last argument highlights the importance of selecting a representative data set to obtain reliable results when data-driven methods are applied in wastewater treatment and why the methodology of data processing description is important.

RQ3: *Following RQ2. Which subset from the total raw data collected would be the most significant for further data-driven tasks? How this subset can be selected? Is it possible to optimize both the parameters considered for prediction (input to the model) and the amount of information (size of the dataset)?*

This question is addressed in Chapter 5. In this chapter, heterogeneous datasets coming from two full-scale PN-A systems are studied. A methodology to select a significant dataset out of the total raw data was proposed. The selection process was based on the definition of a score function subject to conditions which would allow to select a characteristic dataset out of the total raw data. The selection process was towards finding the best set of features to predict the effluent of the full-scale systems. The results show that the optimized dataset achieved highly accurate results when applied to predict the ammonium and nitrate concentrations in the effluent of the systems. Additionally, to show the effectivity of the optimization process, different case-scenarios were evaluated. These case-scenarios not only proved that the optimized subset was better for prediction of the effluent but further techniques were explored. Feature engineering was demonstrated to be a potential tool to create new *predictors* out of the available features to create highly accurate models, however, the application of engineered features to build predictive models in bWWTP has been overlooked, and thus, an important contribution of this work.

RQ4: *Following RQ3. When dealing with heterogeneous datasets. How these datasets can be combined with even smaller datasets; in biological wastewater treatment processes a good example are biomass batch activity tests. How these different sources of data can be combined to create a significant dataset and which tools can be applied to extract knowledge from it?*

This question is addressed in Chapter 6. In this chapter, a new approach to combined different data sources from PN-A systems was developed. Three sources of datasets were studied; online (high density of information), laboratories (medium density of information) and batch activity datasets (low density of information). In a first part, feature selection was applied to reduce the dimensionality of the online and lab datasets. In the second part, the resulting subset was coupled to the small dataset to build the *clustering dataset*. k-means clustering was applied to extract patterns within this data. This methodology was designed to search relations between online and lab features with biomass activity information. The methodology proposed aided in the identification of operational periods when high and low biomass activity occurred. DO and ORP, were discovered as relevant parameters for monitoring AOB and AnAOB activity. Additionally, in SBRs, reaction times above 320 minutes, average DO concentrations of 0.3 mg $L^{-1}$, maximum DO concentrations (in the aeration periods) of 0.8 mg $L^{-1}$, and lower sedimentation times (9.1 minutes) favored high ammonia removal. The information obtained after the application of the new approach helped to extract important knowledge from combined data sources efficiently and that can certainly be applied to similar bWWTP.

RQ5: *Following RQ3. Given heterogeneous datasets from a bWWTP with limited amount of data, and additionally, a process that is not yet well studied from a mechanistic modeling perspective. How can ML methods be applied to extract knowledge and model this complex bWWTP?*

Ensemble learning is a powerful concept from ML that combines *weaker learners* to build a better model, thus, *wisdom of the crowd*. The different ensembles allowed to build highly accurate models in Chapter 4, 5 and 7 and to find the best *predictors* out of a high dimensional dataset (*feature selection* through random forest). The previous experiences motivated the application of ensemble learning to study an advanced wastewater treatment process, more specific, phosphorus removal after secondary clarification. This process has not yet been studied through a modeling perspective, or at least, not in detail. Therefore, in the first part of the study, feature selection was applied to extract relevant interactions between influent, process and different species of phosphorus in the effluent of the process. The results not only agreed with the literature but confirmed the ability of feature selection to study strong interactions in a dataset. In the second part of the study, gradient boosting ensemble which was built from a sequence of nine $\nu$-SVM, was applied to model the effluent composition of the aWWTP, other challenges with this dataset were; i) the limited amount of data to train, and ii) the values of the phosphorus concentrations which were near to zero mg $l^{-1}$. However, the gradient boosting ensemble of $\nu$-SVM, was proved to be effective and better than a single $\nu$-SVM for concentrations of TP lower than 0.1 mg $l^{-1}$.

## 8.2 Contributions

Overall, this work addressed the thorough data analysis of heterogeneous datasets in bWWTP, which is an important gap in the current framework of data-driven approaches in wastewater treatment. An appropriate methodology and suitability of different data-driven methods to approach these datasets was addressed and thoroughly discussed along this work. Listed below are the most important contributions of this work.

- The main contribution of the literature review was to evidence the advantages, disadvantages and challenges of ASM-type models and data-driven approaches. The differentiation between both approaches and current applications in wastewater treatment were addressed. Clearly, the methodology when dealing with data-driven approaches in the current framework presents several problems, mainly with the data processing and description of the tools applied for analysis.
- The importance of data selection in heterogeneous datasets was evidenced. In wastewater treatment, we rely on heterogeneous datasets since they comprise both online and lab data. Thus, to further analyze this information, the right tools and methods should be carefully applied to extract reliable information. The influence of the amount of data for analysis on the outcome of different data-driven methods was demonstrated in this work.
- A methodology to extract a significant subset out of the total raw data was developed. The definition of a score-function, allowed the optimization of a subset which was then used to build highly accurate models.
- The concept of feature engineering was introduced to build better *predictors* in predictive models for wastewater treatment. Although, feature engineering is a well-developed field in data-science, not yet explored in wastewater treatment. Engineered features allowed to build highly accurate models for the prediction of complex bWWTP.
- A methodology was introduced to combine different data sources; online, lab and batch activity datasets, to efficiently extract knowledge.

Ensemble learning was evaluated to extract knowledge and model extremely low levels of phosphorus in advanced wastewater treatment. This process is not yet well studied from a mechanistic modeling perspective. Therefore, the results obtained in this work are an important contribution since not only the models performance was good but additionally the knowledge extracted agreed with the experimental knowledge of the process.

## 8.3 Future perspectives

- One of the main limitations of this work was the application of the knowledge gained from past experiments. A combined effort to test the knowledge gained in actual ongoing processes is necessary. Additionally, the extension of the analysis to similar processes to evaluate if the knowledge gained in this work is particular to the processes studied here or similar patterns eco in comparable bWWTP.

In the context of data quality in wastewater treatment, two problems were identified.

- **Data quality problem I**: Missing data is an important problem in data sources in wastewater treatment, it can affect the accuracy of conclusions drawn as seen in Chapter 4. Although multiple imputation, has proven to be an effective tool to deal with missing values. However, in the context of wastewater treatment, no study has considered approaches to characterize patterns of bias in missing data. To determine the specific features that predict the *missingness* of data. The knowledge of the specific systematic bias patterns in the incidence of missing data in the water sector can help to assess the quality of the conclusions drawn from datasets with missing values.

- **Data quality problem II**: Same data sources may have an acceptable level of quality for some contexts but this quality may be unacceptable for other contexts. However, existing data quality metrics for specific data sources in wastewater treatment (except for sensor data) are not yet addressed or disconnected from the specific contextual characteristics. The need to revise data quality metrics for different sources of data in wastewater treatment is necessary. Especially after demonstrating the impact of the amount of data in different data-driven methods.

## 9 References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A System for Large-Scale Machine Learning. Presented at the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265–283.

Abbasi, M., Abduli, M.A., Omidvar, B., Baghvand, A., 2012. Forecasting Municipal Solid waste Generation by Hybrid Support Vector Machine and Partial Least Square Model. International Journal of Environmental Research 7, 27–38.

Acevedo, B., Borrás, L., Oehmen, A., Barat, R., 2014. Modelling the metabolic shift of polyphosphate-accumulating organisms. Water Research 65, 235–244. https://doi.org/10.1016/j.watres.2014.07.028

Adeogba, E., Barty, P., O'Dwyer, E., Guo, M., 2019. Waste-to-Resource Transformation: Gradient Boosting Modeling for Organic Fraction Municipal Solid Waste Projection. ACS Sustainable Chem. Eng. 7, 10460–10466. https://doi.org/10.1021/acssuschemeng.9b00821

Agrawal, S., 2018. Microbial community analysis during mainstream anaerobic ammonium oxidation (PhD Thesis). Technishe Universität Darmstadt.

Aguado, D., Montoya, T., Borras, L., Seco, A., Ferrer, J., 2008. Using SOM and PCA for analysing and interpreting data from a P-removal SBR. Engineering Applications of Artificial Intelligence 21, 919–930. https://doi.org/10.1016/j.engappai.2007.08.001

Aguado, D., Ribes, J., Montoya, T., Ferrer, J., Seco, A., 2009. A methodology for sequencing batch reactor identification with artificial neural networks: A case study. Computers & Chemical Engineering 33, 465–472. https://doi.org/10.1016/j.compchemeng.2008.10.018

Akratos, C., Papspyros, J., Tsihrintzis, A., 2008. An artificial neural network model and design equations for BOD and COD removal prediction in horizontal subsurface flow constructed wetlands. Chemical Engineering Journal 143, 96–110.

Aksoy, S., Haralick, R.M., 2001. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recognition Letters 22, 563–582. https://doi.org/10.1016/S0167-8655(00)00112-4

Alawi, M., Off, S., Kaya, M., Spieck, E., 2009. Temperature influences the population structure of nitrite-oxidizing bacteria in activated sludge. Environmental Microbiology Reports 1, 184–190. https://doi.org/10.1111/j.1758-2229.2009.00029.x

Alejo, L., Atkinson, J., Guzmán-Fierro, V., Roeckel, M., 2018. Effluent composition prediction of a two-stage anaerobic digestion process: machine learning and stoichiometry techniques. Environ Sci Pollut Res 1–15.

Altman, N.S., 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician 46, 175–185. https://doi.org/10.1080/00031305.1992.10475879

Ante, A., Besche, H., Voss, H., 1994. A Mathematical-Model for Enhanced Biological Phosphorus Removal. Water Sci. Technol. 30, 193–203.

Ardern, E., Lockett, W.T., 1914. Experiments on the oxidation of sewage without the aid of filters. Journal of the Society of Chemical Industry 33, 523–539. https://doi.org/10.1002/jctb.5000331005

Argaman, Y., Papkov, G., Ostfeld, A., Rubin, D., 1999. Single-Sludge Nitrogen Removal Model: Calibration and Verification. Journal of Environmental Engineering 125, 608–617. https://doi.org/10.1061/(ASCE)0733-9372(1999)125:7(608)

Asadi, A., Verma, A., Yang, K., Mejabi, B., 2017. Wastewater treatment aeration process optimization: A data mining approach. Journal of Environmental Management 203, 630–639. https://doi.org/10.1016/j.jenvman.2016.07.047

Bagheri, M., Mirbagheri, S.A., Bagheri, Z., Kamarkhani, A.M., 2015. Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach. Process Safety and Environmental Protection 95, 12–25. https://doi.org/10.1016/j.psep.2015.02.008

Batchelor, B., 1983. Simulation of Single-Sludge Nitrogen Removal. Journal of Environmental Engineering 109, 1–16. https://doi.org/10.1061/(ASCE)0733-9372(1983)109:1(1)

Bi, Z., Takekawa, M., Park, G., Soda, S., Zhou, J., Qiao, S., Ike, M., 2015. Effects of the C/N ratio and bacterial populations on nitrogen removal in the simultaneous anammox and heterotrophic denitrification process: Mathematic modeling and batch experiments. Chemical Engineering Journal 280, 606–613. https://doi.org/10.1016/j.cej.2015.06.028

Bidstrup, S.M., Grady, C.P.L., 1988. SSSP: Simulation of Single-Sludge Processes. Journal (Water Pollution Control Federation) 60, 351–361.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Bissonette, J.A., 1999. Small sample size problems in wildlife ecology: a contingent analytical approach. wbio 5, 65–71. https://doi.org/10.2981/wlb.1999.010

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence, Relevance 97, 245–271. https://doi.org/10.1016/S0004-3702(97)00063-5

Boniecki, P., Dach, J., Pilarski, K., Piekarska-Boniecka, H., 2012. Artificial neural networks for modeling ammonia emissions released from sewage sludge composting. Atmospheric Environment 57, 49–54. https://doi.org/10.1016/j.atmosenv.2012.04.036

Bousquet, O., Luxburg, U. von, Rätsch, G., 2011. Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures. Springer.

Bratby, J., 2016. Coagulation and Flocculation in Water and Wastewater Treatment Second Edition | IWA Publishing, 3rd ed. IWA Publishing, London.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123–140. https://doi.org/10.1007/BF00058655

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Taylor & Francis.

Brockmann, D., Morgenroth, E., 2010. Evaluating operating conditions for outcompeting nitrite oxidizers and maintaining partial nitrification in biofilm systems using biofilm modeling and Monte Carlo filtering. Water Research 44, 1995–2009.

Browne, M.W., 2000. Cross-Validation Methods. Journal of Mathematical Psychology 44, 108–132. https://doi.org/10.1006/jmps.1999.1279

Bürger, R., Careaga, J., Diehl, S., Mejías, C., Nopens, I., Torfs, E., Vanrolleghem, P.A., 2016. Simulations of reactive settling of activated sludge with a reduced biokinetic model. Computers & Chemical Engineering 92, 216–229. https://doi.org/10.1016/j.compchemeng.2016.04.037

Bürger, R., Diehl, S., Nopens, I., 2011. A consistent modelling methodology for secondary settling tanks in wastewater treatment. Water Research 45, 2247–2260. https://doi.org/10.1016/j.watres.2011.01.020

Burrell, P., Keller, J., Blackall, L.L., 1999. Characterisation of the bacterial consortium involved in nitrite oxidation in activated sludge. Water Science and Technology, Biological Nutrient Removal 39, 45–52. https://doi.org/10.1016/S0273-1223(99)00121-3

Buuren, S. van, Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45, 1–67. https://doi.org/10.18637/jss.v045.i03

Cai, J., Xu, K., Zhu, Y., Hu, F., Li, L., 2020. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Applied Energy 262, 114566. https://doi.org/10.1016/j.apenergy.2020.114566

Carpenter, J.R., Kenward, M.G., 2012. The Multiple Imputation Procedure and its Justification, in: Multiple Imputation and Its Application. John Wiley & Sons, Ltd, pp. 37–73. https://doi.org/10.1002/9781119942283.ch2

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40th-year commemorative issue 40, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chang, C.-C., Lin, C.-J., 2002. Training v-Support Vector Regression: Theory and Algorithms. Neural Computation 14, 1959–1977. https://doi.org/10.1162/089976602760128081

Cheremisinoff, N.P., 1997. 4 - Nitrification and denitrification in the activated sludge process, in: Cheremisinoff, N.P. (Ed.), Biotechnology for Waste and Wastewater Treatment. William Andrew Publishing, Westwood, NJ, pp. 151–188. https://doi.org/10.1016/B978-081551409-1.50006-6

Chollet, F., 2015. Keras: Deep learning library for theano and tensorflow. URL: https://keras. io/k 7, T1.

Clark, N.R., Ma'ayan, A., 2011. Introduction to Statistical Methods to Analyze Large Data Sets: Principal Components Analysis. Sci. Signal. 4, tr3–tr3. https://doi.org/10.1126/scisignal.2001967

Corbalá-Robles, L., Picioreanu, C., van Loosdrecht, M.C.M., Pérez, J., 2016. Analysing the effects of the aeration pattern and residual ammonium concentration in a partial nitritation-anammox process. Environmental Technology 37, 694–702. https://doi.org/10.1080/09593330.2015.1077895

Corominas, Ll., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., Poch, M., 2018. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental Modelling & Software, Special Issue on Environmental Data Science. Applications to Air quality and Water cycle 106, 89–103. https://doi.org/10.1016/j.envsoft.2017.11.023

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Machine Learning 20, 273–297. https://doi.org/10.1023/A:1022627411411

Cosenza, A., Mannina, G., Neumann, M.B., Viviani, G., Vanrolleghem, P.A., 2013. Biological nitrogen and phosphorus removal in membrane bioreactors: model development and parameter estimation. Bioprocess. Biosyst. Eng. 36, 499–514. https://doi.org/10.1007/s00449-012-0806-1

Côté, M., Grandjean, B.P.A., Lessard, P., Thibault, J., 1995. Dynamic modelling of the activated sludge process: Improving prediction using neural networks. Water Research 29, 995–1004. https://doi.org/10.1016/0043-1354(95)93250-W

Dapena-Mora, A., Hulle, S.W.V., Campos, J.L., Méndez, R., Vanrolleghem, P.A., Jetten, M., 2004. Enrichment of Anammox biomass from municipal activated sludge: experimental and modelling results. Journal of Chemical Technology & Biotechnology 79, 1421–1428. https://doi.org/10.1002/jctb.1148

Di, Z., Chang, M., Guo, P., Li, Y., Chang, Y., 2019. Using Real-Time Data and Unsupervised Machine Learning Techniques to Study Large-Scale Spatio–Temporal Characteristics of Wastewater Discharges and their Influence on Surface Water Quality in the Yangtze River Basin. Water 11, 1268. https://doi.org/10.3390/w11061268

Dieu, B., 2001. Application of the SCADA system in wastewater treatment plants. ISA Transactions 40, 267–281. https://doi.org/10.1016/S0019-0578(00)00053-7

Dixon, M., Gallop, J., Lambert, S., Healy, J., 2007. Experience with data mining for the anaerobic wastewater treatment process. Environmental Modelling & Software 22, 315–322. https://doi.org/10.1016/j.envsoft.2005.07.031

Dold, P., Ekama, G., Marais, G., 1980. A General Model for the Activated Sludge Process. Progress in Water Technology 12, 47–77. https://doi.org/10.1016/b978-1-4832-8438-5.50010-8

Dold, P.L., Marais, G. v R., 1986. Evaluation of the General Activated Sludge Model Proposed by the IAWPRC Task Group. Water Sci Technol 18, 63–89. https://doi.org/10.2166/wst.1986.0061

Domingo-Félez, C., Smets, B.F., 2019. Modelling electron competition in denitrifying microbial communities through an analogy to electric circuits: simple is better, in: Oral Presentations Proceedings. Presented at the 10th IWA Symposium on Modelling and Integrated Assessment (Watermatex 2019), Copenhagen, Denmark, pp. 566–569.

Dürrenmatt, D.J., Gujer, W., 2012. Data-driven modeling approaches to support wastewater treatment plant operation. Environmental Modelling & Software 30, 47–56. https://doi.org/10.1016/j.envsoft.2011.11.007

Ecob, D., Williamson, J., Hughes, G., Davis, J., 1995. PLC's and SCADA - a water industry experience, in: IEE Colloquium on Application of Advanced PLC (Programmable Logic Controller) Systems with Specific Experiences from Water Treatment (Digest No.1995/112). Presented at the IEE Colloquium on Application of Advanced PLC (Programmable Logic Controller) Systems with Specific Experiences from Water Treatment (Digest No.1995/112), p. 6/1-610. https://doi.org/10.1049/ic:19950742

Ekama, G.A., Wentzel, M.C., 2004. A predictive model for the reactor inorganic suspended solids concentration in activated sludge systems. Water Research 38, 4093–4106. https://doi.org/10.1016/j.watres.2004.08.005

El-Din, A.G., Smith, D.W., 2002. A neural network model to predict the wastewater inflow incorporating rainfall events. Water Research 36, 1115–1126. https://doi.org/10.1016/S0043-1354(01)00287-1

Elmolla, E.S., Chaudhuri, M., Eltoukhy, M.M., 2010. The use of artificial neural network (ANN) for modeling of COD removal from antibiotic aqueous solution by the Fenton process. Journal of Hazardous Materials 179, 127–134. https://doi.org/10.1016/j.jhazmat.2010.02.068

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96. AAAI Press, Portland, Oregon, pp. 226–231.

Fenu, A., Guglielmi, G., Jimenez, J., Spèrandio, M., Saroj, D., Lesjean, B., Brepols, C., Thoeye, C., Nopens, I., 2010. Activated sludge model (ASM) based modelling of membrane bioreactor (MBR) processes: A critical review with special regard to MBR specificities. Water Research 44, 4272–4294. https://doi.org/10.1016/j.watres.2010.06.007

Friedman, J.H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 29, 1189–1232.

Fundneider, T., Alejo, L., Bitter, H., Mathuni, L., Döhler, C., Frederike, R., Pidde, A.V., Lackner, S., 2019. Weitestgehende Phosphorentfernung und Synergieeffekte der Tuch- und Membranfiltration als nachgeschaltete Filtrationsvefahren in der Abwasserhandlung [Advanced phosphorus removal to extremely low levels and synergy effects of cloth and membrane filtration as post filtration process in wastewater treatment]. Gas Wasserfach Wasser Abwasser 65–80.

Furumai, H., Kazmi, A.A., Fujita, M., Furuya, Y., Sasaki, K., 1999. Modeling long term nutrient removal in a sequencing batch reactor. Water Research 33, 2708–2714. https://doi.org/10.1016/S0043-1354(98)00470-9

Fushiki, T., 2011. Estimation of prediction error by using K-fold cross-validation. Stat. Comput. 21, 137–146. https://doi.org/10.1007/s11222-009-9153-8

Fux, C., Siegrist, H., 2003. Nitrogen removal from sludge digester liquids by nitrification / denitrification or partial nitritation / anammox : environmental and economical considerations. Water Science and Technology 50, 19–26.

Fytianos, K., Voudrias, E., Raikos, N., 1998. Modelling of phosphorus removal from aqueous and wastewater samples using ferric iron. Environ. Pollut. 101, 123–130. https://doi.org/10.1016/S0269-7491(98)00007-4

Gao, F., Nan, J., Zhang, X., 2017. Simulating a cyclic activated sludge system by employing a modified ASM3 model for wastewater treatment. Bioprocess Biosyst Eng 40, 877–890. https://doi.org/10.1007/s00449-017-1752-8

Géron, A., 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.

Gilbert, E.M., Agrawal, S., Karst, S.M., Horn, H., Nielsen, P.H., Lackner, S., 2014. Low Temperature Partial Nitritation/Anammox in a Moving Bed Biofilm Reactor Treating Low Strength Wastewater. Environ. Sci. Technol. 48, 8784–8792. https://doi.org/10.1021/es501649m

Gnirss, R., Dittrich, J., 2000. Microfiltration of Municipal Wastewater for Disinfection and Advanced Phosphorus Removal: Results from Trials with Different Small-Scale Pilot Plants. Water Environment Research 72, 602–609. https://doi.org/10.2175/106143000X138184

Graham, J.W., 2012. Missing Data: Analysis and Design, Statistics for Social and Behavioral Sciences. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4614-4018-5

Granata, F., Papirio, S., Esposito, G., Gargano, R., de Marinis, G., 2017. Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators. Water 9. https://doi.org/10.3390/w9020105

Granitto, P.M., Furlanello, C., Biasioli, F., Gasperi, F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems 83, 83–90. https://doi.org/10.1016/j.chemolab.2006.01.007

Griffiths, P., 1994. Modifications to the IAWPRC Task Group general activated sludge model. Water Research 28, 657–664. https://doi.org/10.1016/0043-1354(94)90145-7

Güçlü, D., Dursun, Ş., 2010. Artificial neural network modelling of a large-scale wastewater treatment plant operation. Bioprocess and Biosystems Engineering 33, 1051–1058. https://doi.org/10.1007/s00449-010-0430-x

Guenther, F., Fritsch, S., 2010. neuralnet: Training of neural networks. The R Journal 2, 9.

Gujer, W., Henze, M., Mino, T., Matsuo, T., Wentzel, M.C., Marais, G. v R., 1995. The Activated Sludge Model No. 2: biological phosphorus removal. Water Sci Technol 31, 1–11. https://doi.org/10.2166/wst.1995.0061

Gujer, W., Henze, M., Mino, T., van Loosdrecht, M., 1999. Activated Sludge Model No. 3. Water Sci Technol 39, 183–193. https://doi.org/10.2166/wst.1999.0039

Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y.M., Park, J., Kim, J.H., Cho, K.H., 2015. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. Journal of Environmental Sciences 32, 1–12. https://doi.org/10.1016/j.jes.2015.01.007

Hahsler, M., Piekenbrock, M., Doran, D., 2019. dbscan: Fast Density-Based Clustering with R. Journal of Statistical Software 91, 1–30. https://doi.org/10.18637/jss.v091.i01

Hamed, M.M., Khalafallah, M.G., Hassanien, E.A., 2004. Prediction of wastewater treatment plant performance using artificial neural networks. Environmental Modelling & Software 19, 919–928. https://doi.org/10.1016/j.envsoft.2003.10.005

Han, H., Qiao, J., 2013. Hierarchical Neural Network Modeling Approach to Predict Sludge Volume Index of Wastewater Treatment Process. IEEE Transactions on Control Systems Technology 21, 2423–2431. https://doi.org/10.1109/TCST.2012.2228861

Hao, X., Cao, X., Picioreanu, C., van Loosdrecht, M.C.M., 2005. Model-based evaluation of oxygen consumption in a partial nitrification–Anammox biofilm process. Water Sci Technol 52, 155–160. https://doi.org/10.2166/wst.2005.0195

Hao, X., Heijnen, J.J., Van Loosdrecht, M.C.M., 2002. Model-based evaluation of temperature and inflow variations on a partial nitrification–ANAMMOX biofilm process. Water Research 36, 4839–4849. https://doi.org/10.1016/S0043-1354(02)00219-1

Hao, Xiaodi, Heijnen, J.J., van Loosdrecht, M.C.M., 2002. Sensitivity analysis of a biofilm model describing a one-stage completely autotrophic nitrogen removal (CANON) process. Biotechnol. Bioeng. 77, 266–277. https://doi.org/10.1002/bit.10105

Hao, X., Loosdrecht, M., Meijer, S.C.F., Heijnen, J.J., Qian, Y., 2001. Model-Based Evaluation of Denitrifying P Removal in a Two-Sludge System. Journal of Environmental Engineering 127, 112–118. https://doi.org/10.1061/(ASCE)0733-9372(2001)127:2(112)

Hauduc, H., Gillot, S., Rieger, L., Ohtsuki, T., Shaw, A., Takács, I., Winkler, S., 2009. Activated sludge modelling in practice: an international survey. Water Science and Technology 60, 1943–1951. https://doi.org/10.2166/wst.2009.223

Hauduc, H., Rieger, L., Oehmen, A., Loosdrecht, M.C.M. van, Comeau, Y., Héduit, A., Vanrolleghem, P.A., Gillot, S., 2013. Critical review of activated sludge modeling: State of process knowledge, modeling concepts, and limitations. Biotechnology and Bioengineering 110, 24–46. https://doi.org/10.1002/bit.24624

Hauduc, H., Takács, I., Smith, S., Szabo, A., Murthy, S., Daigger, G.T., Spérandio, M., 2015. A dynamic physicochemical model for chemical phosphorus removal. Water Research 73, 157–170. https://doi.org/10.1016/j.watres.2014.12.053

Hellinga, C., Schellen, A.A.J.C., Mulder, J.W., Loosdrecht, M.C.M.V., Heijnen, J.J., 1998. The SHARON process: An innovative method for ammonium rich wastewater. Water Science and Technology 37, 135–142. https://doi.org/10.1016/S0273-1223(98)00281-9

Henze, M., Grady, C.P.L., Gujer, W., Marais, G.V.R., Matsuo, T., 1987. A general model for single-sludge wastewater treatment systems. Water Research 21, 505–515. https://doi.org/10.1016/0043-1354(87)90058-3

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M.C., 1999. Activated sludge model No.2D, ASM2D. Water Science and Technology 39, 165–182.

Henze, M., Gujer, W., Mino, T., Van Loosdrecht, M., 2000. Activated Sludge Models. IWA Publishing.

Hiatt, W.C., Grady, C.P.L., 2008. An Updated Process Model for Carbon Oxidation, Nitrification, and Denitrification. Water Environment Research 80, 2145–2156. https://doi.org/10.2175/106143008X304776

Hoang, V.Y., Jupsin, H., Le, V.C., Vasel, J.-L., 2012. Modeling of partial nitrification and denitrification in an SBR for leachate treatment without carbon addition. J Mater Cycles Waste Manag 14, 3–13. https://doi.org/10.1007/s10163-011-0033-x

Holenda, B., Domokos, E., Rédey, Á., Fazakas, J., 2008. Dissolved oxygen control of the activated sludge wastewater treatment process using model predictive control. Computers & Chemical Engineering 32, 1270–1278. https://doi.org/10.1016/j.compchemeng.2007.06.008

Hong, S.H., Lee, M.W., Lee, D.S., Park, J.M., 2007. Monitoring of sequencing batch reactor for nitrogen and phosphorus removal using neural networks. Biochemical Engineering Journal 35, 365–370. https://doi.org/10.1016/j.bej.2007.01.033

Horn, H., Hempel, D.C., 1997. Substrate utilization and mass transfer in an autotrophic biofilm system: Experimental results and numerical simulation - Horn - 1997 - Biotechnology and Bioengineering - Wiley Online Library. Biotechnology and Bioengineering 53, 363–371.

Hreiz, R., Latifi, M.A., Roche, N., 2015. Optimal design and operation of activated sludge processes: State-of-the-art. Chemical Engineering Journal 281, 900–920. https://doi.org/10.1016/j.cej.2015.06.125

Huang, Z., Luo, J., Li, X., Zhou, Y., 2009. Prediction of Effluent Parameters of Wastewater Treatment Plant Based on Improved Least Square Support Vector Machine with PSO, in: 2009 First International Conference on Information Science and Engineering. Presented at the 2009 First International Conference on Information Science and Engineering, pp. 4058–4061. https://doi.org/10.1109/ICISE.2009.846

Hulle, S.W.H.V., Vandeweyer, H.J.P., Meesschaert, B.D., Vanrolleghem, P.A., Dejans, P., Dumoulin, A., 2010. Engineering aspects and practical application of autotrophic nitrogen removal from nitrogen rich streams. Chemical Engineering Journal 162, 1–20. https://doi.org/10.1016/j.cej.2010.05.037

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Computing in Science Engineering 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

Hutson, M., 2018. Has artificial intelligence become alchemy? Science 360, 478–478. https://doi.org/10.1126/science.360.6388.478

Iacopozzi, I., Innocenti, V., Marsili-Libelli, S., Giusti, E., 2007. A modified Activated Sludge Model No. 3 (ASM3) with two-step nitrification–denitrification. Environmental Modelling & Software 22, 847–861. https://doi.org/10.1016/j.envsoft.2006.05.009

Insel, G., Hocaoğlu, S.M., Cokgor, E.U., Orhon, D., 2011. Modelling the effect of biomass induced oxygen transfer limitations on the nitrogen removal performance of membrane bioreactor. Journal of Membrane Science 368, 54–63. https://doi.org/10.1016/j.memsci.2010.11.003

Isaacs, S., Hansen, J.A., Schmidt, K., Henze, M., 1995. Examination of the Activated Sludge Model No. 2 with an alternating process. Water Science and Technology, Modelling and Control of Activated Sludge Processes 31, 55–66. https://doi.org/10.1016/0273-1223(95)00180-U

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer Texts in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7

Jiang, T., Myngheer, S., De Pauw, D.J.W., Spanjers, H., Nopens, I., Kennedy, M.D., Amy, G., Vanrolleghem, P.A., 2008. Modelling the production and degradation of soluble microbial products (SMP) in membrane bioreactors (MBR). Water Research 42, 4955–4964. https://doi.org/10.1016/j.watres.2008.09.037

Jiang, T., Sin, G., Spanjers, H., Nopens, I., Kennedy, M.D., van der Meer, W., Futselaar, H., Amy, G., Vanrolleghem, P.A., 2009. Comparison of the modeling approach between membrane bioreactor and conventional activated sludge processes. Water Environ. Res. 81, 432–440. https://doi.org/10.2175/106143008x370377

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349, 255–260. https://doi.org/10.1126/science.aaa8415

Joss, A., Salzgeber, D., Eugster, J., König, R., Rottermann, K., Burger, S., Fabijan, P., Leumann, S., Mohn, J., Siegrist, H.R., 2009. Full-scale nitrogen removal from digester liquid with partial nitritation and anammox in one SBR. Environmental Science and Technology 43, 5301–5306. https://doi.org/10.1021/es900107w

Karpinska, A.M., Bridgeman, J., 2016. CFD-aided modelling of activated sludge systems – A critical review. Water Research 88, 861–879. https://doi.org/10.1016/j.watres.2015.11.008

Kashani, M.N., Shahhosseini, S., 2010. A methodology for modeling batch reactors using generalized dynamic neural networks. Chemical Engineering Journal 159, 195–202. https://doi.org/10.1016/j.cej.2010.02.053

Kim, H., Hao, O.J., McAvoy, T.J., 2001. SBR System for Phosphorus Removal: ASM2 and Simplified Linear Model. Journal of Environmental Engineering 127, 98–104. https://doi.org/10.1061/(ASCE)0733-9372(2001)127:2(98)

Kitchenham, B., 1998. A procedure for analyzing unbalanced datasets. IEEE Trans. Softw. Eng. 24, 278–301. https://doi.org/10.1109/32.677185

Koch, G., Egli, K., Van der Meer, J.R., Siegrist, H., 2000. Mathematical modeling of autotrophic denitrification in a nitrifying biofilm of a rotating biological contactor. Water Science and Technology 41, 191 LP – 198.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence, Relevance 97, 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kuba, T., Murnleitner, E., Loosdrecht, M.C.M. van, Heijnen, J.J., 1996. A metabolic model for biological phosphorus removal by denitrifying organisms. Biotechnology and Bioengineering 52, 685–695. https://doi.org/10.1002/(SICI)1097-0290(19961220)52:6<685::AID-BIT6>3.0.CO;2-K

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28. https://doi.org/10.18637/jss.v028.i05

Kusiak, A., Verma, A., Wei, X., 2013a. A data-mining approach to predict influent quality. Environmental Monitoring and Assessment 185, 2197–2210. https://doi.org/10.1007/s10661-012-2701-2

Kusiak, A., Wei, X., 2014. Prediction of methane production in wastewater treatment facility: a data-mining approach. Ann. Oper. Res. 216, 71–81. https://doi.org/10.1007/s10479-011-1037-6

Kusiak, A., Wei, X., 2012. A data-driven model for maximization of methane production in a wastewater treatment plant. Water Science and Technology 65, 1116–1122. https://doi.org/10.2166/wst.2012.953

Kusiak, A., Zeng, Y., Zhang, Z., 2013b. Modeling and analysis of pumps in a wastewater treatment plant: A data-mining approach. Engineering Applications of Artificial Intelligence 26, 1643–1651. https://doi.org/10.1016/j.engappai.2013.04.001

Ky, R.C., Comeau, Y., Perrier, M., Takacs, I., 2001. Modelling biological phosphorus removal from a cheese factory effluent by an SBR. Water Sci. Technol. 43, 257–264.

Lackner, S., Gilbert, E.M., Vlaeminck, S.E., Joss, A., Horn, H., van Loosdrecht, M.C.M., 2014. Full-scale partial nitritation/anammox experiences – An application survey. Water Research 55, 292–303. https://doi.org/10.1016/j.watres.2014.02.032

Lackner, S., Terada, A., Smets, B.F., 2008. Heterotrophic activity compromises autotrophic nitrogen removal in membrane-aerated biofilms: Results of a modeling study. Water Research 42, 1102–1112. https://doi.org/10.1016/j.watres.2007.08.025

Lackner, S., Thoma, K., Gilbert, E.M., Gander, W., Schreff, D., Horn, H., 2015. Start-up of a full-scale deammonification SBR-treating effluent from digested sludge dewatering. Water Sci Technol 71, 553–559. https://doi.org/10.2166/wst.2014.421

Laureni, M., Weissbrodt, D.G., Villez, K., Robin, O., de Jonge, N., Rosenthal, A., Wells, G., Nielsen, J.L., Morgenroth, E., Joss, A., 2019. Biomass segregation between biofilm and flocs improves the control of nitrite-oxidizing bacteria in mainstream partial nitritation and anammox processes. Water Research 154, 104–116. https://doi.org/10.1016/j.watres.2018.12.051

Le Moullec, Y., Potier, O., Gentric, C., Leclerc, J.P., 2011. Activated sludge pilot plant: Comparison between experimental and predicted concentration profiles using three different modelling approaches. Water Research 45, 3085–3097. https://doi.org/10.1016/j.watres.2011.03.019

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324. https://doi.org/10.1109/5.726791

Lee, D.S., Jeon, C.O., Park, J.M., Chang, K.S., 2002. Hybrid neural network modeling of a full-scale industrial wastewater treatment process. Biotechnology and Bioengineering 78, 670–682. https://doi.org/10.1002/bit.10247

Lee, D.S., Park, J.M., 1999. Neural network modeling for on-line estimation of nutrient dynamics in a sequentially-operated batch reactor. Journal of Biotechnology 75, 229–239. https://doi.org/10.1016/S0168-1656(99)00171-6

Lesouef, A., Payraudeau, M., Rogalla, F., Kleiber, B., 1992. Optimizing Nitrogen Removal Reactor Configurations by On-Site Calibration of the IAWPRC Activated Sludge Model. Water Sci Technol 25, 105–123. https://doi.org/10.2166/wst.1992.0117

Levenspiel, O., 1999. Chemical reaction engineering, 3rd ed. New York.

Liu, Y., Peng, L., Chen, X., Ni, B.-J., 2015. Mathematical Modeling of Nitrous Oxide Production during Denitrifying Phosphorus Removal Process. Environ. Sci. Technol. 49, 8595–8601. https://doi.org/10.1021/acs.est.5b01650

López Garcı a, H., Machón González, I., 2004. Self-organizing map and clustering for wastewater treatment monitoring. Engineering Applications of Artificial Intelligence 17, 215–225. https://doi.org/10.1016/j.engappai.2004.03.004

Lu, X., Pereira, T.D.S., Al-Hazmi, H.E., Majtacz, J., Zhou, Q., Xie, L., Makinia, J., 2018. Model-Based Evaluation of N2O Production Pathways in the Anammox-Enriched Granular Sludge Cultivated in a Sequencing Batch Reactor. Environmental Science & Technology 52, 2800–2809. https://doi.org/10.1021/acs.est.7b05611

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2018. cluster: Cluster Analysis Basics and Extensions.

Magrí, A., Corominas, Ll., López, H., Campos, E., Balaguer, M., Colprim, J., Flotats, X., 2007. A Model for the Simulation of the SHARON Process: pH as a Key Factor. Environmental Technology 28, 255–265. https://doi.org/10.1080/09593332808618791

Manga, J., Ferrer, J., Garcia-Usach, F., Seco, A., 2001. A modification to the Activated Sludge Model No. 2 based on the competition between phosphorus-accumulating organisms and glycogen-accumulating organisms. Water Sci Technol 43, 161–171. https://doi.org/10.2166/wst.2001.0679

Mannina, G., Di Bella, G., Viviani, G., 2010. Uncertainty assessment of a membrane bioreactor model using the GLUE methodology. Biochemical Engineering Journal 52, 263–275. https://doi.org/10.1016/j.bej.2010.09.001

Mattei, M.R., Frunzo, L., D'Acunto, B., Esposito, G., Pirozzi, F., 2015. Modelling microbial population dynamics in multispecies biofilms including Anammox bacteria. Ecological Modelling 304, 44–58. https://doi.org/10.1016/j.ecolmodel.2015.02.007

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133. https://doi.org/10.1007/BF02478259

McKinney, W., 2010. Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference. Presented at the SCIPY 2010, p. 6.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2019. e1071: Misc. Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. TU Wien, Wien.

Mino, T., Liu, W.-T., Kurisu, F., Matsuo, T., 1995. Modelling glycogen storage and denitrification capability of microorganisms in enhanced biological phosphate removal processes. Water Science and Technology, Modelling and Control of Activated Sludge Processes 31, 25–34. https://doi.org/10.1016/0273-1223(95)00177-O

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill.

Mjalli, F.S., Al-Asheh, S., Alfadala, H.E., 2007. Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. Journal of Environmental Management 83, 329–338. https://doi.org/10.1016/j.jenvman.2006.03.004

Monod, J., 1949. The Growth of Bacterial Cultures. Annual Review of Microbiology 3, 371–394. https://doi.org/10.1146/annurev.mi.03.100149.002103

Morgenroth, E., Wilderer, P., 1998. Modeling of enhanced biological Phosphorus removal in a sequencing batch biofilm reactor. Water Science and Technology 37, 583–587.

Mulder, a, Graaf, a a, Robertson, L. a, Kuenen, J.G., 1995. Anaerobic ammonium oxidation discovered in a denitrifying fluidized bed reactor. https://doi.org/10.1111/j.1574-6941.1995.tb00281.x

Murnleitner, E., Kuba, T., vanLoosdrecht, M.C.M., Heijnen, J.J., 1997. An integrated metabolic model for the aerobic and denitrifying biological phosphorus removal. Biotechnol. Bioeng. 54, 434–450.

Naessens, W., Maere, T., Nopens, I., 2012a. Critical review of membrane bioreactor models – Part 1: Biokinetic and filtration models. Bioresource Technology, Membrane Bioreactors (MBRs): State-of-Art and Future 122, 95–106. https://doi.org/10.1016/j.biortech.2012.05.070

Naessens, W., Maere, T., Ratkovich, N., Vedantam, S., Nopens, I., 2012b. Critical review of membrane bioreactor models - Part 2: Hydrodynamic and integrated models. Bioresour. Technol. 122, 107–118. https://doi.org/10.1016/j.biortech.2012.05.071

Ni, B.-J., Pan, Y., Guo, J., Virdis, B., Hu, S., Chen, X., Yuan, Z., 2016. CHAPTER 16:Denitrification Processes for Wastewater Treatment, in: Metalloenzymes in Denitrification. pp. 368–418. https://doi.org/10.1039/9781782623762-00368

Ni, B.-J., Smets, B.F., Yuan, Z., Pellicer-Nàcher, C., 2013. Model-based evaluation of the role of Anammox on nitric oxide and nitrous oxide productions in membrane aerated biofilm reactor. Journal of Membrane Science 446, 332–340. https://doi.org/10.1016/j.memsci.2013.06.047

Ni, B.-J., Xie, W.-M., Liu, S.-G., Yu, H.-Q., Gan, Y.-P., Zhou, J., Hao, E.-C., 2010. Development of a mechanistic model for biological nutrient removal activated sludge systems and application to a full-scale WWTP. AIChE Journal 56, 1626–1638. https://doi.org/10.1002/aic.12066

Nolasco, D.A., Daigger, G.T., Stafford, D.R., Kaupp, D.M., Stephenson, J.P., 1998. The use of mathematical modeling and pilot plant testing to develop a new biological phosphorus and nitrogen removal process. Water Environment Research 70, 1205–1215. https://doi.org/10.2175/106143098X123543

Nopens, I., Torfs, E., Ducoste, J., Vanrolleghem, P.A., Gernaey, K.V., 2015. Population balance models: A useful complementary modelling framework for future WWTP modelling. Water Science and Technology 71, 159–167. https://doi.org/10.2166/wst.2014.500

Nourani, V., Elkiran, G., Abba, S., 2018. Wastewater treatment plant performance analysis using artificial intelligence - an ensemble approach. Water Sci. Technol. 78, 2064–2076. https://doi.org/10.2166/wst.2018.477

Novak, M., Horvat, P., 2012. Mathematical modelling and optimisation of a waste water treatment plant by combined oxygen electrode and biological waste water treatment model. Applied Mathematical Modelling 36, 3813–3825. https://doi.org/10.1016/j.apm.2011.11.028

Oehler, F., Rutherford, J.C., Coco, G., 2010. The use of machine learning algorithms to design a generalized simplified denitrification model. Biogeosciences 7, 3311–3332. https://doi.org/10.5194/bg-7-3311-2010

Oehmen, A., Lopez-Vazquez, C.M., Carvalho, G., Reis, M.A.M., van Loosdrecht, M.C.M., 2010. Modelling the population dynamics and metabolic diversity of organisms relevant in anaerobic/anoxic/aerobic enhanced biological phosphorus removal processes. Water Research 44, 4473–4486. https://doi.org/10.1016/j.watres.2010.06.017

Oles, J., Wilderer, P.A., 1991. Computer Aided Design of Sequencing Batch Reactors Based on the IAWPRC Activated Sludge Model. Water Sci Technol 23, 1087–1095. https://doi.org/10.2166/wst.1991.0560

Oliphant, T.E., 2007. Python for Scientific Computing. Computing in Science Engineering 9, 10–20. https://doi.org/10.1109/MCSE.2007.58

Olsson, G., 2012. ICA and me – A subjective review. Water Research 46, 1585–1624. https://doi.org/10.1016/j.watres.2011.12.054

Pai, T.Y., Tsai, Y.P., Lo, H.M., Tsai, C.H., Lin, C.Y., 2007. Grey and neural network prediction of suspended solids and chemical oxygen demand in hospital wastewater treatment plant effluent. Computers & Chemical Engineering 31, 1272–1281. https://doi.org/10.1016/j.compchemeng.2006.10.012

Pan, Y., Ni, B.-J., Lu, H., Chandran, K., Richardson, D., Yuan, Z., 2015. Evaluating two concepts for the modelling of intermediates accumulation during biological denitrification in wastewater treatment. Water Research 71, 21–31. https://doi.org/10.1016/j.watres.2014.12.029

Pan, Y., Ni, B.-J., Yuan, Z., 2013. Modeling Electron Competition among Nitrogen Oxides Reduction and N $_2$ O Accumulation in Denitrification. Environmental Science & Technology 47, 11083–11091. https://doi.org/10.1021/es402348n

Paul, E., Laval, M.L., Sperandio, M., 2001. Excess sludge production and costs due to phosphorus removal. Environ. Technol. 22, 1363–1371. https://doi.org/10.1080/09593332208618195

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Petersen, B., Vanrolleghem, P.A., Gernaey, K., Henze, M., 2002. Evaluation of an ASM1 model calibration procedure on a municipal–industrial wastewater treatment plant. Journal of Hydroinformatics 4, 15–38. https://doi.org/10.2166/hydro.2002.0003

Picioreanu, C., Pérez, J., van Loosdrecht, M.C.M., 2016. Impact of cell cluster size on apparent half-saturation coefficients for oxygen in nitrifying sludge and biofilms. Water Research 106, 371–382. https://doi.org/10.1016/j.watres.2016.10.017

R Foundation for Statistical Computing, 2016. R: A language and environment for statistical computing. R Core Team, Vienna,Austria.

Ráduly, B., Gernaey, K.V., Capodaglio, A.G., Mikkelsen, P.S., Henze, M., 2007. Artificial neural networks for rapid WWTP performance evaluation: Methodology and case study. Environmental Modelling & Software 22, 1208–1216. https://doi.org/10.1016/j.envsoft.2006.07.003

Randall, C.W., Buth, D., 1984. Nitrite Build-Up in Activated Sludge Resulting from Temperature Effects. Journal (Water Pollution Control Federation) 56, 1039–1044.

Reichert, P., 1994. AQUASIM – A TOOL FOR SIMULATION AND DATA ANALYSIS OF AQUATIC SYSTEMS. Water Sci Technol 30, 21–30. https://doi.org/10.2166/wst.1994.0025

Rieger, L., Gillot, S., Langergraber, G., Ohtsuki, T., Shaw, A., Takacs, I., Winkler, S., 2012. Guidelines for Using Activated Sludge Models. IWA Publishing.

Rieger, L., Koch, G., Kühni, M., Gujer, W., Siegrist, H., 2001. The eawag bio-p module for activated sludge model no. 3. Water Research 35, 3887–3903. https://doi.org/10.1016/S0043-1354(01)00110-5

Rieger, L., Takács, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P.A., Comeau, Y., 2010. Data Reconciliation for Wastewater Treatment Plant Simulation Studies—Planning for High-Quality Data and Typical Sources of Errors. Water Environment Research 82, 426–433.

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408. https://doi.org/10.1037/h0042519

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. arXiv:1901.08971 [cs].

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592. https://doi.org/10.1093/biomet/63.3.581

Russell, Stuart J., Russell, Stuart Jonathan, Norvig, P., Davis, E., 2010. Artificial Intelligence: A Modern Approach. Prentice Hall.

Salama, M.A., Hassanien, A.E., Fahmy, A.A., 2010. Reducing the influence of normalization on data classification, in: 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM). Presented at the 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 609–613. https://doi.org/10.1109/CISIM.2010.5643523

Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. New Support Vector Algorithms. Neural Comput. 12, 1207–1245. https://doi.org/10.1162/089976600300015565

Siegrist, H., Rieger, L., Koch, G., Kühnl, M., Gujer, W., 2002. The EAWAG Bio-P module for activated sludge model No. 3. Water Sci Technol 45, 61–76. https://doi.org/10.2166/wst.2002.0094

Sin, G., Guisasola, A., Pauw, D.J.W.D., Baeza, J.A., Carrera, J., Vanrolleghem, P.A., 2005. A new approach for modelling simultaneous storage and growth processes for activated sludge systems under aerobic conditions. Biotechnology and Bioengineering 92, 600–613. https://doi.org/10.1002/bit.20741

Smets, I.Y., Haegebaert, J.V., Carrette, R., Van Impe, J.F., 2003. Linearization of the activated sludge model ASM1 for fast and reliable predictions. Water Research 37, 1831–1851. https://doi.org/10.1016/S0043-1354(02)00580-8

Smith, S., Takács, I., Murthy, S., Daigger, G.T., Szabó, A., 2008. Phosphate Complexation Model and Its Implications for Chemical Phosphorus Removal. Water Environment Research 80, 428–438. https://doi.org/10.1002/j.1554-7531.2008.tb00349.x

Smola, a J., Scholkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14, 199–222. https://doi.org/Doi 10.1023/B:Stco.0000035301.49549.88

Smolders, G.J.F., Meij, J. van der, Loosdrecht, M.C.M. van, Heijnen, J.J., 1995. A structured metabolic model for anaerobic and aerobic stoichiometry and kinetics of the biological phosphorus removal process. Biotechnology and Bioengineering 47, 277–287. https://doi.org/10.1002/bit.260470302

Smolders, G.J.F., Meij, J. van der, Loosdrecht, M.C.M. van, Heijnen, J.J., 1994. Model of the anaerobic metabolism of the biological phosphorus removal process: Stoichiometry and pH influence. Biotechnology and Bioengineering 43, 461–470. https://doi.org/10.1002/bit.260430605

Soejima, K., Matsumoto, S., Ohgushi, S., Naraki, K., Terada, A., Tsuneda, S., Hirata, A., 2008. Modeling and experimental study on the anaerobic/aerobic/anoxic process for simultaneous nitrogen and phosphorus removal: The effect of acetate addition. Process Biochem. 43, 605–614. https://doi.org/10.1016/j.procbio.2008.01.022

Spérandio, M., Espinosa, M.C., 2008. Modelling an aerobic submerged membrane bioreactor with ASM models on a large range of sludge retention time. Desalination, Selected Papers Presented at the 4th International IWA Conference on Membranes for Water and Wastewater Treatment, 15-17 May 2007, Harrogate, UK. Guest Edited by Simon Judd; and Papers Presented at the International Workshop on Membranes and Solid-Liquid Separation Processes, 11 July 2007, INSA, Toulouse, France. Guest edited by Saravanamuthu Vigneswaran and Jaya Kandasamy 231, 82–90. https://doi.org/10.1016/j.desal.2007.11.040

Sun, W., Liu, M., 2016. Prediction and analysis of the three major industries and residential consumption CO2 emissions based on least squares support vector machine in China. Journal of Cleaner Production 122, 144–153. https://doi.org/10.1016/j.jclepro.2016.02.053

Szabó, A., Takács, I., Murthy, S., Daigger, G.T., Licskó, I., Smith, S., 2008. Significance of Design and Operational Variables in Chemical Phosphorus Removal. Water Environment Research 80, 407–416. https://doi.org/10.2175/106143008X268498

Terada, A., Lackner, S., Tsuneda, S., Smets, B.F., 2007. Redox-Stratification Controlled Biofilm (ReSCoBi) for Completely Autotrophic Nitrogen Removal: The Effect of Co-versus Counter-Diffusion on Reactor Performance. Biotechnology and Bioengineering 97, 40–51.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63, 411–423. https://doi.org/10.1111/1467-9868.00293

Torregrossa, D., Leopold, U., Hernández-Sancho, F., Hansen, J., 2018. Machine learning for energy cost modelling in wastewater treatment plants. Journal of Environmental Management 223, 1061–1067. https://doi.org/10.1016/j.jenvman.2018.06.092

van der Walt, S., Colbert, S.C., Varoquaux, G., 2011. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science Engineering 13, 22–30. https://doi.org/10.1109/MCSE.2011.37

Van Loosdrecht, M.C.M., Henze, M., 1999. Maintenance, endogeneous respiration, lysis, decay and predation. Water Sci Technol 39, 107–117. https://doi.org/10.2166/wst.1999.0024

Van Loosdrecht, M.C.M., Lopez-Vazquez, C.M., Meijer, S.C.F., Hooijmans, C.M., Brdjanovic, D., 2015. Twenty-five years of ASM1: past, present and future of wastewater treatment modelling. Journal of Hydroinformatics 17, 697–718. https://doi.org/10.2166/hydro.2015.006

Vangsgaard, A.K., Mauricio-Iglesias, M., Gernaey, K.V., Smets, B.F., Sin, G., 2012a. Sensitivity analysis of autotrophic N removal by a granule based bioreactor: Influence of mass transfer versus microbial kinetics. Bioresource Technology 123, 230–241. https://doi.org/10.1016/j.biortech.2012.07.087

Vangsgaard, A.K., Mauricio-Iglesias, M., Gernaey, K. V., Smets, B.F., Sin, G., 2012b. Sensitivity analysis of autotrophic N removal by a granule based bioreactor: Influence of mass transfer versus microbial kinetics. Bioresource Technology 123, 230–241. https://doi.org/10.1016/j.biortech.2012.07.087

Vanhooren, H., Meirlaen, J., Amerlinck, Y., Claeys, F., Vangheluwe, H., Vanrolleghem, P. a, 2003. WEST : modelling biological wastewater treatment. Journal of Hydroinformatics 27–50.

Vannucci, M., Colla, V., 2018. Advanced Neural Networks Systems for Unbalanced Industrial Datasets, in: Esposito, A., Faudez-Zanuy, M., Morabito, F.C., Pasero, E. (Eds.), Multidisciplinary Approaches to Neural Computing, Smart Innovation, Systems and Technologies. Springer International Publishing, Cham, pp. 181–189. https://doi.org/10.1007/978-3-319-56904-8_18

Vanrolleghem, P.A., Spanjers, H., Petersen, B., Ginestet, P., Takacs, I., 1999. Estimating (combinations of) Activated Sludge Model No. 1 parameters and components by respirometry. Water Sci. Technol. 39, 195–214. https://doi.org/10.1016/S0273-1223(98)00786-0

Vázquez-Padín, J.R., Mosquera-Corral, A., Campos, J.L., Méndez, R., Carrera, J., Pérez, J., 2010. Modelling aerobic granular SBR at variable COD/N ratios including accurate description of total solids concentration. Biochemical Engineering Journal 49, 173–184. https://doi.org/10.1016/j.bej.2009.12.009

Vega De Lille, M., Berkhout, V., Fröba, L., Gro\s s, F., Delgado, A., 2015. Ammonium estimation in an ANAMMOX SBR treating anaerobically digested domestic wastewater. Chemical Engineering Science 130, 109–119. https://doi.org/10.1016/j.ces.2015.03.018

Verma, A., Wei, X., Kusiak, A., 2013. Predicting the total suspended solids in wastewater: A data-mining approach. Engineering Applications of Artificial Intelligence 26, 1366–1372. https://doi.org/10.1016/j.engappai.2012.08.015

Vlaeminck, S.E., Terada, A., Smets, B.F., De Clippeleir, H., Schaubroeck, T., Bolea, S., Demeestere, L., Mast, J., Boon, N., Carballa, M., Verstraete, W., 2010. Aggregate size and architecture determine microbial activity balance for one-stage partial nitritation and anammox. Applied and Environmental Microbiology 76, 900–909. https://doi.org/10.1128/AEM.02337-09

Volcke, E.I.P., Hulle, S.W.H.V., Donckels, B.M.R., Loosdrecht, M.C.M. van, Vanrolleghem, P.A., 2005. Coupling the SHARON process with Anammox: Model-based scenario analysis with focus on operating costs. Water Science and Technology 52, 107–115. https://doi.org/10.2166/wst.2005.0093

Volcke, E.I.P., Picioreanu, C., De Baets, B., van Loosdrecht, M.C.M., 2012. The granule size distribution in an anammox-based granular sludge reactor affects the conversion-Implications for modeling. Biotechnology and Bioengineering 109, 1629–1636. https://doi.org/10.1002/bit.24443

Wan, X., Baeten, J.E., Volcke, E.I.P., 2019. Effect of operating conditions on N2O emissions from one-stage partial nitritation-anammox reactors. Biochemical Engineering Journal 143, 24–33. https://doi.org/10.1016/j.bej.2018.12.004

Wang, H., Yang, F., Luo, Z., 2016. An experimental study of the intrinsic stability of random forest variable importance measures. BMC Bioinformatics 17, 60. https://doi.org/10.1186/s12859-016-0900-5

Wang, Y., Geng, J., Guo, G., Wang, C., Liu, S., 2011. N2O production in anaerobic/anoxic denitrifying phosphorus removal process: The effects of carbon sources shock. Chemical Engineering Journal 172, 999–1007. https://doi.org/10.1016/j.cej.2011.07.014

Wanner, O., Gujer, W., 1986. A multispecies biofilm model. Biotechnology and Bioengineering 28, 314–328. https://doi.org/10.1002/bit.260280304

Wentzel, M.C., Ekama, G.A., Marais, G. v. R., 1992. Processes and Modelling of Nitrification Denitrification Biological Excess Phosphorus Removal Systems – A Review. Water Science and Technology 25, 59–82. https://doi.org/10.2166/wst.1992.0114

Wiener, M., Liaw, A., 2002. Classification and Regression by randomForest. R news 2, 18–22.

Wilkinson, L., Task Force on Statistical Inference, 1999. Statistical Methods in Psychology Journals: Guidelines and Explanations. American Psychologist 54, 594–604.

Wisniewski, K., Kowalski, M., Makinia, J., 2018. Modeling nitrous oxide production by a denitrifying-enhanced biologically phosphorus removing (EBPR) activated sludge in the presence of different carbon sources and electron acceptors. Water Research 142, 55–64. https://doi.org/10.1016/j.watres.2018.05.041

Witten, I.H., Frank, E., Hall, M., 2011. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier. https://doi.org/10.1016/C2009-0-19715-5

Wouters-Wasiak, K., Héduit, A., Audic, J.M., Lefèvre, F., 1994. Real-time control of nitrogen removal at full-scale using oxidation reduction potential. Water Sci Technol 30, 207–210. https://doi.org/10.2166/wst.1994.0192

Wrage, N., Velthof, G.L., van Beusichem, M.L., Oenema, O., 2001. Role of nitrifier denitrification in the production of nitrous oxide. Soil Biology and Biochemistry 33, 1723–1732. https://doi.org/10.1016/S0038-0717(01)00096-7

Wu, J., Yan, G., Zhou, G., Xu, T., 2014. Model predictive control of biological nitrogen removal via partial nitrification at low carbon/nitrogen (C/N) ratio. Journal of Environmental Chemical Engineering 2, 1899–1906. https://doi.org/10.1016/j.jece.2014.08.007

Wu, X., Yang, Y., Wu, G., Mao, J., Zhou, T., 2016. Simulation and optimization of a coking wastewater biological treatment process by activated sludge models (ASM). Journal of Environmental Management 165, 235–242. https://doi.org/10.1016/j.jenvman.2015.09.041

Wyffels, S., Van Hulle, S.W.H., Boeckx, P., Volcke, E.I.P., Van Cleemput, O., Vanrolleghem, P. a., Verstraete, W., 2004a. Modeling and simulation of oxygen-limited partial nitritation in a membrane-assisted bioreactor (MBR). Biotechnology and Bioengineering 86, 531–542. https://doi.org/10.1002/bit.20008

Wyffels, S., Van Hulle, S.W.H., Boeckx, P., Volcke, E.I.P., Van Cleemput, O., Vanrolleghem, P. a., Verstraete, W., 2004b. Modeling and simulation of oxygen-limited partial nitritation in a membrane-assisted bioreactor (MBR). Biotechnology and Bioengineering 86, 531–542. https://doi.org/10.1002/bit.20008

Xie, B., Ma, Y., Wan, J., Wang, Y., Guan, Z., 2017. An accuracy model for on-line prediction of effluent ammonia nitrogen in anammox treatment system based on PCA-BP algorithm. 2017 2nd IEEE

International Conference on Computational Intelligence and Applications, ICCIA 2017 2017, 402–406. https://doi.org/10.1109/CIAPP.2017.8167248

Xu, X.-J., Chen, C., Wang, A.-J., Ni, B.-J., Guo, W.-Q., Yuan, Y., Huang, C., Zhou, X., Wu, D.-H., Lee, D.-J., Ren, N.-Q., 2017. Mathematical modeling of simultaneous carbon-nitrogen-sulfur removal from industrial wastewater. Journal of Hazardous Materials 321, 371–381. https://doi.org/10.1016/j.jhazmat.2016.08.074

Yagci, N., Insel, G., Artan, N., Orhon, D., 2004. Modelling and calibration of phosphate and glycogen accumulating organism competition for acetate uptake in a sequencing batch reactor. Water Sci Technol 50, 241–250. https://doi.org/10.2166/wst.2004.0382

Yagci, N., Insel, G., Tasli, R., Artan, N., Randall, C.W., Orhon, D., 2006. A new interpretation of ASM2d for modeling of SBR performance for enhanced biological phosphorus removal under different P/HAc ratios. Biotechnology and Bioengineering 93, 258–270. https://doi.org/10.1002/bit.20701

Yang, K., Li, Z., Zhang, H., Qian, J., Chen, G., 2010. Municipal wastewater phosphorus removal by coagulation. Environmental Technology 31, 601–609. https://doi.org/10.1080/09593330903573223

Yang, S.-S., Pang, J.-W., Guo, W.-Q., Yang, X.-Y., Wu, Z.-Y., Ren, N.-Q., Zhao, Z.-Q., 2017. Biological phosphorus removal in an extended ASM2 model: Roles of extracellular polymeric substances and kinetic modeling. Bioresource Technology 232, 412–416. https://doi.org/10.1016/j.biortech.2017.01.048

Yeo, I.-K., Johnson, R.A., 2000. A New Family of Power Transformations to Improve Normality or Symmetry. Biometrika 87, 954–959.

Zeng, Raymond J., Yuan, Z., Keller, J., 2003. Enrichment of denitrifying glycogen-accumulating organisms in anaerobic/anoxic activated sludge system. Biotechnology and Bioengineering 81, 397–404. https://doi.org/10.1002/bit.10484

Zeng, R. J., Yuan, Z.G., Keller, J., 2003. Model-based analysis of anaerobic acetate uptake by a mixed culture of polyphosphate-accumulating and glycogen-accumulating organisms. Biotechnol. Bioeng. 83, 293–302. https://doi.org/10.1002/bit.10671

Zhang, Z., Kusiak, A., Zeng, Y., Wei, X., 2016. Modeling and optimization of a wastewater pumping system with data-mining methods. Applied Energy 164, 303–311. https://doi.org/10.1016/j.apenergy.2015.11.061

Zhao, H., Hao, O.J., McAvoy, T.J., 1999. Approaches to modeling nutrient dynamics: ASM2, simplified model and neural nets. Water Science and Technology, Modelling and microbiology of activated sludge processes 39, 227–234. https://doi.org/10.1016/S0273-1223(98)00788-4

Zhao, J., Huang, J., Guan, M., Zhao, Y., Chen, G., Tian, X., 2016. Mathematical simulating the process of aerobic granular sludge treating high carbon and nitrogen concentration wastewater. Chemical Engineering Journal 306, 676–684. https://doi.org/10.1016/j.cej.2016.07.098

Zheng, X., Plume, S., Ernst, M., Croué, J.-P., Jekel, M., 2012. In-line coagulation prior to UF of treated domestic wastewater – foulants removal, fouling control and phosphorus removal. Journal of Membrane Science 403–404, 129–139. https://doi.org/10.1016/j.memsci.2012.02.051

Zhou, M., Gong, J., Yang, C., Pu, W., 2013. Simulation of the performance of aerobic granular sludge SBR using modified ASM3 model. Bioresource Technology 127, 473–481. https://doi.org/10.1016/j.biortech.2012.09.076

Zipper, T., Fleischmann, N., Haberl, R., 1998. Development of a new system for control and optimization of small wastewater treatment plants using oxidation-reduction potential (ORP). Water Sci Technol 38, 307–314. https://doi.org/10.2166/wst.1998.0225

Zuthi, M.F.R., Guo, W.S., Ngo, H.H., Nghiem, L.D., Hai, F.I., 2013. Enhanced biological phosphorus removal and its modeling for the activated sludge and membrane bioreactor processes. Bioresource Technology 139, 363–374. https://doi.org/10.1016/j.biortech.2013.04.038

# 10 Appendix

## 10.1 Softwares based on ASM type models

Together with the evolution of ASM models, the computational tools and softwares available for the implementation and solving of ASM equations appeared. Nowadays there is a variety of commercial and free software packages. Table A.1 summarizes information from commercial and free softwares commonly used both for learning and applied research: GPS-X, Simba, Sumo, BioWin, Aquasim, ASIM and WEST.

Table A. 1 Software packages available for modeling biological processes in wastewater treatment (ASM)

| Software | Distr. | Progr. language | Country | Free Lic. | Operating system |
|---|---|---|---|---|---|
| GPS-X | Hydromantis | ACSL, connection to Matlab | United States of America | No | Unix and Windows |
| Simba | Ifak-In control solutions | Matlab/Simulink, C# | Germany/Canada | No | Windows |
| SUMO | Dynamita | C# (graphical interface), SumoSlang, compiled in C++ | France | No | Windows |
| WBioWin | EnviroSim | Embarcadero Delphi ,C++ and Fortran. | France | No | Windows |
| Aquasim | Eawag | C++ and XVT (graphical interface) | Switzerland | Yes | Unix and Windows |
| Asim | Eawag | C++ | Switzerland | Yes and commercial version | Windows |
| WEST | MIKE by DHI | C++ | Belgium | No | Unix and Windows |

GPS-X is distributed by Hydromantis in Hamilton, Ontario and the software has a number of models for activated sludge and treatment works modeling such as ASM1, ASM2, ASM2d, ASM3, a temperature dependent version of ASM1. The GPS-X software has automated the input file functions. The input file for many programs is often a large data file with various wastewater inputs fractionated into 16 chemical parameters as required by the IWA models. SIMBA on the other hand was developed in Germany. Initially, Simba was built over Matlab/Simulink, however nowadays not necessary. Simba as GPS-X is loaded with IWA models. SUMO was developed by Dr. Imre Takacs and distributed by Dynamita in France and is a powerful open process source, multipurpose simulation environment developed for wastewater treatment process modeling. BioWin is distributed by Envirosim, ASM models are loaded in BioWin already, additionally various reactor configurations are available. Aquasim and ASIM were developed by EAWAG, and Aquasim is by far one of the most applied softwares in research for biological processes in wastewater (Reichert, 1994). ASIM (Activated Sludge SIMulation

Program) is a simulation program, which allows the simulation of different biological wastewater treatment systems with up to 10 different reactors in series (aerobic, anoxic and anaerobic), including sludge return and internal recirculation streams. The program allows for the definition of process control loops (simple proportional controllers and on/off type controllers) and dynamic simulation of load variation (diurnal load variation, temperature variation, variation of operational parameters such as aeration, excess sludge removal, recycle rates, among others. Finally, WEST was developed in Belgium and today is distributed by MIKE, DHI (Vanhooren et al., 2003).

## 10.2 Classification of missing values

### 10.2.1 Missing completely at random

Let $\boldsymbol{Var_i} = (Var_{i,1}, Var_{i,2}, Var_{i,j}, \dots, Var_{i,n})$ be a variable in a dataset of $m$ variables: $\{\boldsymbol{Var_1}, \boldsymbol{Var_2}, \boldsymbol{Var_i}, \dots, \boldsymbol{Var_m}\}$, clearly $i = 1, \dots, m$. In the definition of $\boldsymbol{Var_i}$, $n$ denotes the intended number of observations to collect from variable $\boldsymbol{Var_i}$, i.e. $Var_{i,j}$ and $j = 1, \dots, n$. For each individual observation $j$ and variable $i$, we can define a parameter $\boldsymbol{R_i} = (R_{i,1}, R_{i,2}, R_{i,j}, \dots, R_{i,n})$ such that the value of $R_{i,j} = 1$ if $Var_{i,j}$ is observed and $R_{i,j} = 0$ if $Var_{i,j}$ is missing. Then, the missing value mechanism is defined as,

$$\Pr(\boldsymbol{R_i} | \boldsymbol{Var_i}) \qquad \text{Def. A. 1}$$

Def 3.1, refers to the probability of observing variable $\boldsymbol{i}$'s data given the potentially unseen values in $\boldsymbol{Var_i}$. The previous definition will facilitate the description of the types of missing values in this and upcoming sub-sections.

Given a dataset with missing values, data are missing completely at random (MCAR) if the probability of a value in the dataset being missing is unrelated to the phenomena that caused the *missingness*.

$$\Pr(\boldsymbol{R_i} | \boldsymbol{Var_i}) = \Pr(\boldsymbol{R_i}) \qquad \text{Def. A. 2}$$

If data are missing by design, because of an equipment failure or because the samples are performed in different frequencies compared to other variables, such data are classified as MCAR. The statistical advantage of data that are MCAR is that the analysis remains unbiased (Rubin, 1976). Power may be lost in the design, but the estimated parameters are not biased by the absence of the data.

Missing at random

Data are referred as missing at random (MAR) if given the observed data, the probability distribution of $\boldsymbol{R_i}$ is independent of the missing data. If for any $\boldsymbol{i}$, we define $\boldsymbol{Var_{i,O}} \subseteq \boldsymbol{Var_i}$ as the subset of observed values in $\boldsymbol{Var_i}$, then MAR can be defined as,

$$\Pr(\boldsymbol{R_i} | \boldsymbol{Var_i}) = \Pr(\boldsymbol{Var_{i,O}}) \qquad \text{Def. A. 3}$$

Under MAR the chance of observing a value will depend on its value. Graham, (2012) defines MAR as an accessible *missingness*, because the researcher had access to the cause of *missingness*. For example, in the context of bWWTP, when monitoring a lab-scale sequencing batch reactor (SBR), sensors are placed in the reactor to measure parameters such as pH or temperature (T). Due to the nature of the operation of SBR systems, there is a period of time during a cycle where the reactor will be empty and the values registered by the sensors (if they are not in contact with the medium), will be deleted and not considered for analysis. The nature of the *missingness* of the pH and T sensors values in this example will be MAR, since they will depend on the cycle of the SBR. Depending on the analysis method, these data can still induce parameter bias in analyses due to the contingent emptiness of values. However, if the parameter is estimated with methods such as multiple imputation, this imputation method will provide asymptotically unbiased estimates.

### 10.2.2 Missing not at random

If data is neither MCAR or MAR, then the data is Missing Not At Random (MNAR). In MNAR the probability of missing observations is independent of the observed data.

$$\Pr(\boldsymbol{R_i}|\boldsymbol{Var_i}) \neq \Pr(\boldsymbol{Var_{i,O}}) \qquad \text{Def. A. 4}$$

Graham, (2012) defines MNAR as inaccessible *missingness* because the cause of *missingness* has not been *measured* and is therefore not available for analysis. Examples of MNAR in the context of bWWTP are very frequent, such as irregular sampling of pollutants (there is not a fixed pattern on the sampling of certain pollutants).

## 10.3 Figures and tables

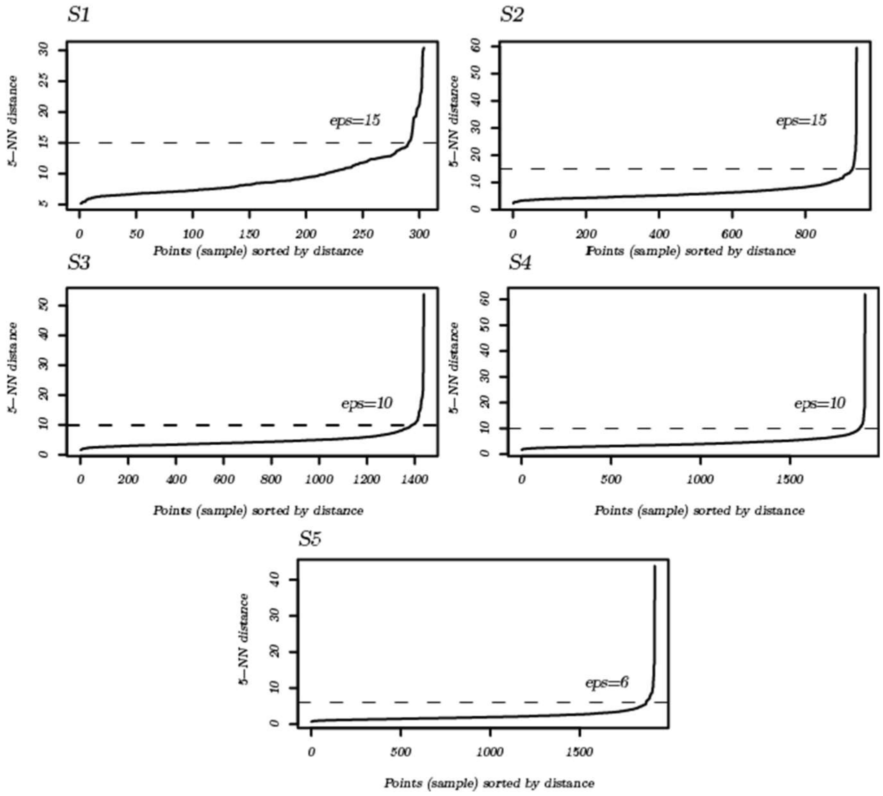


Figure A. 1 Selection of eps value in DBSCAN clustering: winter season

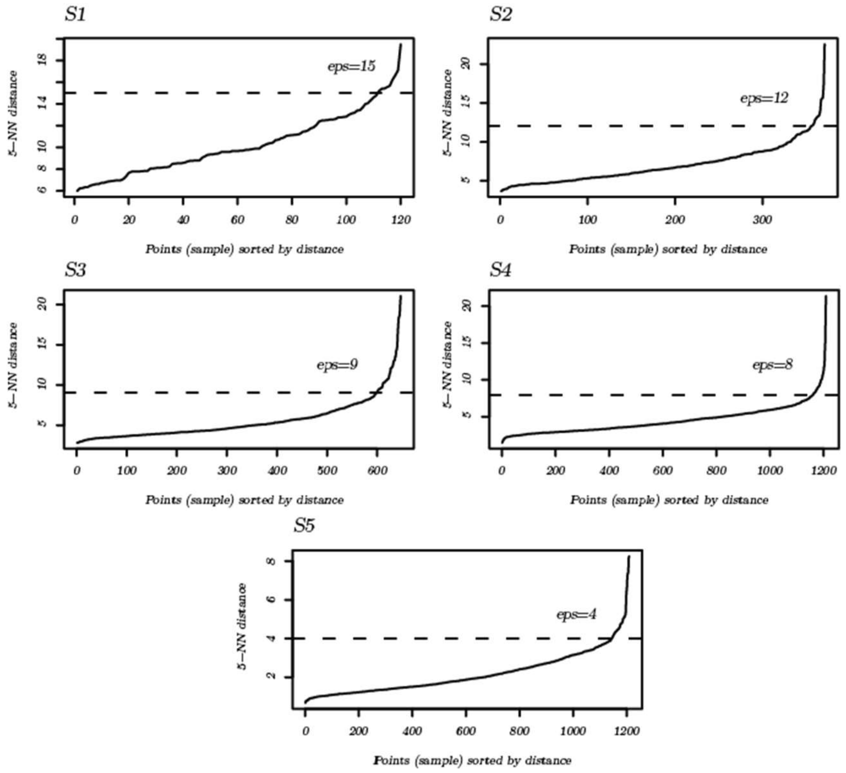


Figure A. 2 Selection of eps value in DBSCAN clustering: summer season

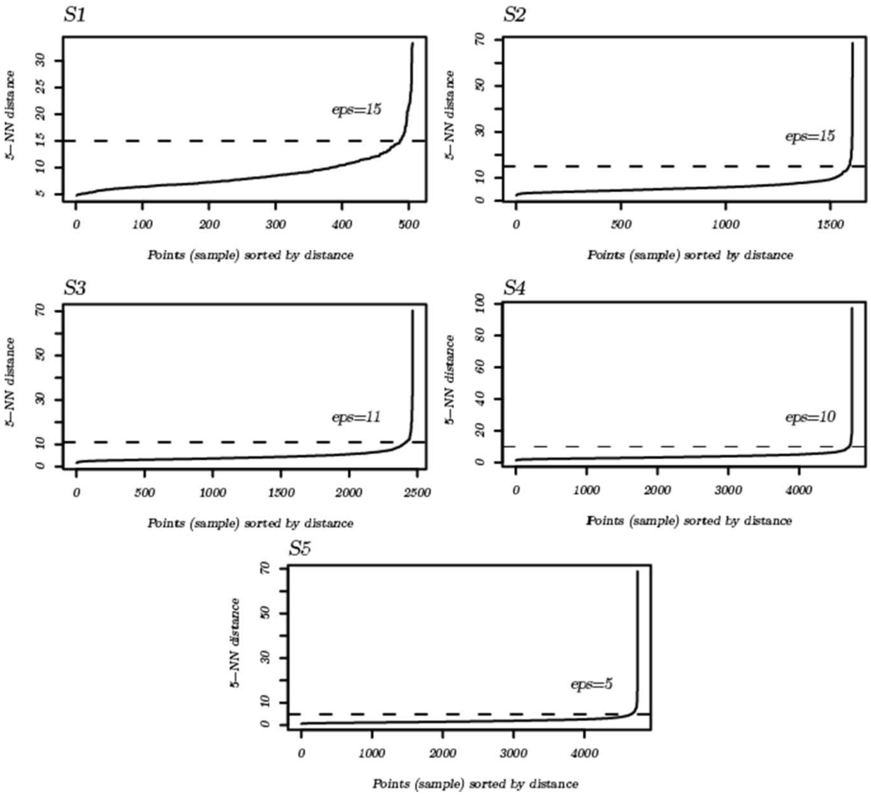


Figure A. 3 Selection of eps value in DBSCAN clustering: all seasons

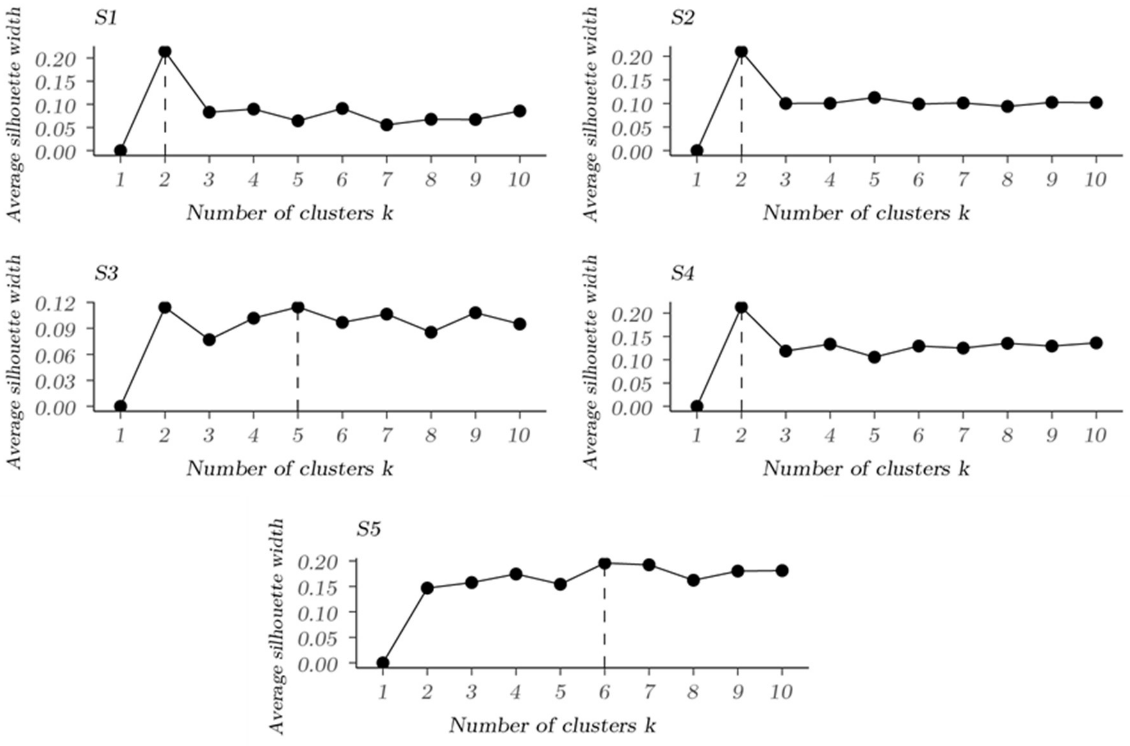


Figure A. 4 Selection of clusters through silhouette method in k-means: winter

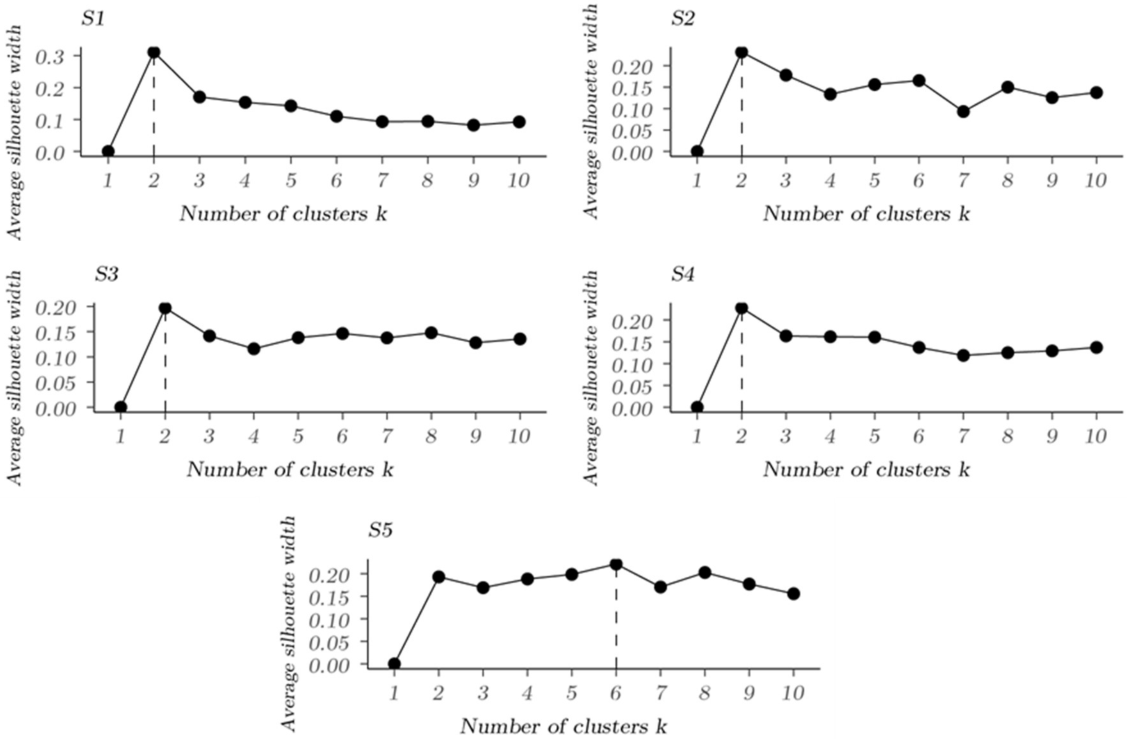


Figure A. 5 Selection of clusters through silhouette method in k-means: summer

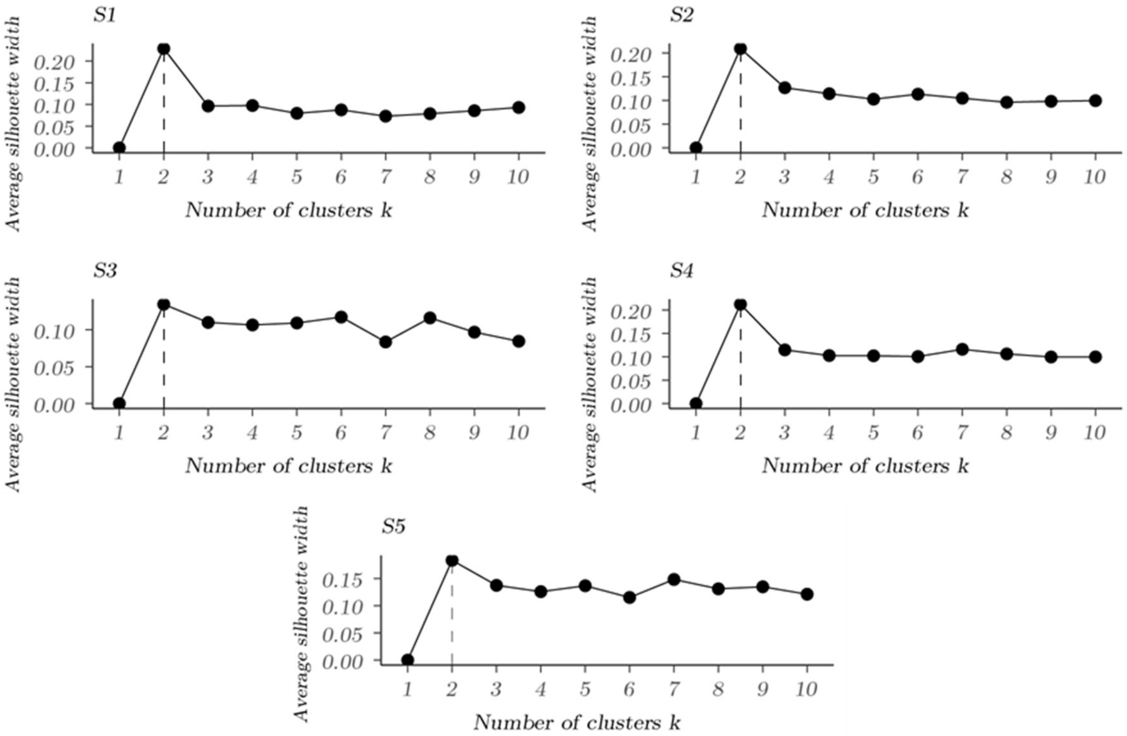


Figure A. 6 Selection of clusters through silhouette method in k-means: summer

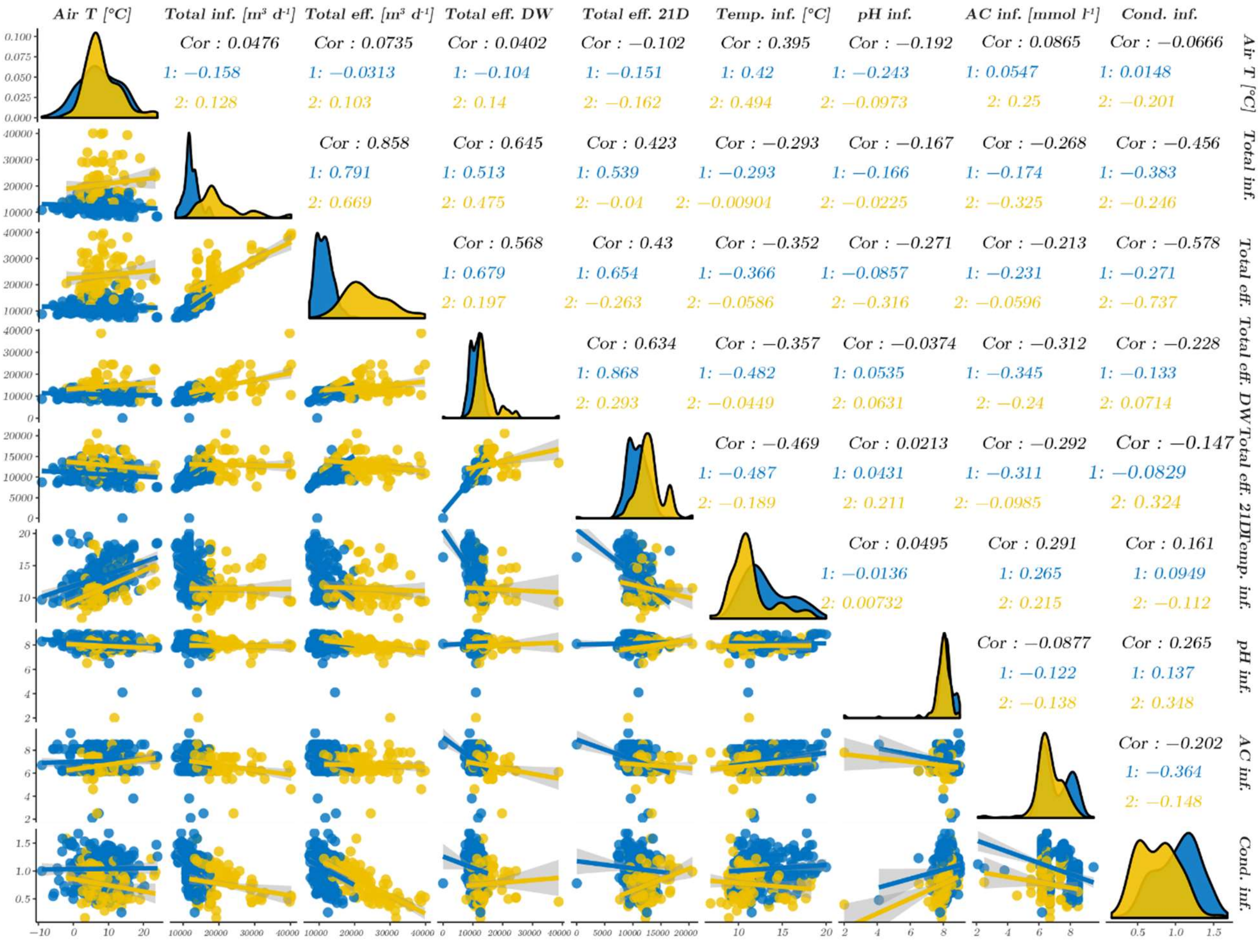


Figure A. 7 k-means clustering analysis results for some parameters for the winter seasons (subgroup 1).

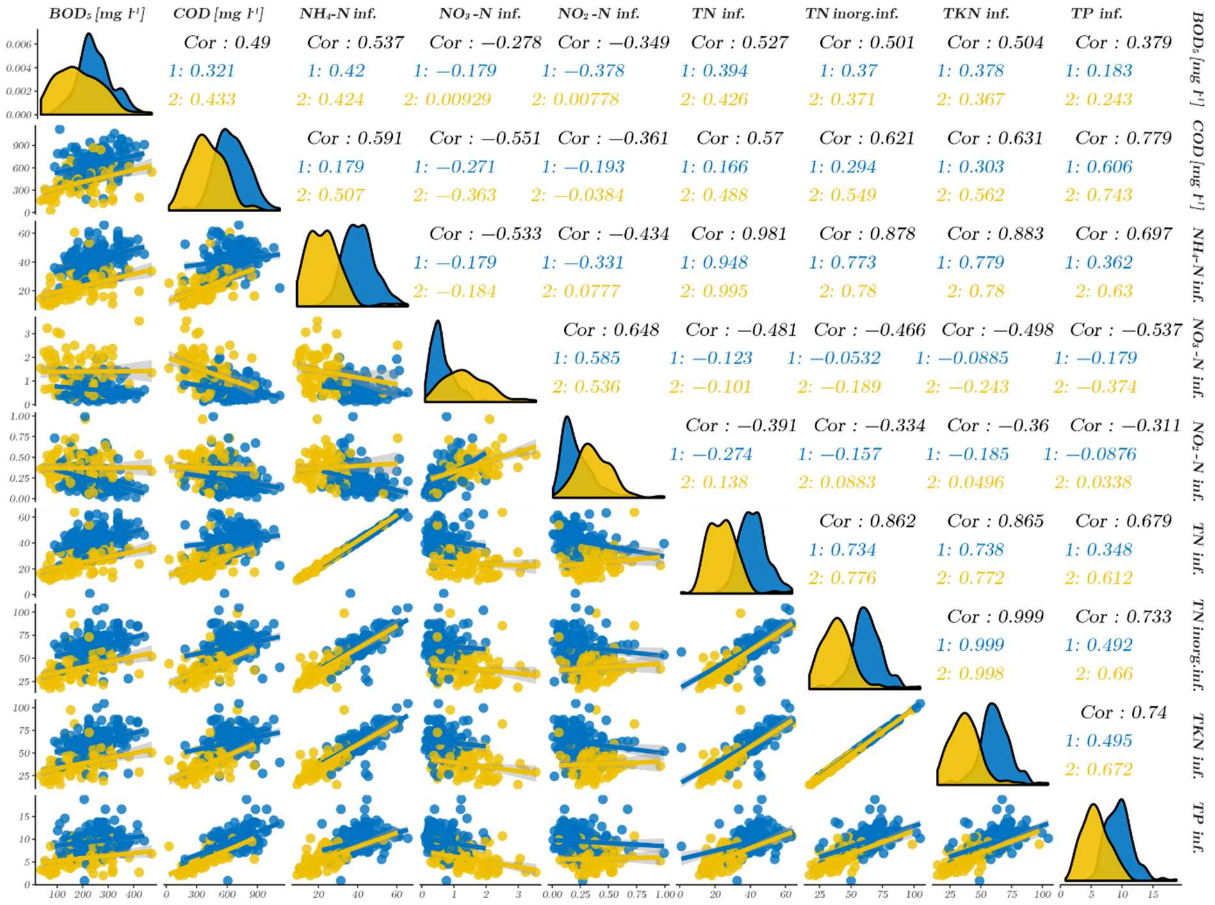


Figure A. 8 k-means clustering analysis results for some parameters for the winter seasons (subgroup 1).

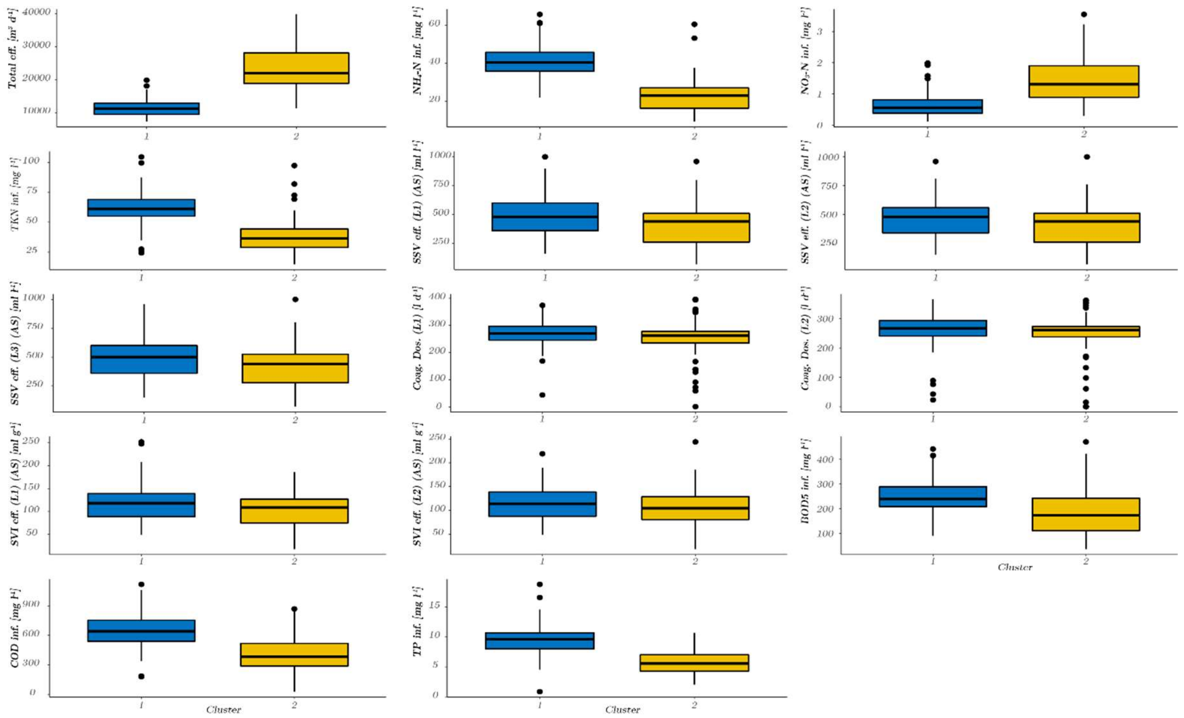


Figure A. 9 Box-plot of different parameters according to clusters in S1 (winter). Lower and upper box boundaries 25th and 75th percentiles, respectively, line inside box median, outliers are identified and scattered in black.

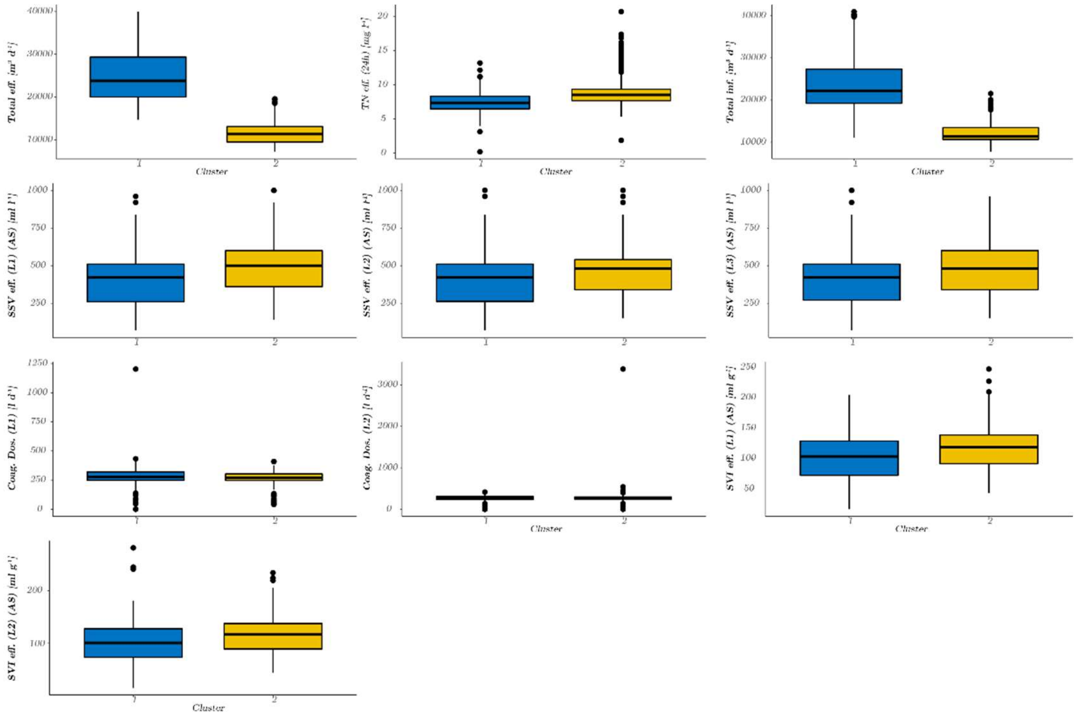


Figure A. 10 Box-plot of different parameters according to clusters in S2 (winter). Lower and upper box boundaries 25th and 75th percentiles, respectively, line inside box median, outliers are identified and scattered in black.

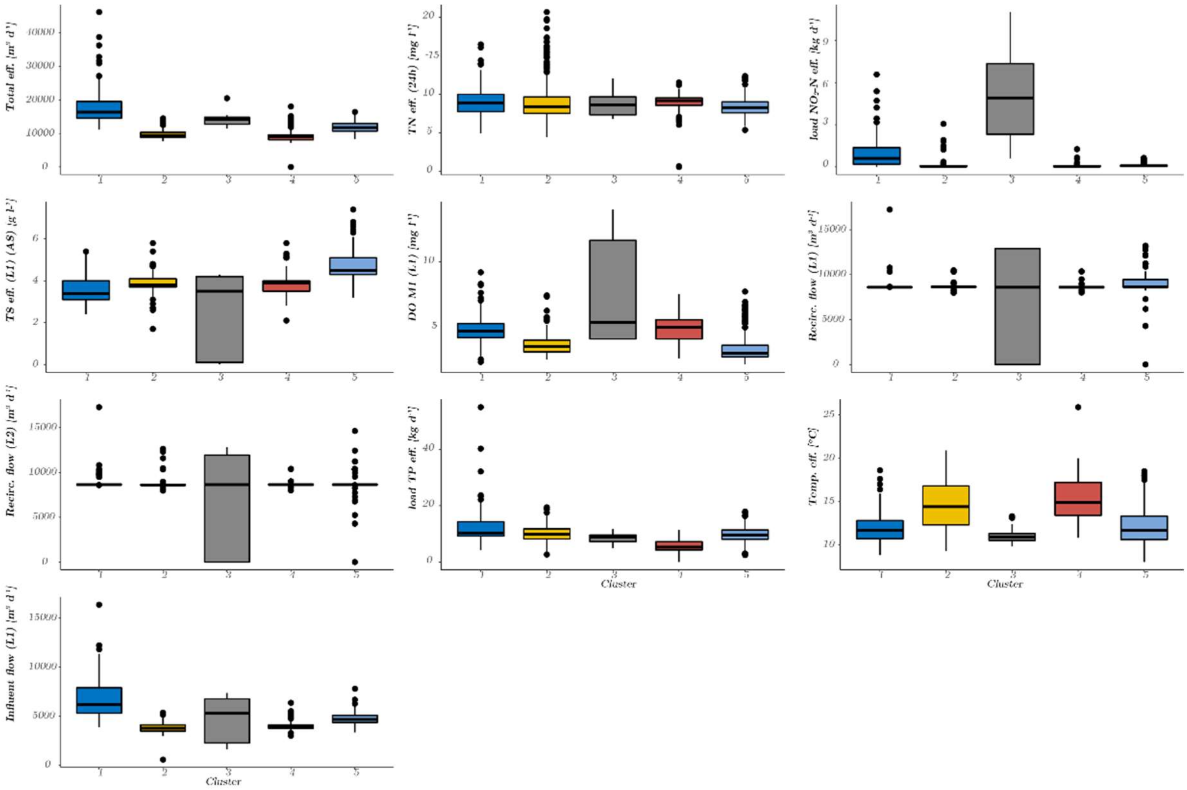


Figure A. 11 Box-plot of different parameters according to clusters in S3 (winter). Lower and upper box boundaries 25th and 75th percentiles, respectively, line inside box median, outliers are identified and scattered in black.

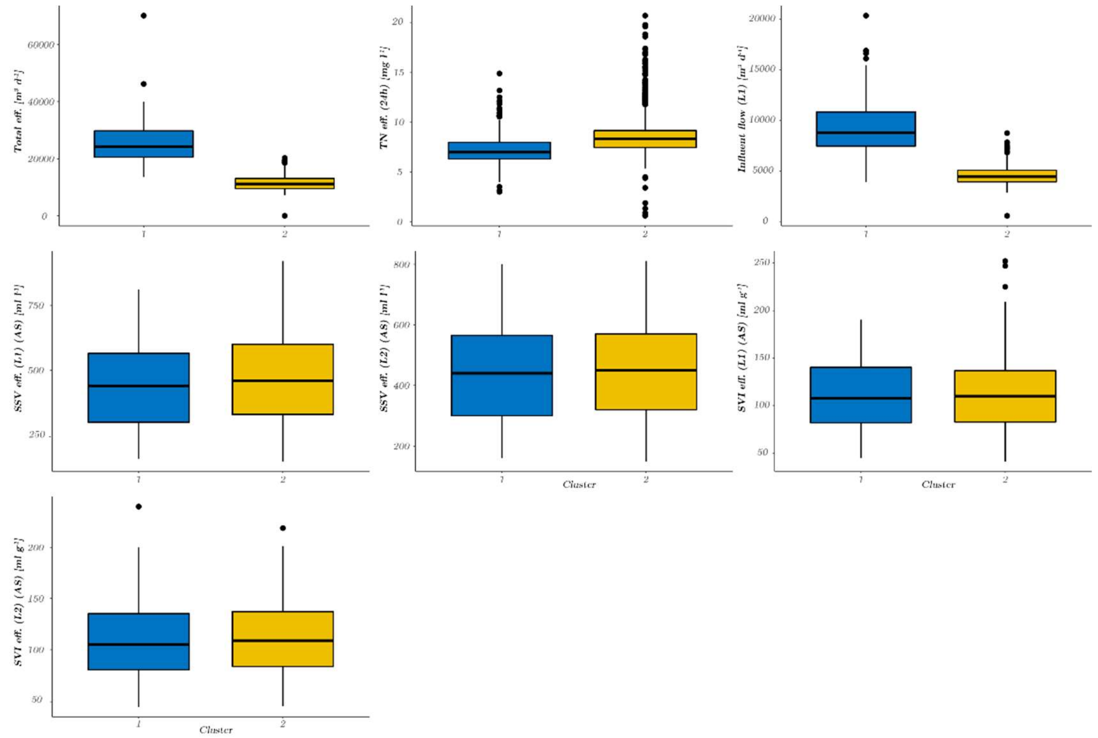


Figure A. 12 Box-plot of different parameters according to clusters in S4 (winter). Lower and upper box boundaries 25th and 75th percentiles, respectively, line inside box median, outliers are identified and scattered in black.

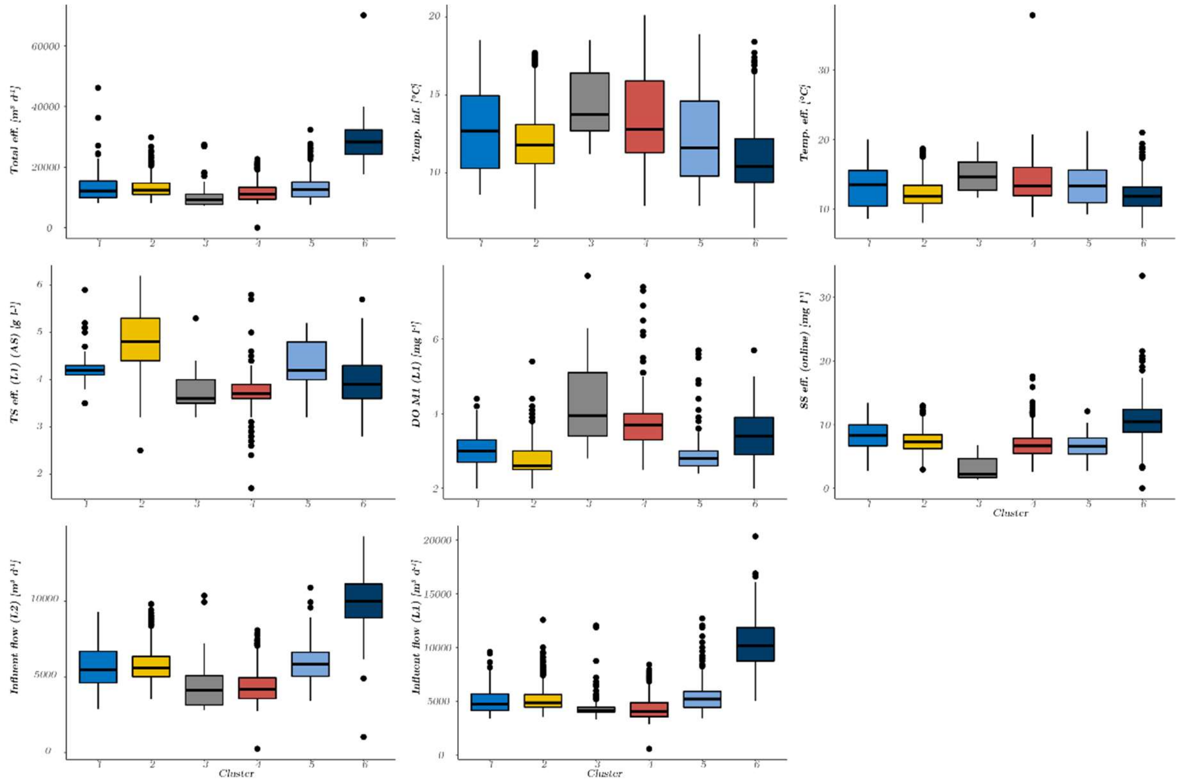


Figure A. 13 Box-plot of different parameters according cluster in S5 (winter). Lower and upper box boundaries 25th and 75th percentiles, respectively, line inside box median, outliers are identified and scattered in black.

Table A. 2 List of Abbreviations SBR reactors

| Abbreviations | Description | Units |
|---|---|---|
| Aer.Time per Cycle | Aeration time per SBR cycle | [min cycle$^{-1}$] |
| Anammox | Anaerobic ammonium oxidation | Ad./Adimensional |
| AnAOB | Anaerobic ammonium oxidizing bacteria | [mg(N)g(oTS)$^{-1}\cdot$h$^{-1}$] |
| ANN | Artificial neural networks | Ad. |
| AOB | Ammonia oxidizing bacteria | mg($NH_4$-N) g(oTS)$^{-1}\cdot$h$^{-1}$ |
| ASM1 | Activated sludge model No. 1 | Ad. |
| ASM3 | Activated sludge model No. 3 | Ad. |
| COD | Chemical oxygen demand | [mg L$^{-1}$] |
| COD filt. Eff. | Effluent chemical oxygen demand filtered (soluble) concentration | [mg L$^{-1}$] |
| COD filt. Inf. | Influent chemical oxygen demand filtered (soluble) concentration | [mg L$^{-1}$] |
| COD hom. Eff. | Effluent chemical oxygen demand homogeneous concentration | [mg L$^{-1}$] |
| COD hom. Inf. | Influent chemical oxygen demand homogeneous concentration | [mg L$^{-1}$] |
| Conductivity Reac. | Conductivity reactor | [mS cm-1] |
| Cycle Time | Cycle time of SBR | [min] |
| DO | Dissolved oxygen | [mg L$^{-1}$] |
| DO (Max) | Maximum dissolved oxygen concentration | [mg L$^{-1}$] |
| DO Average (ON) | Average dissolved oxygen concentration | [mg L$^{-1}$] |
| ERM | Empirical Risk minimization | Ad. |
| FF-ANN | Feed Forward artificial neural networks | Ad. |
| MAD | Mean average deviation | Ad. |
| MAPE | Mean absolute percentage error | Ad. |
| MAR | Missing values at random | Ad. |
| MSE | Mean squared error | Ad. |
| NH4-N Eff. | Ammonia concentration effluent | [mg N L-1] |
| NH4-N Inf. | Ammonia concentration in the effluent | [mg N L-1] |
| $NO_2$-N Eff. | Nitrite concentration effluent | [mg N L-1] |
| $NO_3$-N Eff. | Nitrate concentration effluent | [mg N L-1] |
| NOB | Nitrite oxidizing bacteria | [mg($NO_3$-N)g(oTS)$^{-1}\cdot$h$^{-1}$] |
| ORP max. | Maximum oxidation reduction potential | [mV] |
| ORP min. | Minimum oxidation reduction potential | [mV] |
| oTS Reac. | Organic Total solids Reac. | [%] |
| pH Reac. | pH reactor | Ad. |
| PN-A | Partial nitritation anammox | Ad. |
| Q Inf. | Volumetric flow influent | [m$^3$L$^{-1}$] |
| Reac. Time | Reaction time | [min cycle$^{-1}$] |
| RFE | Recursive feature elimination | Ad. |
| RM | Risk minimization | Ad. |
| RMSE | Root mean squared error | Ad. |
| SBR | Sequencing batch reactor | Ad. |
| SBRA | Sequencing batch reactor A | Ad. |
| SBRB | Sequencing batch reactor B | Ad. |
| Sedim. Time | Sedimentation time | [min] |
| SRM | Structural risk minimization | Ad. |
| SVI Reac. | Sludge volume index reactor | [mg L$^{-1}$] |
| SVM | Support vector machines | Ad. |
| Temperature Reac. | Temperature reactor | [°C] |
| Time of Air OFF | Aeration time blower OFF per aeration cycle. | [min] |
| Time of Air ON | Aeration time blower ON per aeration cycle. | [min] |
| TS Eff. | Total solids effluent | [mg L$^{-1}$] |
| TS Inf. | Total solids influent | [mg L$^{-1}$] |
| TS Reac. | Total Solids concentration in the reactor | [mg L$^{-1}$] |
| WWTP | WWTP | Ad. |
| ORP | Difference between Maximum and Minimum oxidation reduction potential | [mV] |

Table A. 3 Summary of measured parameters; average, maximum, minimum, standard deviation and number (No.) of observations during the operation of the SBRA and SBRB systems.

| | SBR$_A$ | | | | | SBR$_B$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Parameter/Variable** | **Mean** | **Max.** | **Min.** | $\sigma^2$ | **Obs.** | **Mean** | **Max.** | **Min.** | $\sigma^2$ | **Obs.** |
| Total Solids Influent [mg L$^{-1}$] | 236.8 | 480.0 | 28.0 | 76.7 | 123 | 238.8 | 480.0 | 80.0 | 75.2 | 120 |
| Volumetric Inflow [m$^3$L$^{-1}$] | 74.9 | 120.0 | 20.0 | 25.4 | 604 | 92.4 | 1000.0 | 135.0 | 51.5 | 612 |
| COD hom. Influent [mg L$^{-1}$] | 676.9 | 1000.0 | 340.0 | 103.4 | 79 | 674.4 | 1000.0 | 340.0 | 104.7 | 81 |
| COD filt. Influent [mg L$^{-1}$] | 323.3 | 460.0 | 170.0 | 47.7 | 79 | 321.3 | 460.0 | 135.0 | 51.5 | 81 |
| NH4-N influent [mg N L$^{-1}$] | 959.5 | 1173.0 | 798.0 | 100.6 | 143 | 955.7 | 1173.0 | 798.0 | 100.4 | 145 |
| Total Solids Reactor [g L$^{-1}$] | 2.2 | 4.6 | 0.2 | 0.9 | 146 | 2.7 | 4.9 | 1.8 | 0.7 | 150 |
| Organic Total Solids Reactor [%] | 67.0 | 77.0 | 55.5 | 6.1 | 62 | 71.5 | 77.5 | 64.5 | 4.1 | 63 |
| Sludge Volume Index Reactor [mL g$^{-1}$] | 72.5 | 120.0 | 33.3 | 21.3 | 131 | 78.6 | 135.0 | 42.0 | 26.2 | 133 |
| Total Solids Effluent [mg L$^{-1}$] | 297.5 | 3300.0 | 1.4 | 618.9 | 100 | 383.4 | 4000.0 | 0.5 | 637.0 | 111 |
| COD hom. Effluent [mg L$^{-1}$] | 403.5 | 3600.0 | 46.7 | 644.3 | 62 | 593.4 | 4160.0 | 109.0 | 781.1 | 56 |
| COD filt. Effluent [mg L$^{-1}$] | 121.8 | 205.0 | 34.0 | 39.8 | 74 | 127.0 | 230.0 | 35.3 | 41.8 | 73 |
| NH$_4$-N Effluent [mg N L$^{-1}$] | 209.2 | 594.0 | 48.3 | 84.5 | 262 | 227.4 | 435.0 | 36.0 | 97.8 | 260 |
| NO$_3$-N Effluent [mg N L$^{-1}$] | 81.1 | 297.0 | 0.2 | 72.9 | 262 | 105.1 | 255.0 | 20.0 | 45.8 | 259 |
| NO$_2$-N Effluent[mg N L$^{-1}$] | 0.4 | 7.6 | 0.0 | 0.9 | 263 | 0.8 | 25.0 | 0.0 | 2.4 | 261 |
| Sedimentation Time [min] | 15.5 | 30.0 | 1.0 | 10.3 | 612 | 21.3 | 40.0 | 8.0 | 6.6 | 616 |
| Reaction Time [min cycle$^{-1}$] | 325.9 | 335.0 | 242.0 | 9.2 | 612 | 324.6 | 346.0 | 321.0 | 3.7 | 616 |
| Cycle Time [min] | 359.9 | 360.0 | 283.0 | 3.1 | 612 | 360.0 | 360.0 | 360.0 | 0.0 | 616 |
| Time air (ON) [min] | 8.1 | 65.0 | 2.0 | 9.1 | 612 | 9.1 | 67.0 | 3.0 | 10.9 | 609 |
| Time air (OFF) [min] | 9.4 | 20.0 | 0.0 | 3.4 | 612 | 9.5 | 62.0 | 0.0 | 4.1 | 609 |
| Aeration time per Cycle [min cycle $^{-1}$] | 107.3 | 260.0 | 34.0 | 52.8 | 615 | 129.6 | 268.0 | 0.0 | 40.5 | 616 |
| pH Reactor [-] | 7.3 | 7.9 | 6.5 | 0.2 | 610 | 7.3 | 8.3 | 7.0 | 0.2 | 600 |
| Temperature Reactor [°C] | 31.1 | 36.0 | 25.8 | 2.6 | 608 | 31.3 | 36.6 | 26.0 | 2.8 | 593 |
| Conductivity Reactor [mS cm$^{-1}$] | 3049.4 | 4899.4 | 1372.8 | 694.0 | 602 | 3747.0 | 8729.9 | 2477.2 | 955.6 | 600 |
| DO Average (ON) [mg L$^{-1}$] | 0.3 | 0.6 | 0.0 | 0.1 | 601 | 0.2 | 0.5 | 0.0 | 0.1 | 600 |
| DO (Max) [mg L$^{-1}$] | 0.7 | 1.8 | 0.0 | 0.3 | 596 | 0.8 | 2.1 | 0.0 | 0.3 | 599 |
| ORP max. [mV] | 90.7 | 203.0 | -25.0 | 31.1 | 601 | 72.2 | 207.0 | -61.0 | 35.1 | 586 |
| ORP min. [mV] | -61.8 | 28.0 | -178.0 | 27.7 | 600 | -69.8 | 101.0 | -252.0 | 43.7 | 595 |
| ORP [mV] | 152.2 | 239.0 | 5.0 | 33.3 | 599 | - | - | - | - | - |

Table A. 4 Batch activity test datasets for Reactors SBR$_A$ (left) and SBR$_B$ (right)

| Date | SBR$_A$ | | | SBR$_B$ | | |
|---|---|---|---|---|---|---|
| | AOB<br>mg(NH$_4$-N)<br>g(oTS)$^{-1}$ · h$^{-1}$ | NOB<br>[mg(NO$_3$-N)<br>g(oTS)$^{-1}$ · h$^{-1}$] | AnAOB<br>[mg(N)<br>g(oTS)$^{-1}$ · h$^{-1}$] | AOB<br>mg(NH$_4$-N)<br>g(oTS)$^{-1}$ · h$^{-1}$ | NOB<br>[mg(NO$_3$-N)<br>g(oTS)$^{-1}$ · h$^{-1}$] | AnAOB<br>[mg(N)<br>g(oTS)$^{-1}$ · h$^{-1}$] |
| 04.11.2011 | 2.69 | 4.60 | 6.11 | 2.27 | 7.56 | 6.32 |
| 30.11.2011 | 3.43 | 2.08 | 7.54 | 4.25 | 7.92 | 2.23 |
| 14.12.2011 | 5.63 | 1.22 | 6.91 | 2.41 | 4.41 | 5.24 |
| 28.12.2011 | 8.99 | 1.06 | 6.70 | 4.18 | 3.92 | 2.90 |
| 12.01.2012 | 12.13 | 1.36 | 6.65 | 4.39 | 8.88 | 2.27 |
| 28.01.2012 | 12.58 | 0.90 | 7.96 | 8.01 | 12.94 | 3.64 |
| 08.02.2012 | 4.88 | 0.32 | 7.88 | 6.35 | 5.03 | 6.35 |
| 23.02.2012 | 1.15 | 0.00 | 10.02 | 2.04 | 6.91 | 5.34 |
| 07.03.2012 | 1.33 | 1.48 | | 7.19 | 6.37 | 7.00 |
| 21.03.2012 | 3.92 | 2.72 | 11.26 | 4.55 | 4.83 | 6.04 |
| 05.04.2012 | 4.03 | 1.54 | 6.76 | 4.72 | 3.71 | 2.26 |
| 17.04.2012 | 0.79 | 1.34 | 5.71 | 0.23 | 1.83 | 1.60 |
| 25.04.2012 | 7.41 | 2.18 | 8.24 | 8.32 | 0.82 | 2.50 |
| 10.05.2012 | 1.24 | 1.76 | 6.45 | 2.28 | 6.38 | 2.00 |
| 16.05.2012 | 2.09 | 3.35 | 5.08 | 5.37 | 0.82 | 5.75 |
| 06.06.2012 | 3.28 | 4.91 | 6.69 | 11.14 | 3.08 | 5.83 |
| 26.06.2012 | 3.00 | 2.46 | 2.18 | 15.36 | 5.00 | 12.50 |
| 17.07.2012 | 4.19 | 3.31 | 9.61 | 7.11 | 0.16 | 9.80 |
| 07.08.2012 | | 5.28 | 6.11 | | 2.03 | 4.95 |
| 05.09.2012 | 4.88 | 5.61 | 0.00 | 19.08 | 8.62 | |
| 23.10.2012 | | | 9.48 | | | 12.68 |
| 15.11.2012 | 16.94 | 4.72 | 12.05 | 21.79 | 3.96 | 12.42 |
| 05.02.2013 | 19.50 | 3.33 | 14.51 | 11.11 | 3.33 | 10.23 |
| 09.05.2013 | 4.74 | 4.10 | 6.89 | 5.44 | 2.93 | 6.19 |
| 28.05.2013 | 15.43 | 4.97 | 13.40 | 14.76 | 4.16 | 9.70 |

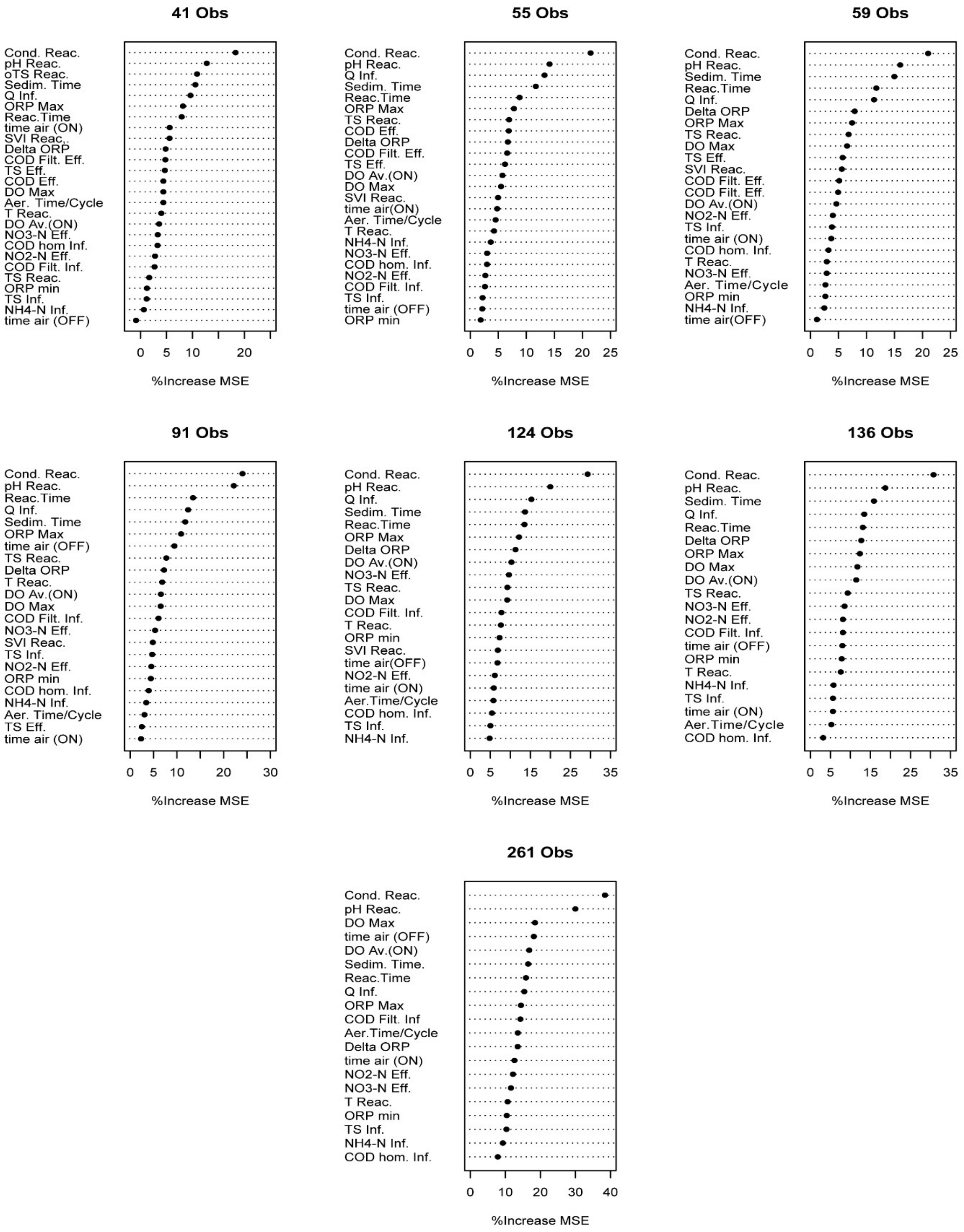


Figure A. 14 Ranking of importance obtained from the Feature Selection with Random Forest: NH4-N Eff (SBRA)

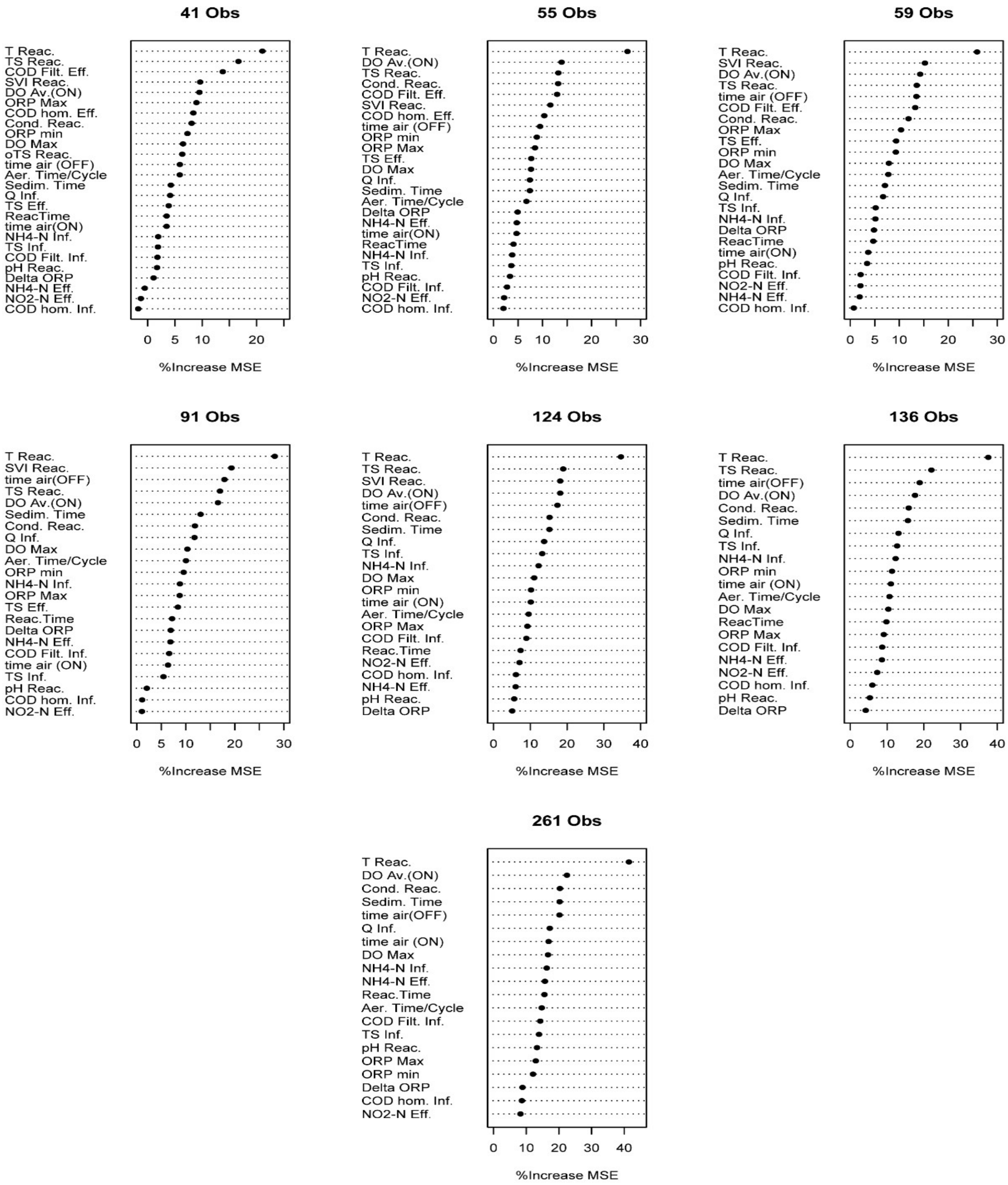


Figure A. 15 Ranking of importance obtained from the Feature Selection with Random Forest: NO3-N Eff.(SBRA)

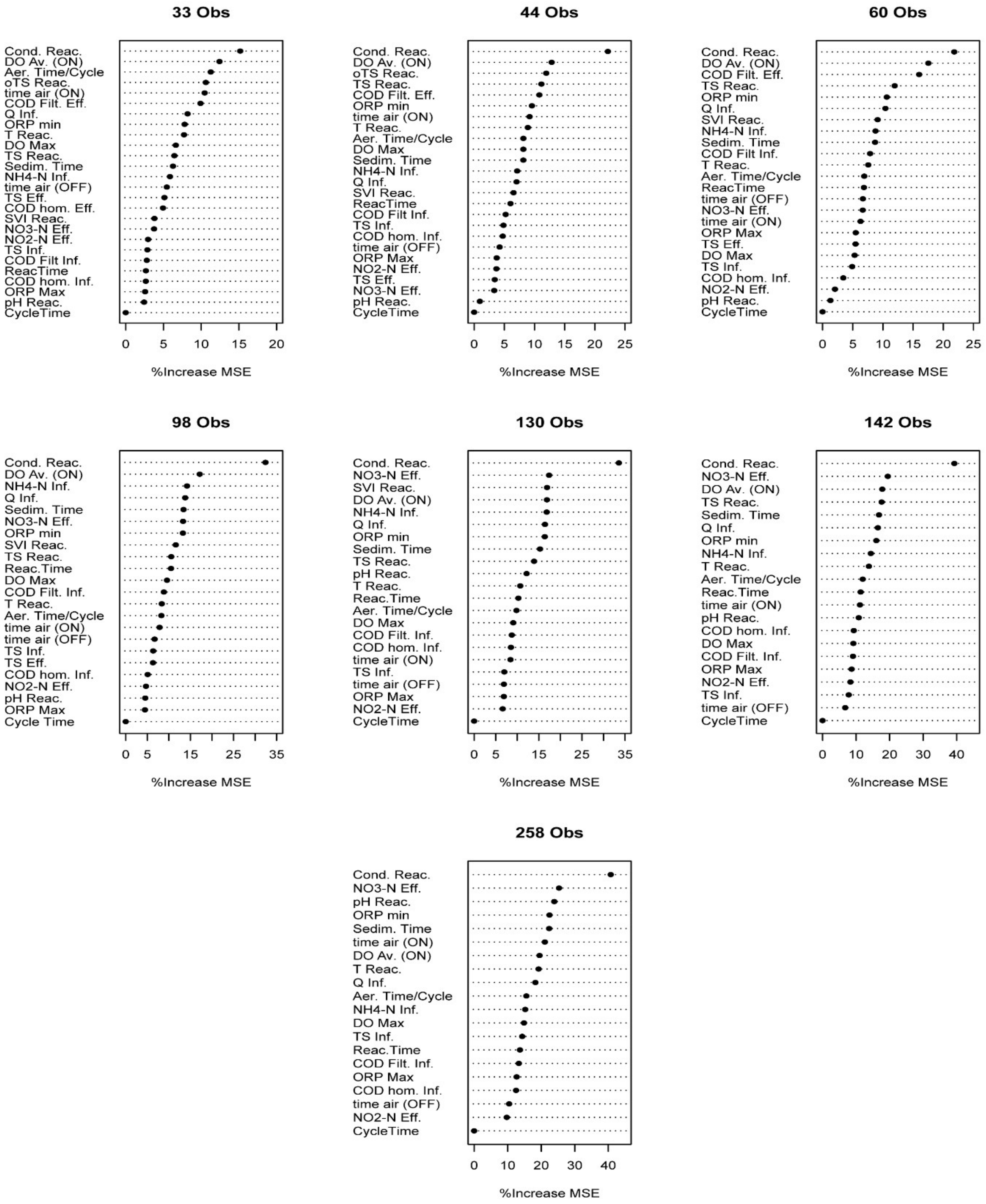


Figure A. 16 Ranking of importance obtained from the Feature Selection with Random Forest: $NH_4$-N Eff. ($SBR_B$)

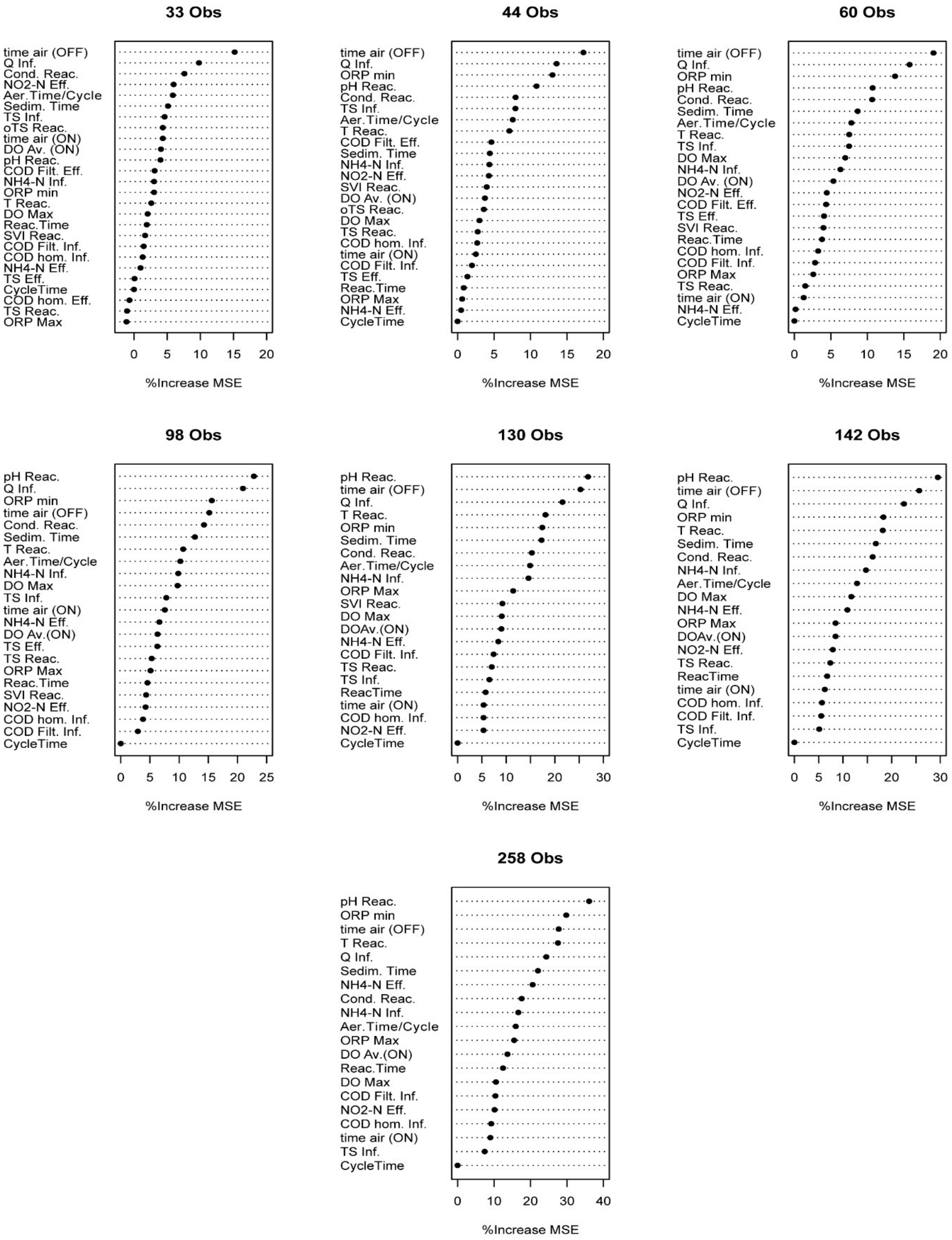


**Figure A. 17** Ranking of importance obtained from the Feature Selection with Random Forest: $NO_3$-N Eff. ($SBR_B$)

**Table A. 5** Influence of different variables on the *%Inc. MSE* parameter (*Random Forest*)

| Parameters | $SBR_A$ | | $SBR_B$ | |
|---|---|---|---|---|
| | $NH_4$-N Eff. | $NO_3$-N Eff. | $NH_4$-N Eff. | $NO_3$-N Eff. |
| Cond. Reac. | >10% | >10% | >10% | >10% |
| Q inf. | >10% | >10% | >10% | >10% |
| ORP min | <5% | | >10% | >10% |
| pH Reac. | >10% | <5% | <5% | >10% |
| Sedim. Time | >10% | >10% | | >10% |
| Temp Reac. | | >10% | | >10% |
| time air OFF | | >10% | | >10% |
| Aer time /Cycle | <5% | | | >10% |
| Reac. Time | >10% | <5% | >10% | <5% |
| Cycle time | | | <5% | <5% |
| SVI Reac. | | >10% | | <5% |
| COD hom. Inf. | <5% | <5% | | <5% |
| COD filt. Inf. | | <5% | | <5% |
| TS Reac. | | >10% | >10% | |
| DO Av. ON | | >10% | >10% | |
| $NH_4$-N Inf. | <5% | | >10% | |
| $NO_3$-N Eff. | | | >10% | |
| $NO_2$-N Eff. | <5% | <5% | <5% | |
| DO Max | | >10% | | |
| ORP | | <5% | | |
| $NH_4$-Eff. | | <5% | | |
| time air ON | | <5% | | |
| ORP Max | >10% | | | |
| TS Inf. | <5% | | | |

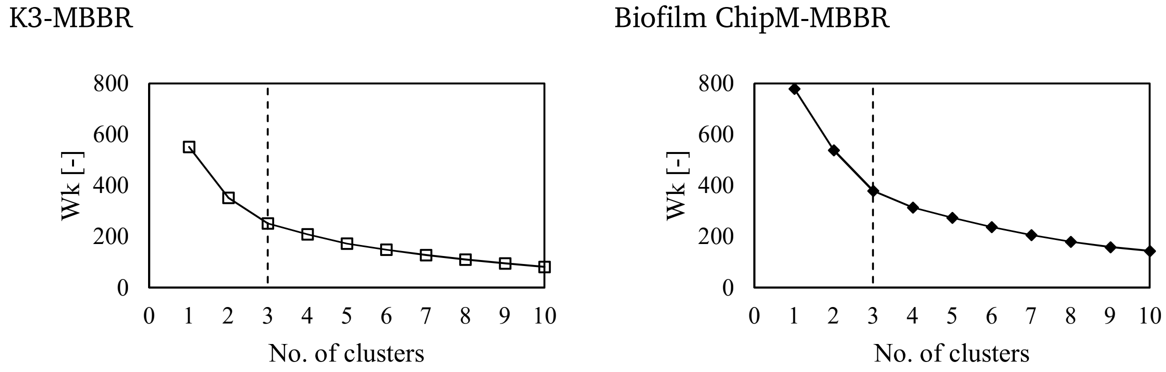


**Figure A. 18** Selection of number of clusters in k-means clustering. The within clusters sum of squares (Wk) stabilizes after 3 clusters (slope becomes constant).

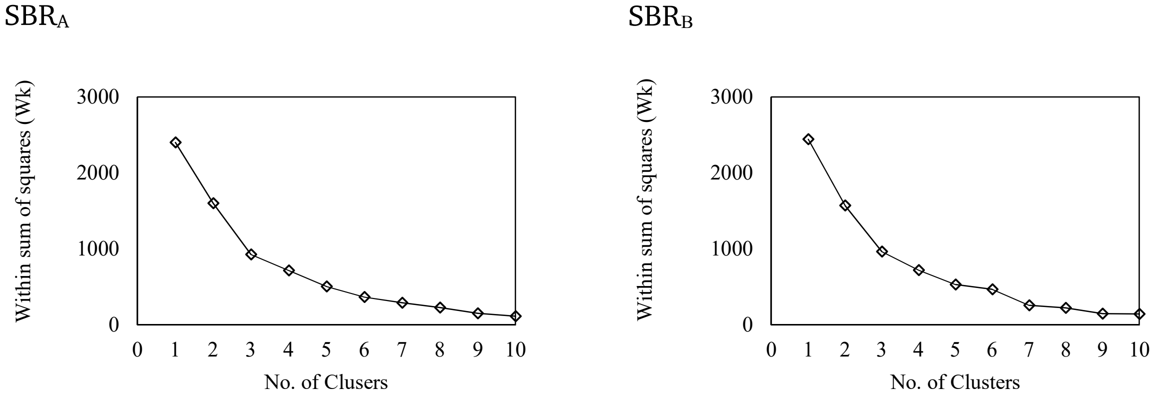


Figure A. 19 Elbow-method criterion for SBRB. The within clusters sum of squares (k-means clustering parameter) is plotted against different number of clusters.

Table A. 6 Coordinates from the clusters centroids obtained from the k-means clustering analysis developed for K3 and Chip M moving bed biofilm reactors.

| Parameter | K3 | | | Chip M | | |
|---|---|---|---|---|---|---|
| | Clus. 1 | Clus. 2 | Clus. 3 | Clus. 1 | Clus. 2 | Clus. 3 |
| AnAOB [mg N $L^{-1}$ $d^{-1}$] | 76.61 | 11.59 | 29.5 | 265.8 | 87.4 | 21.7 |
| AOB [mg N $L^{-1}$ $d^{-1}$] | 41.9 | 8.2 | 42.3 | 102.7 | 238.3 | 16.6 |
| NOB [mg N $L^{-1}$ $d^{-1}$] | 108.2 | 34.2 | 70.7 | 192.4 | 168.5 | 45.1 |
| HB [mg $L^{-1}$ $d^{-1}$] | 0.0 | 35.5 | 0.0 | 0.0 | 0.0 | 42.4 |
| NH4-N Inf. [mg N $L^{-1}$] | 75 | 47.19 | 45.55 | 103.20 | 47.95 | 47.84 |
| NO2-N Inf. [mg N $L^{-1}$] | 2.56 | 2.02 | 5.01 | 6.00 | 2.87 | 0.96 |
| NO3-N Inf. [mg N $L^{-1}$] | 2.11 | 0.31 | 3.45 | - | - | - |
| COD Inf. [mg $L^{-1}$] | 0 | 45 | 0 | 0 | 0 | 42 |
| COD Filt. Inf. [mg $L^{-1}$] | 0 | 29 | 0 | 0 | 0 | 28 |
| BOD5 Inf. [mg $L^{-1}$] | 0 | 25 | 0 | 0 | 0 | 20 |
| COD Eff. [mg $L^{-1}$] | 0 | 35.25 | 0 | 0 | 0 | 44.94 |
| pH Reac. | 7.16 | 7.35 | 7.18 | - | - | - |
| DO Reac. [mg $L^{-1}$] | 0.30 | 0.39 | 0.20 | 0.45 | 0.21 | 0.53 |
| ORP Reac. [mV] | 323 | 98 | 200 | 335 | 251 | 44 |
| Cond. Reac. [mS $cm^{-1}$] | 1.64 | 1.42 | 1.09 | 1.80 | 1.76 | 1.08 |

| | | | | | | |
|---|---|---|---|---|---|---|
| HRT [d] | 1.92 | 0.95 | 1.9 | 1.3 | 2.5 | 1.1 |
| NH4-N Eff. [mg N L$^{-1}$] | 9.04 | 5.18 | 13.33 | 8.00 | 6.13 | 8.66 |
| NO2-N Eff. [mg N L$^{-1}$] | 0.41 | 4.52 | 14.50 | 2.28 | 2.07 | 5.98 |
| NO3-N Eff. [mg N L$^{-1}$] | 25.32 | 20.42 | 9.01 | 37.6 | 10.4 | 13.8 |
| Temp. Reactor [°C] | 17.8 | 13.5 | 10.7 | 20.3 | 14.0 | 14.4 |

Table A. 7 Coordinates for the clusters centroids found through the clustering datasets for SBRA and SBRB.

| Parameters | SBR$_A$ | | | SBR$_B$ | | |
|---|---|---|---|---|---|---|
| | Clus.1 | Clus. 2 | Clus. 3 | Clus.1 | Clus. 2 | Clus. 3 |
| AOB [mg $NH_4$-N g (oTS)$^{-1}$ · h$^{-1}$] | 6.4 | 10.9 | 3.2 | 12.4 | 4.6 | 4.2 |
| NOB [mg $NO_3$-N g (oTS) $^{-1}$ · h$^{-1}$] | 1.4 | 4.3 | 2.5 | 3.2 | 6.9 | 2.7 |
| AnAOB [mg N g (oTS) $^{-1}$ · h$^{-1}$] | 7.5 | 9.4 | 6.5 | 9.5 | 4.7 | 2.8 |
| Q Inf. [m$^3$ d$^{-1}$] | 46.2 | 86.9 | 52.5 | 110.8 | 65.2 | 40.1 |
| $NH_4$-N Inf. [mg N L$^{-1}$] | 997.5 | 943.3 | 1052.3 | 940.3 | 998.0 | 1092.6 |
| ReacTime [min] | 322.5 | 327.7 | 316.1 | - | - | - |
| Aer. Time/ Cycle [min] | - | - | - | 117.3 | 168.0 | 83.2 |
| Time air ON [min] | - | - | - | 6.0 | 11.4 | 6.4 |
| Time air OFF [min] | 8.3 | 8.2 | 12.0 | 9.9 | 7.6 | 15.9 |
| ORP min [mV] | | | | -91.4 | -20.7 | -62.2 |
| ORP max [mV] | 58.2 | 104.7 | 90.6 | - | - | - |
| Sed. Time [min] | 25.1 | 9.1 | 29.5 | 20.7 | 22.6 | 30.0 |
| pH Reactor [-] | 7.3 | 7.3 | 7.4 | 7.3 | 7.2 | 7.5 |
| Temperature Reac. [°C] | 28.4 | 32.5 | 31.8 | 32.3 | 28.9 | 31.3 |
| Conductivity Reactor [mS cm$^{-1}$] | 2254.6 | 2967.4 | 3920.8 | 3597.1 | 4056.0 | 5724.7 |
| DO Av. ON [mg L$^{-1}$] | 0.1 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 |
| DO Max [mg L$^{-1}$] | 0.3 | 0.8 | 0.7 | - | - | - |
| TS Reactor [g L$^{-1}$] | 0.7 | 2.6 | 2.7 | 2.7 | 3.0 | 2.6 |
| $NH_4$-N Eff. [mg N L$^{-1}$] | 207.7 | 140.7 | 283.4 | 151.2 | 296.3 | 297.1 |
| $NO_3$-N Eff. [mg N L$^{-1}$] | 5.2 | 126.9 | 103.5 | 94.2 | 102.2 | 178.2 |

## 10.4 Feature engineering: Yeo-Johnson transformation

Yeo-Johnson transformation is defined for numeric feature with a domain in ℝ, this transformation is appropriate for reducing skewness and to approximate normality (Yeo and Johnson, 2000).

$$\mathrm{X}_{\mathrm{transformed}} = \begin{cases} \frac{(X+1)^{\lambda}-1}{\lambda} & if\ \lambda \neq 0\ and\ X \geq 0 \\ \ln(X+1) & if\ \lambda = 0\ and\ X \geq 0 \\ -\frac{(-X+1)^{2-\lambda}-1}{2-\lambda} & if\ \lambda \neq 0\ and\ X < 0 \\ -\ln(-X+1) & if\ \lambda = 2\ and\ X < 0 \end{cases} \qquad \text{Equation A. 1}$$

*Ama Sua*
*Ama Llulla*
*Ama Quella*
(*Quechua*)

Don’t be a thief
Don’t be a liar
Don’t be lazy
(*English*)

Zur Autorin:

Luz Alejo wurde 1991 in Concepción, Chile, geboren. Sie studierte Chemieingenieurwesen an der Universidad de Concepción und erhielt 2015 ihren Abschluss in Verfahrenstechnik. Sie nahm aktiv an Forschungsprojekten in der Fakultät für Chemieingenieurwesen ihrer Alma Mater teil und präsentierte ihre Arbeiten auf nationalen und internationalen Konferenzen und Symposien. Während ihres Studiums an der Universidad de Concepción erhielt sie Stipendien für Forschung in der Minenindustrie sowie für Studienaufenthalte im Ausland. Ende 2017 begann sie ihre Doktorarbeit im IWAR-Institut am Lehrstuhl für Abwasserwirtschaft unter der Leitung von Prof. Dr. Susanne Lackner und Dr. John Atkinson von der Universität Adolfo Ibañez in Santiago de Chile. Seit ihrer Ankunft am IWAR-Institut im Jahr 2017 beschäftigte sich Luz Alejo mit der Forschung im Zusammenhang mit datenintensiven Fragestellungen und der Anwendung datengestützter Ansätze im Wassersektor. Sie analysierte verschiedene Datenquellen, untersuchte die Qualität der Daten und entwickelte Methoden für die Datenanalyse im Wassersektor im Hinblick auf die Entdeckung von Wissen.

Zum Inhalt:

Das maschinelle Lernen (ML) ist eines der am schnellsten wachsenden technischen Gebiete, das an der Schnittstelle von Informatik und Statistik liegt und den Kern der künstlichen Intelligenz (KI) und der Datenwissenschaft bildet. Im Bereich der Abwasserbehandlung ging der Ursprung der umfangreichen Datengenerierung mit der Automatisierung von Kläranlagen (WWTP) einher. Die Informationen von Kläranlagen, die erzeugt und aufgezeichnet werden, umfassen komplexe und heterogene Datenquellen; online von Sensoren, on/off Steuerdaten von Pumpen und Geräten und off-line Messungen von Laboratorien. Sensoren sind in der Lage, Messungen alle paar Sekunden aufzuzeichnen und so täglich Tausende von Datenpunkten zu generieren. Die in Laboratorien bei der Abwasserbehandlung erzeugten Daten sind entscheidend für die Bewertung der Wasserqualität in jedem biologischen Abwasserbehandlungsprozess (bWWTP) und oft auch für die Validierung der Sensorinformationen. Aufgrund der Kosten und des Zeitaufwands ist die Häufigkeit der Probenahmen für Labormessungen im Vergleich zu den Sensoren jedoch oft drastisch reduziert. Die daraus resultierende Datenbank (aus Sensoren und Labors) beinhaltet daher unterschiedliche Probenahmehäufigkeiten und somit einen sehr heterogenen Datensatz. Diese Arbeit zeigt, wie wichtig die Datenauswahl in Kläranlagen-Datensätzen ist, um zuverlässige Informationen zu erhalten. In dieser Arbeit wurden verschiedene Methoden zur Analyse heterogener Datensätze im Wassersektor entwickelt.

ISBN 978-3-940897-65-7