UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos

Amanda Vitória Santos

IMPROVING MEMBRANE FOULING CONTROL AND AMMONIA REMOVAL ON MEMBRANE BIOREACTORS FROM A DATA-DRIVEN APPROACH

Belo Horizonte
2022

Amanda Vitória Santos

IMPROVING MEMBRANE FOULING CONTROL AND AMMONIA REMOVAL ON MEMBRANE BIOREACTORS FROM A DATA-DRIVEN APPROACH

Tese apresentada ao Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutora em Saneamento, Meio Ambiente e Recursos Hídricos.

Área de concentração: Meio Ambiente

Linha de pesquisa: Caracterização, prevenção e controle da poluição

Orientador: Profa. Dra. Sílvia Maria Alves Correa Oliveira

Coorientador: Profa. Dra. Míriam Cristina Santos Amaral

Belo Horizonte
2022

Santos, Amanda Vitória.
S237i Improving membrane fouling control and ammonia removal on membrane bioreactors from a data-driven approach [recurso eletrônico] / Amanda Vitória Santos. – 2022.
1 recurso online (167 f.: il., color.) : pdf.

Orientadora: Sílvia Maria Alves Correa Oliveira.
Coorientadora: Míriam Cristina Santos Amaral.

Tese (doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Anexos: f. 150-167.

Bibliografia: f. 136-149.
Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia sanitária - Teses. 2. Meio ambiente - Teses.3. Análise de componentes principais - Teses. 4. Controle estatístico do processo - Teses. 5. Inteligência artificial - Teses. 6. Redes Neurais (Computação) - Teses. I. Oliveira, Silvia Maria Alves Corrêa. II. Amaral, Míriam Cristina Santos. III. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.

CDU: 628(043)

Ficha catalográfica elaborada pelo bibliotecário Reginaldo César Vital dos Santos CRB/6 2165
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS
GERAIS[ESCOLA DE ENGENHARIA]
COLEGIADO DO CURSO DE GRADUAÇÃO / PÓS-GRADUAÇÃO EM [SANEAMENTO, MEIO AMBIENTE E RECURSOSHÍDRICOS]

FOLHA DE APROVAÇÃO

["Improving Membrane Fouling Control And Ammonia Removal On Membrane Bioreactors From AData-driven Approach"]

AMANDA VITÓRIA SANTOS

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

Profa Sílvia Maria Alves Correa Oliveira - Orientadora

Profa Miriam Cristina Santos Amaral Moravia - coorientadora

Prof. Eduardo Coutinho de Paula

Prof. Argimiro Resende Secchi

Prof. Cristiano Christofaro Matosinhos

Prof. Eduardo Lucas Subtil

Aprovada pelo Colegiado do PG SMARH

Versão Final aprovada por

Profa. Priscilla Macedo Moura

Coordenadora

Profª. Sílvia Maria Alves Corrêa Oliveira

Orientadora

Belo Horizonte, 28 de outubro de 2022.

Documento assinado eletronicamente por **Silvia Maria Alves Correa Oliveira**, **Professora do Magistério Superior**, em 31/10/2022, às 11:27, conforme horário oficial de Brasília, com fundamentono art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Coutinho de Paula**, **Professor do Magistério Superior**, em 31/10/2022, às 14:33, conforme horário oficial de Brasília, com fundamento no art. 5ºdo Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristiano Christofaro Matosinhos**, **Usuário Externo**, em 31/10/2022, às 15:07, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Miriam Cristina Santos Amaral Moravia**, **Professora do Magistério Superior**, em 31/10/2022, às 15:24, conforme horário oficial de Brasília, com fundamentono art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Lucas Subtil**, **Usuário Externo**, em 04/11/2022, às19:05, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Argimiro Resende Secchi**, **Usuário Externo**, em 17/11/2022, às 17:09, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Priscilla Macedo Moura**, **Coordenador(a) de curso de pós-graduação**, em 18/11/2022, às 10:46, conforme horário oficial de Brasília, com fundamento no art. 5ºdo Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando ocódigo verificador **1860096** e o código CRC **FD1FD8D0**.

Referência: Processo nº 23072.263932/2022-37 SEI nº 1860096

À Marcelinha, uma parte de mim.

AGRADECIMENTOS

Agradeço às agências CNPq, CAPES e FAPEMIG, à UFMG e ao PPG-SMARH pelos recursos oferecidos para a realização deste trabalho e para minha formação profissional.

Agradeço à refinaria de petróleo pelo fornecimento dos dados de monitoramento.

Agradeço às minhas orientadoras, Profa. Sílvia Oliveira e Profa. Míriam Amaral, pelo apoio e suporte; e aos membros avaliadores da banca de defesa, Prof. Argimiro Secchi, Prof. Eduardo Suptil, Prof. Cristiano Matosinhos e Prof. Eduardo Coutinho por terem aceitado o convite para participar da avaliação e pelas contribuições ao trabalho.

Agradeço aos colegas dos grupos de pesquisa GEAPS Membranas e GETEDA pelo convívio, pelas dicas partilhadas, pelas conversas e pelo companheirismo que tornaram os dias mais leves.

Agradeço aos meus pais por terem possibilitado que eu vivesse esse e tantos outros importantes momentos e pelos ensinamentos que transcendem o acadêmico, que ensinam a viver. Obrigada por, mesmo nem sempre entendendo, nunca terem deixado de me apoiar nos caminhos que eu escolhi.

Agradeço à minha irmã por ser minha maior certeza de que tudo sempre vai ficar bem. Obrigada por compartilhar a vida comigo e por ter me dado o presente mais valioso do mundo. Agradeço também ao Pedroca, por ser calma e sensatez em uma família um tanto quanto agitada. Afinal, não teria penta sem Cafu!

Agradeço ao Moisés, por ser minha fonte inesgotável de coragem, paz e felicidade. Por dividir os dias, o cansaço, as alegrias, as frustrações e as conquistas. Sem seu apoio diário, não seria possível chegar até aqui. Obrigada por me impulsionar.

Agradeço ao Bauducco, por preencher todos os dias com o amor mais puro e sincero.

Agradeço com muito carinho à Letícia, por me mostrar que os desafios podem ser carregados como balões.

Agradeço a todos os meus amigos por serem luz, alegria e cor. De maneira especial, agradeço à Laurie por juntas sermos mar. Ao Éder e à Katiane, por terem feito a mudança de cidade ser

mais leve. À Anne e ao Paulin, por terem dividido as alegrias e dificuldades da vida acadêmica desde o início da caminhada. À Bella, por ser porto seguro.

Agradeço aos meus avós por serem inspiração e força.

E, por fim, mas acima de tudo, agradeço a Deus, por estar sempre comigo, iluminando meu caminho e guiando meus passos. Agradeço ao Senhor por ter abençoado todos eles até aqui. E que os próximos que virão também assim sejam.

A todos vocês, meu amor.

“Without data, you are just another person with an opinion”.
W. Edwards Deming

RESUMO

Biorreatores com membrana (BRM) têm sido aplicados com sucesso no tratamento de esgotos e de efluentes industriais. No entanto, apesar de sua ampla aplicação, esta tecnologia ainda é restrita pela incrustação da membrana. Como o controle da incrustação é uma tarefa complexa e que demanda a investigação de um grande conjunto de variáveis frequentemente correlacionadas, a aplicação de técnicas de inteligência artificial (AI) e de aprendizado de máquina (ML) é uma boa alternativa para melhor monitorar e controlar a incrustação nesses sistemas. Além disso, o reuso da água tratada é um importante desafio para as indústrias contemporâneas e, consequentemente, alcançar efluentes tratados de alta qualidade é fundamental. Em especial para as refinarias de petróleo, a remoção de amônia é um árduo desafio, que também está relacionado a uma série de variáveis que impactam no desempenho do BRM. Portanto, modelos de AI/ML são também uma boa alternativa para monitorar e controlar a remoção de amônia. Logo, este trabalho se propõe a auxiliar a definição de estratégias para melhor controle da incrustação da membrana e da remoção de amônia em BRM aplicando técnicas de AI/ML, especificamente Análise por Componentes Principais (ACP), Redes Neurais Artificiais (RNA) e Controle Estatístico de Processos Multivariado (CEPM). Para tanto, dados de monitoramento de um BRM em escala piloto atuando em uma refinaria de petróleo foram considerados. Os modelos foram desenvolvidos em R e aplicados para investigar diferentes relações entre variáveis, modelar o comportamento do sistema e detectar e diagnosticar falhas relacionadas à incrustação e à baixa remoção de amônia, a fim de compreender suas principais causas e propor estratégias eficientes para o seu controle. O modelo ACP identificou as variáveis filtrabilidade do lodo, temperatura e número de dias sem limpeza química como as mais influentes na incrustação da membrana e foi eficaz na previsão do desempenho do BRM (R^2 = 0,71 e Q^2 = 0,78), permitindo detectar amostras atípicas e identificar problemas operacionais. As cartas de controle T^2 e Q detectaram 100 e 96%, respectivamente, da operação com baixa permeabilidade da membrana, ressaltando sua alta capacidade de detecção de falhas. As cartas de controle também foram capazes de alertar preventivamente sobre a diminuição da permeabilidade, logo elas podem ser utilizadas para guiar a tomada de decisão em relação ao controle da incrustação, orientando, por exemplo, quando realizar limpezas químicas e/ou dosar melhoradores de permeabilidade. Em relação à remoção de amônia, os modelos RNA e ACP identificaram que a concentração de óleos e graxas e a demanda química de oxigênio (DQO) afluentes, em conjunto com a permeabilidade da membrana, contribuem para menores remoções de amônia, enquanto tempo de retenção de lodo e temperatura estão relacionados a maiores remoções. O modelo RNA também previu efetivamente a remoção de amônia a partir de um conjunto de condições operacionais, com R^2 igual a 0,87. Além disso, a carta de controle Q detectou 100% da operação com remoções inferiores a 85%, o que poderia possibilitar uma atuação mais efetiva no sistema, por exemplo ajustando a temperatura e/ou mantendo maiores tempos de retenção do lodo, e evitar que a remoção desse poluente atingisse níveis mais baixos. Portanto, conclui-se que a modelagem de BRM por meio de AI e ML é uma interessante ferramenta para monitorar, compreender e prever o comportamento do sistema. Os modelos construídos a partir de RNA, ACP e CEPM podem ser aplicados como ferramentas de apoio à tomada de decisão quanto ao controle da incrustação da membrana e melhora da remoção de amônia, contribuindo para melhores desempenhos e, consequentemente, para operações mais eficientes de BRM.

Palavras-chave: Tratamento de efluentes. Inteligência Artificial (IA). Redes Neurais Artificiais (RNA). Análise de Componentes Principais (ACP). Controle Estatístico de Processos Multivariado (CEPM).

ABSTRACT

Membrane bioreactors (MBR) have been successfully applied in the treatment of domestic and industrial wastewater. However, despite its wide application, this technology is still restricted by membrane fouling. As membrane fouling control is a complex task that demands the investigation of a large set of frequently correlated variables, Artitifical Intelligence (AI) and Machine Learning (ML) techniques are interesting alternatives to better monitor and control membrane fouling in these systems. In addition, water reuse is an important challenge for nowadays industries and, thus, achieving high quality treated effluents is critical. Especially for oil refineries, the removal of ammonia is an arduous task that is also influenced by many correlated variables that impact MBR performance. Therefore, AI/ML models are a equally promising alternative to monitor and control ammonia removal. In this context, the modelling techniques Artificial Neural Networks (ANN), Principal Component Analysis (PCA) and Multivariate Statistical Process Control (MSPC) have been highlighted in the literature. Therefore, this work aims to help define strategies for better control of membrane fouling and ammonia removal in MBR by applying PCA, ANN and MSPC. For that, monitoring data from a pilot-scale MBR operating in an oil refinery were considered. The models were developed in R and applied to investigate different relations between variables, to model the behavior of the system and to detect and diagnose failures related to membrane fouling and low ammonia removal capacity, in order to understand their main causes and propose efficient strategies for their control. The PCA model identified the variables sludge filterability, temperature and number of days without chemical cleaning as the most influential on membrane fouling and it was effective in predicting the MBR performance ($R^2 = 0.71$ and $Q^2 = 0.78$), making it possible to detect atypical samples and identify operational problems. T^2 and Q multivariate control charts detected 100 and 96%, respectively, of the operation with low membrane permeability, underlining their high fault detection capacity. The control charts were also able to provide preventive warnings about the decrease in membrane permeability so they can be used to support decision-making regarding membrane fouling control, guiding, for example, when to perform chemical cleanings or to dose membrane permeability improvers. Regarding ammonia removal, ANN and PCA models identified that the influent concentration of oil and grease and chemical oxygen demand (COD), together with membrane permeability, contribute to lower ammonia removals, while sludge retention time and temperature are related to higher removals. ANN model also effectively predicted ammonia removal from a set of input operating conditions, with R^2 equal to 0.87. Furthermore, Q control chart detected 100% of the operation with removals below 85%, which could allow a more effective action on the system, for example by adjusting the temperature and/or maintaining longer sludge retention times, and preventing the ammonia removal from reaching lower levels. Hence, it can be concluded that MBR modeling through AI and ML is an interesting tool to monitor, understand and predict the behavior of the system. Models built from ANN, PCA and MSPC can be applied as decision support tools regarding membrane fouling control and improved ammonia removal, contributing to better performances and, consequently, to more efficient MBR operations.

Keywords: Wastewater Treatment. Artificial Intelligence (AI). Artificial Neural Networks (ANN). Principal Components Analysis (PCA). Multivariate Statistical Process Control (MSPC).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Wastewater Treatment Technologies	
CAS	Conventional Activated Sludge
DM	Dynamic Membranes
MBBR	Moving Bed Biofilm Reactor
MBR	Membrane Bioreactors
MF	Microfiltration
MSP	Membrane Separation Processes
OMBR	Osmotic Membrane Bioreactor
PAC	Powdered Activated Carbon
PES	Polyether Sulfone
UF	Ultrafiltration
Statistical Techniques and Parameters	
AI	Artificial Intelligence
ANFIS	Adaptive Network-based Fuzzy Inference System
ANN	Artificial Neural Networks
AOC	Abnormal Operating Conditions
CL	Central Line
DModX	Distance to the model
DNN	Deep Neural Networks
ELM	Extreme Learning Machine
FA	Factor Analysis
FL	Fuzzy Logic
FNN	Feed-forward Neural Networks
GA	Genetic Algorithm
IoT	Internet of Things
IQR	Interquartile Range
KNN	K-Nearest Neighbours
LCL	Lower Control Limit
LSTM	Long-short Term Memory Network
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MCUSUM	Multivariate Cumulative Sum
MEWMA	Multivariate Exponential Weighted Moving Average
MI	Moving Interval
ML	Machine Learning
MSE	Mean Squared Error
MSPC	Multivariate Statistical Process Control
MSPC-PCA	MSPC based on PCA
NOC	Normal Operating Conditions
PC	Principal Components
PCA	Principal Components Analysis
PLS	Partial Least Square
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
SGD	Stochastic Gradient Descent

SNN	Spiking Neural Networks
SPC	Statistical Process Control
SPE	Squared Prediction Errors
SVM	Suport Vector Machine
UCL	Upper Control Limit
UWL	Upper Warning Limit

Physicochemical and Biological Parameters

AOB	Ammonia-Oxidizing Bacteria
BOD	Biochemical Oxygen Demand
CF	Concentration Factor
COD	Chemical Oxygen Demand
CST	Capillary Suction Time
CTA	Cellulose Triacetate
DO	Dissolved Oxygen
DOM	Dissolved Organic Matter
DS	Dray Solution
EfOM	Effluent Organic Matter
EPS	Extracellular Polymeric Substances
HRT	Hydraulic Retention Time
MLSS	Mixed Liquor Suspended Solids
MLVSS	Mixed Liquor Volatile Suspended Solids
OG	Oil and Grease
OLR	Organic Load Rate
SDWC	Sequential Days Without Cleaning
SFI	Sludge Filtration Index
SMP	Soluble Microbial Products
SRT	Sludge Retention Time
TMP	Transmembrane Pressure
TOC	Total Organic Carbon
VFA	Volatile Fatty Acids

Others

CAGR	Compound Annual Growth Rate
DFC	Delft Filtration Characterization
EPA	Environmental Protection Agency
ESE	Environmental Science and Engineering
FBG	Fluid Bed Granulation
FT	Filter Test
MQL	Method Quantification Limit
NIR	Near-infrared
PES	Polyether Sulfone
TTF	Time to Filter

APRESENTAÇÃO

For better organization, this document has been divided into five chapters: i) Contextualization; ii) Artificial Neural Networks, Principal Components Analysis and Multivariate Statistical Process Control: theoretical foundation and applications; iii) Improving membrane fouling control; iv) Improving ammonia removal; v) Final Considerations.

In Chapter I, the theme addressed in the work is presented, as well as the context in which it is inserted. The importance of membrane bioreactors (MBR) for industrial wastewater treatment, the seriousness of membrane fouling and ammonia removal and how the application of data-driven approaches like artificial intelligence (AI) and machine learning (ML) is a promising alternative for better monitoring and control of industrial processes and, thus, for process improvement, are presented and discussed. The relevance and innovation of the work are also highlighted and its hypothesis and goals are presented. In Chapter II, a comprehensive literature review is presented. It includes the theoretical foundation of the AI/ML techniques Artificial Neural Networks (ANN), Principal Component Analysis (PCA) and Multivariate Statistical Process Control (MSPC) and the presentation of important papers on the application of ANN and PCA for monitoring and controlling membrane fouling and MBR performance and the application of MSPC to detect and diagnose operating faults in various industrial processes. In Chapter III, the occurrence of membrane fouling on MBR is assessed. PCA and MSPC models were applied to understand and predict membrane fouling, as well as to identify and diagnose its occurrence. Based on the results, efficient strategies for its control are recommended. Similarly, in Chapter IV, ANN, PCA and MSPC models were applied to investigate ammonia removal on MBR. The models were also able to map and predict the system's behavior and to guide efficient strategies for its improvement. The results obtaneid with all three models are presented and discussed. Finally, all the results obtained within the study are discussed in an integrated manner in Chapter V, which also presents the conclusions and recommendations of the work.

SUMÁRIO

I. CONTEXTUALIZATION

1 INTRODUCTION

Contamination of natural water by inappropriate disposal of industrial residues and wastewater is currently one of the greatest environmental harms. The presence of emerging and persistent pollutants in industrial wastewater is an important global concern that has been leading to increasingly stringent environmental regulations. Furthermore, the demand for water reuse is growing fast, especially in industries. Therefore, the application of highly efficient wastewater treatment technologies that provide high quality treated water and allow its reuse has been increasingly sought. In fact, in recent works, the efficiency has been suggested as the major sustainability parameter for selecting a technology among the current alternatives (KAMALI *et al.*, 2019).

Membrane bioreactors (MBR) are currently considered a highly efficient technology (ZANDI *et al.*, 2019), as they combine biological treatment with membrane separation (usually micro- or ultrafiltration – MF or UF, respectively) and stand out for the high effluent quality achieved. Besides, other MBR advantages can be highlighted, like high efficiency in removing micro- and persistent organic pollutants, small industrial area requirement, and low sludge production (JUDD, 2016). MBR technology has been developed for wastewater treatment for over three decades (YAMAMOTO *et al.*, 1989) and through these years they have been widely applied, especially for municipal (HU *et al.*, 2020) and different typologies of industrial wastewater, such as textile (YURTSEVER; SAHIKAYA; ÇINAR, 2020), brewery (LU *et al.*, 2019), dairy (SONG; LIU, 2019) and oil refinery (HUANG *et al.*, 2020; MOSER *et al.*, 2019; OLIVEIRA; VIANA; AMARAL, 2020), and have been demonstrating great performances.

Oil refineries, in particular, have been increasingly applying MBR since they can really benefit from water reuse due to the large amount of water needed. However, to enable the treated water reuse at the boilers, the effluent ammonia concentration must be sufficiently low and thus understanding and controlling the factors that impact the most on its removal is essential, as well as being able to predict it (ZHANG; CHEN; JIANG, 2022). The use of artificial intelligence (AI) can notably contribute to this matter, since it is able to successfully realize feature extraction and correlation analysis of input and output data, achieving more efficient pattern classification and logical regression tasks than traditional methods (BAGHERI; AKBARI; MIRBAGHERI, 2019).

Moreover, although MBR efficiency has been consolidated by several applications and despite the great advances already achieved in its operation, membrane fouling is still a serious issue that decreases the process performance and leads to permeate flux decline, which in turn results in higher operating costs (DU *et al.*, 2020) and increases energy requirement, prejudicing the current search for more energy efficient technologies (MIRRA *et al.*, 2020). For the treatment of complex industrial wastewater, such as those from oil refineries, membrane fouling is even more challenging. Thus, an often adopted strategy in industries is the intensive use of backwashing and chemical cleaning. However, despite mitigating membrane fouling, these procedures also damage the efficiency of the process: backwashing reduces the unit productivity, since the operation is stopped and the permeate is consumed; and chemical cleaning, besides reducing the unit productivity due to the stoppage of operation, also reduces membranes lifetime. Therefore, establishing the best frequency of backwashing and chemical cleanings guided by the system performance is very interesting.

There are several studies on membrane fouling involving MBR in the literature, as computed by Meng *et al*. (2017) in their review paper. According to the authors, understanding fouling on MBR is a complex task though, since the comprehension of fouling mechanisms requires knowledge of physical, chemical and biological phenomena as well as how they interact and therefore a large set of variables that are often strongly correlated must be investigated. This way, AI stands out once more as a very promising alternative for the investigation of membrane fouling on MBR, as discussed by Bagheri *et al*. (2019), since it can be applied to extract relevant information from monitoring datasets (KAMALI *et al.*, 2020).

Machine learning (ML) is a subset of AI based on mathematical and statistical algorithms that can be used to predict the output data by finding the relationship or rule of known data. Thus, ML models can solve complex problems and achieve difficult modelling, for example nonlinear distribution and multidimensional space distribution processes (ZHONG *et al.*, 2021). There are several ML techniques that can be applied for monitoring MBR performance and the most common models include support vector machine (SVM), fuzzy logic (FL), random forest (RF), factor analysis (FA), extreme learning machine (ELM), genetic algorithm (GA), among others. Chang *et al*. (2022) tested seven ML algorithms to model both water flux and salinity of a lab-scale osmotic membrane bioreactor (OMBR) and their results have demonstrated the promise of ML for investigating these systems. Among many ML techniques though, Artificial Neural Networks (ANN), Principal Components Analysis (PCA) and Multivariate Statistical Process

Control (MSPC) have been standing out in recently published papers for their high performances at modelling complex systems (ALKMIM *et al.*, 2020; SANTOS *et al.*, 2021).

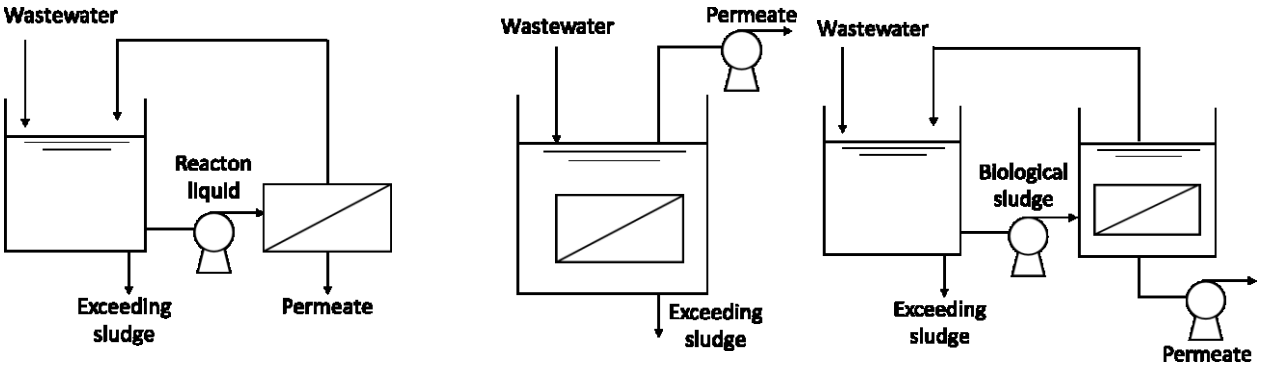
Therefore, this work aims to improve membrane fouling control and ammonia removal on MBR through the application of different AI/ML models. A pilot-scale MBR applied to the treatment of a real oil refinery wastewater was used as a case study and ANN, PCA and MSPC models were developed. The MBR was monitored during five years, so it was possible to assess its historical behavior, and the investigated variables were chosen so relevant information could be obtained on both biodegradation and membrane separation mechanisms. The models were applied to investigate the relations between different variables and to identify and diagnose operating faults related to the occurrence of membrane fouling and to low percentages of ammonia removal, in order to comprehend their main causes and to propose efficient strategies for their control. This way, the work is expected to contribute to a more efficient operation of MBR, which is currently a highly important wastewater treatment technology.

2 BACKGROUND

2.1 MBR: important wastewater treatment systems

MBR technology is based on the combination of biological treatment processes with membrane separation. Therefore, two main mechanisms are involved in the treatment: the organic matter is degraded by microorganisms (aerobically or anaerobically) and the biomass is separated from the treated effluent by membranes (usually MF or UF). MBR can be operated in two main configurations: pressurized or submerged (Figure 1). In the first case, the reaction liquid is pumped into the membrane module, normally with hollow fiber, flat sheet or tubular membrane configuration, and thus the applied pressure is the driving force for permeation. For this reason, operating pressures on this configuration are high, which leads to higher energy costs; however, maintenance and membrane cleaning are simpler, as it is an external module, and the permeate fluxes achieved are generally higher. In submerged MBR in turn, the membrane is submerged in the biomass, either inside the biological tank itself or in a separate tank with recirculation of the sludge between the tanks, and the permeate is removed by suction. This way, it is possible to operate at considerably lower pressures, which decreases energy demand and mitigates the occurrence of membrane fouling (JUDD, 2011).

Figure 1 - MBR usual configurations: (a) pressurized and (b) submerged.

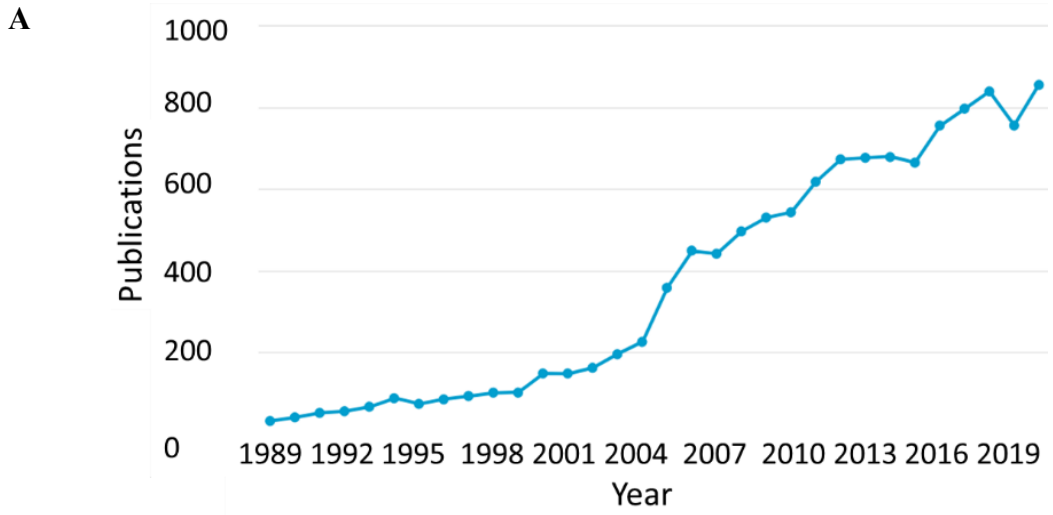


Compared with the conventional treatment process using activated sludge (CAS), MBR is able to produce better quality treated effluents and presents higher efficiency in removing micro- and persistent organic pollutants, due to the high performance of membranes in the retention of solids with low molecular weight and to the higher sludge retention time (SRT) of this system. Since in MBR the SRT is independent of the hydraulic retention time (HRT), it is possible to

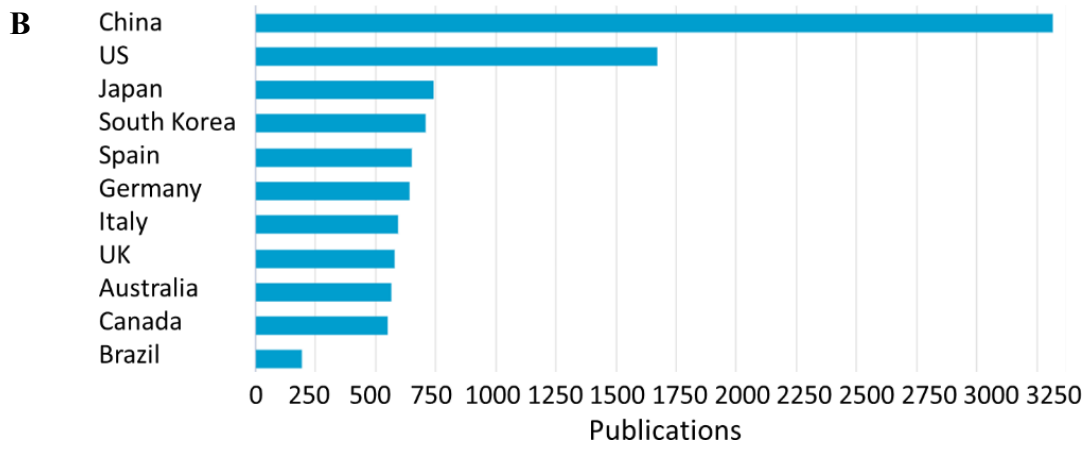
adopt a higher SRT than usual in CAS, which allows a better sludge acclimatization, hence a greater capacity to remove recalcitrant compounds. Additionally, the complete retention of the biomass ensures greater diversity of microorganisms in the biological tank, which contributes to the higher removal of persistent pollutants. Besides these advantages, MBR present smaller industrial area requirement, lower sensitivity to organic load rate (OLR) variation and lower sludge production (JUDD, 2011).

Due to its many important advantages over conventional systems, the application of MBR for the treatment of several types of wastewater has been extensively studied in the literature, like municipal wastewater (DING *et al.*, 2020; HU *et al.*, 2020), landfill leachate (AMARAL, 2016; LEBRON *et al.*, 2021) and different industrial wastewater, such as dairy (FRAGA *et al.*, 2017), textile (YURTSEVER; SAHINKAYA; ÇINAR, 2020), petrochemical (ALKMIM *et al.*, 2017; BAYAT *et al.*, 2015; KARRAY *et al.*, 2020; SAMBUSITI *et al.*, 2020), pharmaceutical (CHEN *et al.*, 2020), among others. The results reported by different works are of such interest that the publication of papers concerning MBR systems has been increasingly growing in the last three decades. Indeed, a search for publications related to MBR over the last thirty years in Scopus database identified 12,178 publications[1], with an average annual increase in the number of publications of 12% (Figure 2). It is noticeable the high influence of China, responsible for about 30% of publications (Brazil accounts for about 2%).

Figure 2 - a) Number of publications on MBR since 1989 and (b) distribution of the publications by country.



[1]Publications were searched in Scopus database using the query: TITLE-ABS-KEY ("membrane bioreactor" OR "MBR") AND PUBYEAR > 1989, on May 2021.



Source: Scopus database.

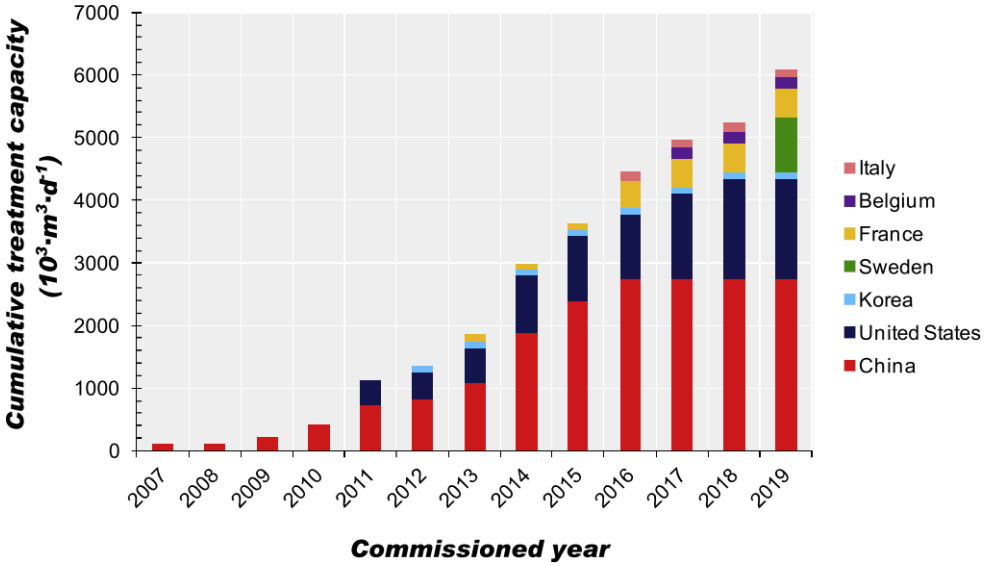
Furthermore, MBR systems have been of interest not only to the academic community, but also to the industrial sector and they have been increasingly applied for municipal and industrial wastewater treatment applications in large scale (JUDD, 2016; XIAO *et al.*, 2019). According to BCC Research, an important company on market research, MBR global market totaled $425 million in 2014 and reached $1.9 billion in 2018, becoming five times bigger in only four years. BCC research also estimates that the global market should grow to reach $3.8 billion by 2023, at a compound annual growth rate (CAGR) of 14.7% for the period of 2018-2023. According to them, MBR market is growing faster than both the larger market for wastewater treatment equipments and the market for other membrane systems (BCC RESEARCH, 2019).

Besides the increase in number, MBR plants have also been exponentially increased in scale. Whereas the largest plant installed at the turn of the millennium had a capacity of 13,000 m^3/d, by 2004 there were two plants of more than 40,000 m^3/d. Beijing Wenyu River plant, built in 2007, was the first one to reach a 100,000 m^3/d capacity (super-large scale) and since then the number of super-large MBR continued to increase all around the world to reach more than sixty super-large plants worldwide by 2019, mainly in China (JUDD, 2021; THE MBR SITE, 2021).

In Brazil, the largest MBR plant, with a 56,000 m^3/d capacity, is located in Aquapolo Ambiental S.A., in São Paulo (SP). Aquapolo is the largest wastewater reuse project in the South America, and the fifth largest of its kind in the world (KULLMANN; LAWRENCE; COSTA, 2021). The facility transforms municipal wastewater into treated water to supply the ABC Petrochemical Complex (MACHADO, 2019), and MBR technology was identified as the most cost-effective solution to upgrade the existing municipal wastewater treatment plant (WWTP) to meet São Paulo's pressing demand for industrial reuse wastewater.

Figure 3 displays the development of super-large MBR plants around the world and over the years. By 2019, it was also estimated that MBR technology provided around 5% of the world's municipal wastewater treatment capacity (JUDD, 2021).

Figure 3 - MBR treatment capacity over the years.



Source: (MENG *et al.*, 2017).

However, although MBR has been extensively studied and improved and its effectiveness has been consolidated by several applications, both academic and industrial, membrane fouling is still a severe drawback for the process efficiency, since it decreases the treated effluent quality, reduces the permeate flux, and increases the operating costs. In order to further improve MBR performance and to broader its application for municipal and industrial wastewater treatment, membrane fouling must be then effectively controlled.

2.2 Membrane fouling on MBR

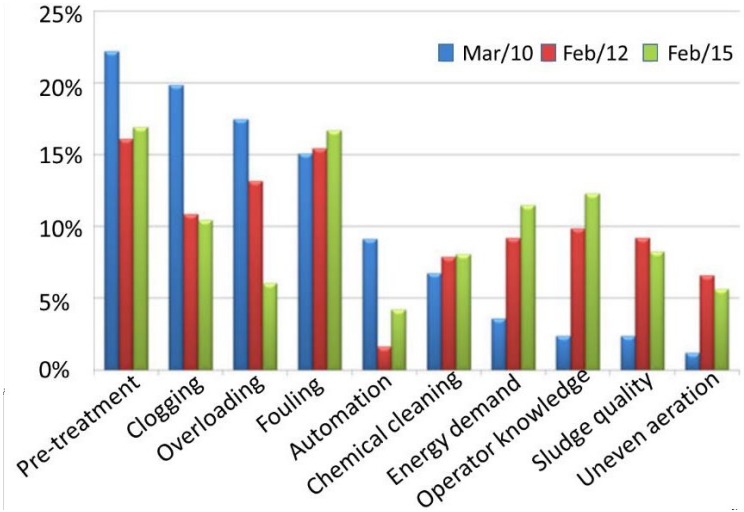
Membrane fouling is caused due to adsorption of solute molecules onto the membrane surface, obstruction of pores by suspended particles or deposit of suspended material onto the membrane surface, forming a cake (JUDD, 2011). Numerous factors related to MBR feed, membrane and biomass characteristics and operating conditions impact on membrane fouling occurrence and severity (LE-CLECH; CHEN; FANE, 2006). Several works have addressed membrane fouling characterization (like composition and morphological characteristics), upgrading of operating conditions (like SRT, HRT and permeate flow); and strategies for membrane fouling control (DU *et al.*, 2020). As discussed in these works, however, understanding membrane fouling on MBR systems is more complicated than on other membrane systems due to their complexity

and heterogeneity.

Since membrane fouling causes the restriction, occlusion or blocking of membrane pores at the surface of the membrane, it prejudices the membrane rejection capability, decreasing the treated effluent quality (JUDD, 2008). Besides, membrane fouling causes permeate flux decline, which in turn makes it necessary to proceed with chemical cleanings and replacements of the membrane more often, resulting in higher operating costs. Furthermore, it increases the energy requirement of the process, due to higher pressure demand, which impacts negatively on the process sustainability (KAMALI *et al.*, 2019).

For the treatment of complex industrial wastewater, such as those from oil refineries, membrane fouling is even more challenging. A survey was performed by The MBR Site in 2015 based on two questions: i) 'In your experience, what are the main technical issues or limitations that prevent MBR working as they should?'; and ii) 'In your opinion, how will MBR technology develop in the future?'. Respondents to the survey, 85% of whom were practitioners, identified membrane fouling as the greatest challenge to MBR operation (Figure 4). Comparing the results for the same question in previous researches, it is possible to observe that issues like clogging and overloading have become less important over the years, while fouling remainders as a major challenge. Pretreatment and energy demand, both related to membrane fouling, were also mentioned as important challenges. For the second question, membrane fouling was among the five most common keywords found in the answers, behind industrial, cost, energy and reuse (JUDD; JUDD, 2015).

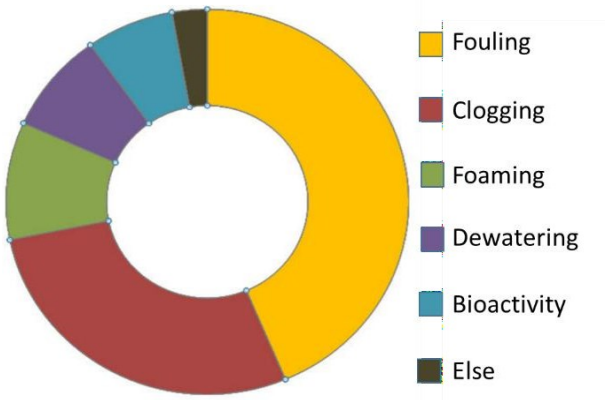
Figure 4 - MBR survey based on the question 'In your experience, what are the main technical issues or limitations that prevent MBR working as they should?'.



Source: The MBR Site, 2015.

In 2016, a similar survey was conducted, when it was asked ‘In your experience, what is the biggest challenge posed by MBR sludge?’ and once again membrane fouling was pointed out as the foremost challenge (Figure 5) (JUDD; JUDD, 2016).

Figure 5 - MBR survey based on the question ‘In your experience, what is the biggest challenge posed by MBR sludge?’.

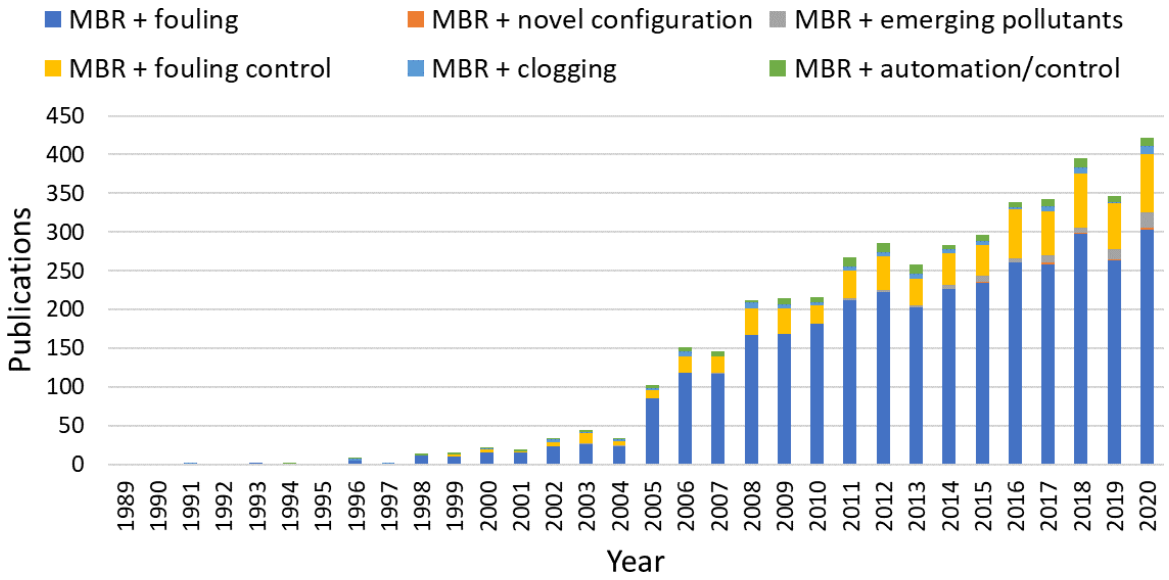


Source: The MBR Site, 2016.

Therefore, for a more efficient and economical operation of MBR systems, comprehending and controlling membrane fouling are priority demands and several recent studies involving MBR in the literature focus on membrane fouling, hoping to overcome this limitation. A search for publications related to membrane fouling on MBR over the last thirty years in Scopus database identified 3,591 publications[2], around 30% of the results found for MBR in general. Figure 6 displays the number of publications on different recent aspects of MBR (membrane fouling, fouling control, clogging, automation, novel configurations, and emerging pollutants) over the last thirty years. There is a clear predominance of interest for membrane fouling and membrane fouling control in published works in the last years. Current issues such as emerging pollutants and automation have also been increasingly studied in recent years, but they are still less approached than membrane fouling, highlighting the relevance of this matter.

[2]Publications were searched in Scopus database using the query: TITLE-ABS-KEY ("membrane bioreactor" OR “MBR” AND “fouling”) AND PUBYEAR > 1989, on May 2021.

Figure 6 - Publications on different recent aspects of MBR over the last thirty years.



Publications were searched in Scopus database using the queries: TITLE-ABS-KEY ("membrane bioreactor" OR "MBR" AND "fouling") AND PUBYEAR > 1989; TITLE-ABS-KEY ("membrane bioreactor" OR "MBR" AND "fouling control") AND PUBYEAR > 1989; TITLE-ABS-KEY ("membrane bioreactor" OR "MBR" AND "novel configuration") AND PUBYEAR > 1989; TITLE-ABS-KEY ("membrane bioreactor" OR "MBR" AND "clogging") AND PUBYEAR > 1989; TITLE-ABS-KEY ("membrane bioreactor" OR "MBR") AND ("emerging pollutants" OR "emerging contaminants") AND PUBYEAR > 1989; and TITLE-ABS-KEY ("membrane bioreactor" OR "MBR") AND ("automation" OR "process control") AND PUBYEAR > 1989, on May 2021.

2.2.1 State-of-art

In order to compile what has been discussed by the global academic community over the years regarding membrane fouling on MBR, some of the review papers published over the last decades on this matter and their principal insights are presented below.

Chang *et al.* (2002) was the first paper of this type found in our research. By the turn of the millennium, much research was already being done to investigate membrane fouling on MBR. However, well-structured theories on membrane fouling were not yet available because of the highly heterogeneous nature of the system. Nevertheless, three factors were related to the occurrence and extent of fouling in MBR: biomass characteristics, membrane characteristics, and operating conditions. Besides, some fouling control strategies, such as low-flux operation, high-shear slug flow aeration, periodical permeate backflushing, intermittent suction operation and addition of powdered activated carbon, were already being assessed, without a consensus though.

Researchers were still pursuing consensus on the exact fouling phenomena in MBR then and Le-Clech, Chen and Fane (2006) observed that a large number of publications indicated soluble

microbial products (SMP) to be one of the main parameters affecting MBR fouling. Strategies for physical and chemical cleaning were considered under-reported in the literature, which was related to the complex interactions between different fouling parameters. Understanding the nature of MBR foulants and their interactions with the membrane material was considered thus critical to provide new directions for membrane cleaning agents and protocols and for fouling mitigation strategies on MBR.

Given the difficulty in understanding foulants nature though, several works began to focus on the visualization and characterization of membrane fouling in MBR. Meng *et al.* (2010) discussed the advantages and limitations of approaches used by that time for (i) visualization of cake morphology; (ii) analysis of chemical composition; and (iii) identification of microbial community structure. The authors concluded that although a number of advances had been achieved enabling a clearer picture of fouling layer, membrane fouling behavior was still a mystery, which thereby reflected the inadequacy of existing methods for membrane fouling layer characterization, highlighting the need to develop specific techniques for MBR fouling study for its more comprehensive and reliable characterization.

Gkotsis *et al.* (2014) extensively reviewed the fundamentals of membrane fouling and the most used mitigation strategies on MBR aiming to address recent developments. Several biomasses' (e.g. mixed liquor volatile suspended solid - MLVSS, extracellular polymeric substances – EPS and SMP) and membrane's characteristics (e.g. porosity and hydrophobicity) were reviewed, but the results reported were often contradictory. Floc size, though, was pointed as a major fouling cause and chemical cleaning was pointed as a major control strategy; however, it was also related to the decrease in membrane lifetime. According to the authors, future research should focus on new cleaning methods and emerging technologies, like forward osmosis MBR. Also, they asked if further research should focus on comprehending fouling mechanisms or move to more macroscopic approaches such as mathematical modelling based on empirical relations.

Although previous review papers had presented different aspects of MBR fouling, biofouling had only been simply or partially reviewed, thus Deng *et al.* (2016) focused on this matter, addressing biofilm formation, influence factors and control approaches. They concluded that sludge properties (like sludge filterability), play a critical role in biofouling. SMP and EPS concentration and floc size were again highlighted by their great influence on membrane fouling

occurrence. Besides, adding sponge or flocculants into MBR was considered a good strategy for biofouling reduction and the development of integrated MBR with novel flocculants was suggested.

By this time, MBR was already considered a well-established technology with many full-scale units around the world treating municipal and industrial wastewater. However, membrane fouling and energy consumption were still serious challenges and, thus, developments in energy reduction, fouling control and novel configurations were being pursuit and were compiled by Krzeminski *et al.* (2017). The authors observed that these advances were concentrated on aeration, control systems, surface modifications, module configuration and novel fouling mitigation methods, as mechanical cleaning (electric field or membrane vibration). Most of the novel configurations were focused on hybrid systems. Stable flux production for long term operation and effective and/or low-energy membrane cleaning procedures were still needed, as well as cost-effective, washing chemicals resistant and antifouling membranes. Therefore, advances in material science were required.

Similarly, Bagheri and Mirbagheri (2018) discussed the many membrane fouling mitigation strategies that had been studied. According to the authors, over the last decade much effort had been made on employing novel technologies for fouling control on MBR, focusing on the improvement of the operating conditions and on the use of chemical agents to mitigate membrane fouling. However, these methods did not provide a sustainable solution for the problem. Most of the more recent studies thus had been working on using nanomaterials, cell entrapment, and biologically- and electrically-based methods to mitigate membrane fouling and the novel strategies had been showing high performances. However, the sustainable control of membrane fouling required employing more than one single approach and their application for large-scale MBR needed more research.

Finally, in more recent years, based on the new era of vast data generation and giant computer processing capacity, several works started to evaluate the study of membrane fouling and its control using data science techniques. Bagheri, Akbari and Mirbagheri (2019) reviewed the application of several AI and ML techniques for a better understanding and control of membrane fouling. Clustering analysis, ANN, FL, model trees, genetic programming, image recognition, and feature selection were found to be powerful techniques. GA and particle swarm optimization were also proven to be successfully applied for process optimization. The authors

concluded that AI and ML models can be applied to intelligently monitor and control membrane fouling, reducing the operating costs by allowing to take the best action when necessary.

From these papers it is clear that, despite the great amount of works studying membrane fouling in MBR, understanding and consequently controlling it is still a hard task. Since MBR combine biodegradation with membrane separation processes, the interaction between many physical, chemical and biological phenomena must be evaluated to understand fouling mechanisms in these systems. Consequently, a large set of strongly correlated variables must be investigated and analyzed, which makes the application of data-driven approaches an interesting alternative.

2.3 Ammonia removal on MBR

Besides membrane fouling, controlling ammonia removal so it remains on a stable condition and satisfactory level and do not present much variation is also an important challenge for the treatment of wastewaters containing ammonia by MBR systems. Generally, ammonia removal on MBR is achieved through nitrification and, since nitrifying bacteria are more sensitive to environmental factors, they tend to have lower growth rates than heterotrophic organisms. Thus, the treatment of wastewater by MBR is often limited by the removal of ammonia and it becomes important to understand the variables that influence the process.

Nitrification is the process of biological oxidation of ammonia by chemoautotrophic bacteria that use CO_2 as a carbon source. It occurs in two steps: in the first step, ammonia is converted to nitrite by ammonia-oxidizing bacteria (AOB), such as those from the genus *Nitrossomonas*, and, in the second, nitrite is converted to nitrate by nitrite-oxidizing bacteria (NOB), such as those from the genus *Nitrobacter* (von SPERLING; CHERNICHARO, 2006). It is relevant to underline that nitrification is to be understood as removal of ammonia, but not of nitrogen, since it does not result in the removal of nitrogen, but only in its conversion from ammonia to nitrate.

The growth rate of nitrifying microorganisms, especially *Nitrossomonas*, is very slow and much smaller than the growth rate of heterotrophic microorganisms responsible for the stabilization of the carbonaceous matter, reaching up to five times smaller (SHARMA; AHLERT, 1977). For this reason, in a biological treatment system where nitrification is desired, like in the treatment of ammonia-rich wastewaters, SRT should be high enough such that it enables the development of nitrifying bacteria before they are washed out from the system. This is of major importance to ensure satisfactory ammonia removal (von SPERLING; CHERNICHARO, 2006).

The environmental factors that most influence the growth rate of the nitrifying organisms and, as a consequence, the oxidation rate of ammonia are temperature, pH, dissolved oxygen (DO) and the presence of toxic or inhibiting substances. Temperature and pH are both related to the kinetics of the growth rate reaction of the nitrifying bacteria, which can be expressed in terms of Monod's relation, presented in Eqn. 1 (von SPERLING; CHERNICHARO, 2006):

$$\mu = \mu_{max} \left[\frac{NH_4^+}{K_N + NH_4^+}\right] \quad (1)$$

Where:

μ = specific growth rate of the nitrifying bacteria (d^{-1});
μ_{max} = maximum specific growth rate of the nitrifying bacteria (d^{-1})
NH_4^+ = ammonia concentration, expressed in terms of nitrogen (mgL^{-1})
K_N = half-saturation constant (mgL^{-1})

In general, higher values of temperature lead to higher growth rates and, according to Downing (1978), the nitrification rate is at its optimal and approximately constant in the pH range from 7.2 to 8.0. From Monod's relation, it is also possible to note that the lower the concentration of ammonia in the reactor, the lower the growth rate. Therefore, for higher influent concentrations of ammonia, higher ammonia removals are expected.

DO is a crucial factor to maintain the nitrifying bacteria activity and thus the United States Environmental Protection Agency (EPA) recommends that the DO concentration in the reactor should not be reduced to less than 2 mg/L (EPA, 1993). As for toxic substances, they can seriously inhibit the growth of nitrifying bacteria, mainly *Nitrossomonas*, which are more sensitive. Among several substances known to be inhibiters, there are sulphide, phenol, cyanide and oil and grease (INGLEZAKIS *et al.*, 2017; NORIEGA-HEVIA *et al.*, 2020).

Another crucial factor for nitrifying bacteria growth rate and thus ammonia removal is the ratio between carbon and nitrogen available in the system (C/N ratio). Greater availabities of organic matter, expressed mainly by the chemical oxygen demand (COD), cause higher growth of heterotrophic bacteria and result in a strong competition between them and the nitrifying ones for substrate and DO. The increase in the C/N ratio thus favors heterotrophic bacteria growth whereas limits the nitrifying ones, compromising ammonia removal (ÆSØY; ØDEGAARD; BENTZEN, 1998).

2.3.1 State-of-art

Aiming to provide an overview of the advances and limitations regarding ammonia removal on MBR, some review papers published over the last two decades on this matter and their principal insights are presented below.

At the beginning of the millennium, submerged MBR had emerged to overcome disadvantages of the pressurized MBR, like the loss of microorganism activity due to the high-speed shearing flow and the force of the pump that used to destroy the MBR biomass. However, the economics of MBR were still a barrier for its widespread application and it was thus necessary to optimize the design of these processes. Yang and Fan (2007) reviewed the back then advances on MBR design and operation and discussed aspects of properly determining HRT and SRT, sludge concentration optimization, the removal of nitrogen and phosphorus, membrane fouling control and the analysis of processing economics. The authors said that because the amount of nitrifying bacteria is small and their levels of activity low, the ammonia removal is unsatisfactory at the beginning of MBR operation. With process progress (and SRT increase), sludge concentration increases and so the amount of such bacteria. The authors also stated that DO is a limiting factor for nitrification, which is also sensitive to changes in ambient temperature, SRT, pH and OLR.

Sun *et al*. (2010), in turn, was concerned about nitrogen removal from domestic wastewaters with low C/N ratios, which was often limited because organic carbon is a limiting factor for denitrification. The authors reviewed then innovative bacterial nitrogen removal pathways such as shortcut nitrification/denitrification, simultaneous nitrification/denitrification and anaerobic ammonium oxidation (Anammox) process and concluded that MBR, among other technologies, was effective in supporting the innovative biological nitrogen removal pathways. During their investigation, the authors also concluded that AOB and NOB have different physiological characteristics and responses to environmental factors, like: AOB present slower growth rate and greater DO dependency than NOB. Besides, the papers approached by the authors demonstrate that influent COD, DO and floc size play an important role on nitrification/denitrification. Finally, the authors stated that in submerged MBR, good nitrification performance can be attained due to the high concentration of nitrifying biomass, long SRT, and the aerobic condition provided by the intensive aeration used to mitigate membrane fouling.

Due to its many advantages, MBR were rapidly becoming the technology of choice over CAS treatment systems. One important advantage was the retention of sufficient amount of nitrifying bacteria, which makes it feasible for MBR to achieve strong tolerance against the shock loads with stable and highly efficient ammonia removal. Then, several studies were being published on the nitrifier ecophysiology. Many techniques were being employed over the years aiming to understand the nitrifying community and its interaction within MBR systems. Therefore, Awolusi, Kumari and Bux (2015) focused on their review paper on the identification of optimal operational and environmental conditions for efficient nitrification in MBR. From the papers studied, the authors observed that pH plays a major role in ammonia removal on MBR than on CAS. A particular work (HE; XUE; WANG, 2009) showed that when the influent pH was acidic (approximately 4.8), the ammonia removal rate was 56%, whereas an increased removal up to 99% was observed when pH was neutral (7.2) and a decrease to 75% was noted when the pH increased to about 9.7. Therefore, efficient nitrification within the MBR falls within pH range of 7.5–8.5. Kim *et al*. (2008) found that when temperature increased from 20 to 30 ºC, ammonia oxidation proceeded from 0.25 to 1.33 gN/gVSS thereby indicating a high correlation of temperature with ammonia oxidation. Reports on activity of nitrifier at very low temperature and DO levels though indicate that nitrifier are capable of adapting to extreme conditions such as low temperature. As for SRT, the authors did not recommend to operate above 40 days since it could influence nitrification adversely. Low C/N ratio was considered to favor nitrification, whereas higher ratios support the heterotrophs. In a study of membrane-aerated biofilm reactor, nitrification efficiency of 93% was achieved at C/N ratio equals to 5 but at C/N ratio equals to 6, increased heterotrophic bacteria growth was observed with resultant inhibition of nitrifier (LIU *et al.*, 2010).

Currently, although the application of MBR for municipal and industrial wastewater secondary treatment (i.e., organic matter reduction) is well established, its evaluation for nitrogen removal still needs research. For this reason, Mao *et al*. (2020) presented a review paper that provides an overview of MBR process configurations for the removal of nitrogen based on conventional nitrogen-removal pathways (i.e., nitrification/denitrification) as well as alternative ones, such as Anammox. The authors reviewed papers that covered a wide range of system configurations, including immersed or side-stream MBR, single or multichamber processes, and the application of fixed and moving bed biofilms. Their findings indicated that operating variables play an important role in controlling nitrogen removal, especially feed composition (particularly C/N ratio), membrane characteristics, SRT and HRT. The authors also discussed that the modeling

of nitrogen-removing MBR systems can enable the optimization of system performance, and thus is a useful tool for reducing costs. The modeling structures typically include two major parts: i) biological models which focus on the bulk suspended biomass; and ii) physical models which focus on the cake layer formation on the surface of the membrane and have been applied successfully to predict the overall performance of small-scale MBR systems.

Therefore, based on the review papers, one can notice that, although several challenges about ammonia removal on MBR remain (e.g., membrane fouling, cost, and energy consumption), a number of opportunities exist, such as new reactor configurations, new microbial pathways and the development of more comprehensive models, like the data-driven ones, to promote better understanding, monitoring and control of the process. These opportunities may lead thus to the broader application of MBR processes for nitrogen removal from municipal and industrial wastewater in the future.

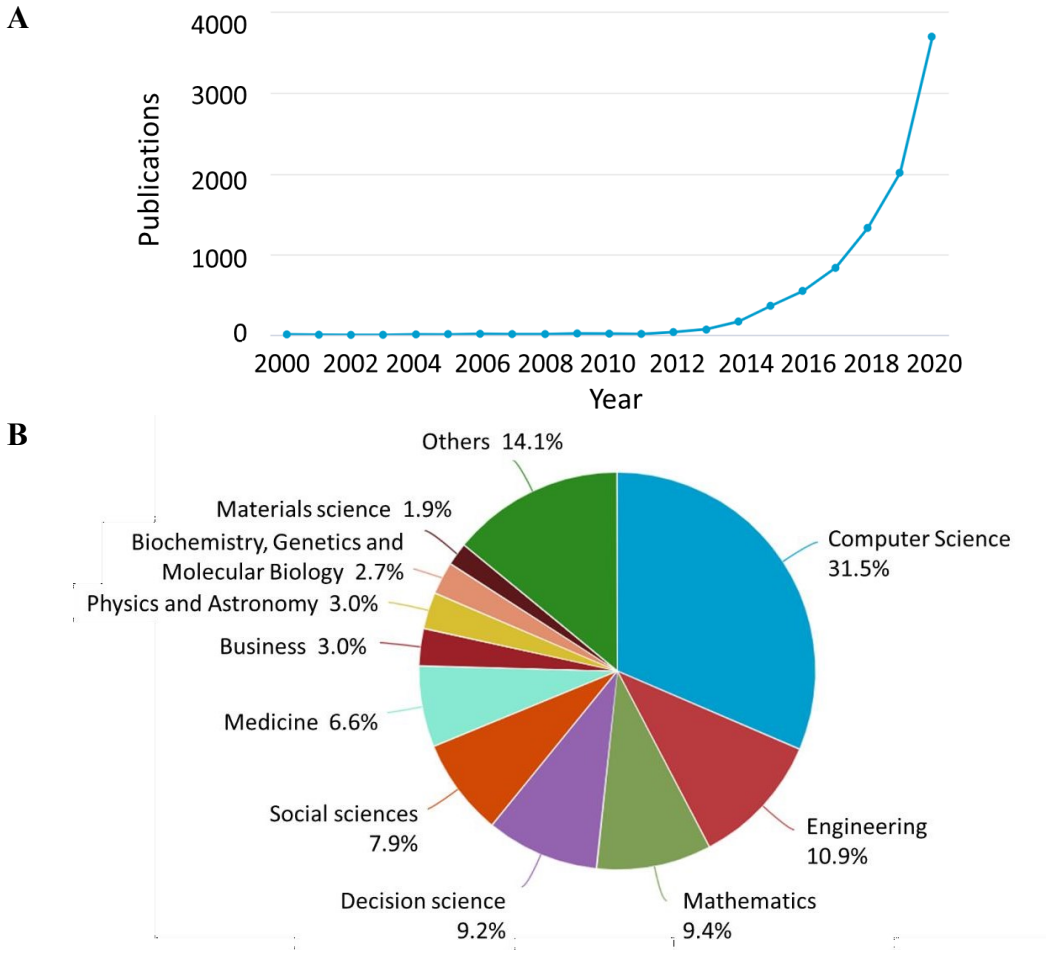
2.4 Data-driven approaches

With the advancement of the Internet of Things (continuous increase of connected devices - IoT) and the increasingly automation of processes, massive amounts of data are now available, a phenomenon known as big data (CHEN *et al.*, 2019). These data sets include trillions of words of text, billions of images, and billions of hours of speech and video, as well as vast amounts of genomic data, vehicle tracking data, clickstream data, social network data, industrial monitoring data and so on. Data volume has thus far exceeded the capacity of manual analyses and, at the same time, computers have become far more powerful and algorithms have been developed to enable broader and deeper analyses. The convergence of all these factors has given rise to data science, a set of fundamental principles, processes and methods for understanding phenomena via the automated analysis of data (RUSSEL; NORVIG, 2022).

The main goal of data science is improving decision-making and the benefits of data-driven decision-making are countless, promoting a much more assertive control of processes and therefore improving their performance (RABAN; GORDON, 2020). A key feature of data science is the extraction of relevant information and knowledge from data, which is also called data mining (VASSAKIS; PETRAKIS; KOPANAKIS, 2018). There are hundreds of data mining techniques that can be used as monitoring tools, since they allow to effectively determine the complex relations between input and output variables on a process (KAMALI *et al.*, 2020).

The growth of interest in data science recently is impressive and the exponential increase in the number of publications on this matter in Scopus database, with an average annual increase of near 80% since 2014 (Figure 7a), demonstrates how data science is relevant for the analysis of processes nowadays in several distinct areas of knowledge (Figure 7b). Works applying data science on engineering stand out, contributing to more than 10% of the published papers, and the works state that data science has been successfully applied for engineering purposes.

Figure 7 - a) Number of publications on data science since 2000; and (b) Distribution of papers by area of knowledge.



Source: Scopus database.
Publications were searched in Scopus database using the query: TITLE-ABS-KEY ("data science") AND PUBYEAR > 2000, on May, 2021. Others accounts for: energy; environmental science; Earth and planetary sciences; agricultural and biological sciences; chemistry; chemical engineering; arts and humanities; neuroscience; economics; pharmacology; health; psychology; nursing; immunology and microbiology; veterinary; and dentistry.

The field of environmental science and engineering (ESE) has also been impacted by the rapid advancement in analytical tools and monitoring technologies and massive expansion in quantity and complexity of data, which demand more advanced computational data analyses approaches beyond traditional statistical tools. Data science approaches, especially AI and ML, have shown

promise in solving complex data patterns in ESE, because of their powerful fitting abilities (ZHONG *et al.*, 2021).

The application of AI and ML modelling is widespreading so fastly that only in 2022 more than 91.000 papers have already been published on these techniques, according to a search in Scopus database[3]. Once more, engineering applications stand out, accounting for nearly 20% of the publications, behind only Computer Science field. Therefore, their application for better monitoring and controlling membrane fouling and ammonia removal on MBR wastewater treatment systems is highly interesting and can contribute to improve the technology efficiency.

2.4.1 AI fundamentals

AI is a branch of computer science that deals with the simulation of human intelligence behavior in computers or machines (BAGHERI; AKBARI; MIRBAGHERI, 2019). Any technique that enables a computational system to mimic human intelligence is a kind of AI. With the advances in computer power, large amounts of data, and theoretical understanding, AI techniques have received high attention and have become an essential part of many studies. Technically, AI is the intelligence displayed by machines in perceiving the environment by them and in taking actions that maximize the chance of successfully achieving the intended goals. This learning process based on experience compares to the natural intelligence demonstrated by humans and other animals (SHAO *et al.*, 2022).

The notion of AI can be traced back to the Middle Ages, however back them the understanding of AI was mostly related to myths. There were many legends about using witchcraft or alchemy to give consciousness to inanimate matter such as the Takwin of Jabir, the golem of Judah Loew and of Homunculus of Paracelsus and the Greek bronze man Talos – an artificially intelligent man-machine created to protect the island of Crete from invaders (RUSSEL; NORVIG, 2022).

In the 1940s and 1950s though, neurological studies showed that the brain was a neuronal neural network that emits with or without pulses, triggering discussions among a few scientists from mathematics, psychology and engineering who began to explore the possibility of an artificial

[3] Publications were searched in Scopus database using the query: TITLE-ABS-KEY ("artificial intelligence" AND "machine learning"), on September 2022.

brain. In 1943, neurologist Warren McCulloch and mathematician Walter Pitts co-authored a book that combines mathematics and algorithms, establishes neural networks and mathematical models, and simulates human thinking activities (SHAO *et al.*, 2022). Alan Turing, in 1950, published an article in which he described how to create intelligent machines and test their intelligence (TURING, 1950). The test, that would sidestep the philosophical vagueness of the question 'Can a machine think?', became known as the Turing Test and is still used as a criterion for judging whether a machine is intelligent. A computer passes the test if a human interrogator, after posing some written questions, cannot tell whether the written responses come from a person or from a computer.

Since then, AI has been increasingly improved and applied to solve the most different problems on several fields and areas of knowledge and its rapid development has dramatically changed our way of production and life. Indeed, according to Shao *et al.* (2022), AI has become the new arena of the new round of scientific and technological revolution and industrial transformation and is a noteworthy breakthrough to seize the opportunity for future development. Zhang, Zhu, and Su (2020) suggested that AI development can be conceptually divided into three stages: i) symbol AI, also called knowledge-driven approach; ii) data-driven approach, based on deep learning; and iii) the third generation AI that combines knowledge- and data-driven approaches.

A quick way to summarize the milestones in AI history is listing the Turing Award winners: Marvin Minsky, in 1969, and John McCarthy, in 1971, defined the field foundations based on representation and reasoning; Allen Newell and Herbert Simon, in 1975, developed symbolic models of problem solving and human cognition; Ed Feigenbaum and Raj Reddy, in 1994, built expert systems that encode human knowledge to solve real-world problems; Judea Pearl, in 2011, developed probabilistic reasoning techniques that deal with uncertainty in a principled manner; and finally Yoshua Bengio, Geoffrey Hinton, and Yann LeCun, in 2019, stated deep learning (multilayer neural networks) a critical part of modern computing (RUSSEL; NORVIG, 2022).

Indeed, ML and, more recently, deep learning are nowadays important subareas of AI. Their development was strongly impulsionated by the remarkable advances in computing power and the emergence of big data, as these factors have led to the development of learning algorithms specially designed to take advantage of very large datasets and to the increasing interest in AI among scientists, companies, investors, governments, the media, and the general public.

2.4.2 ML principles

There is sometimes confusion between the terms AI and ML. ML is a subfield of AI that studies the ability to improve performance based on experience. Some AI systems use ML methods to achieve competence and some do not.

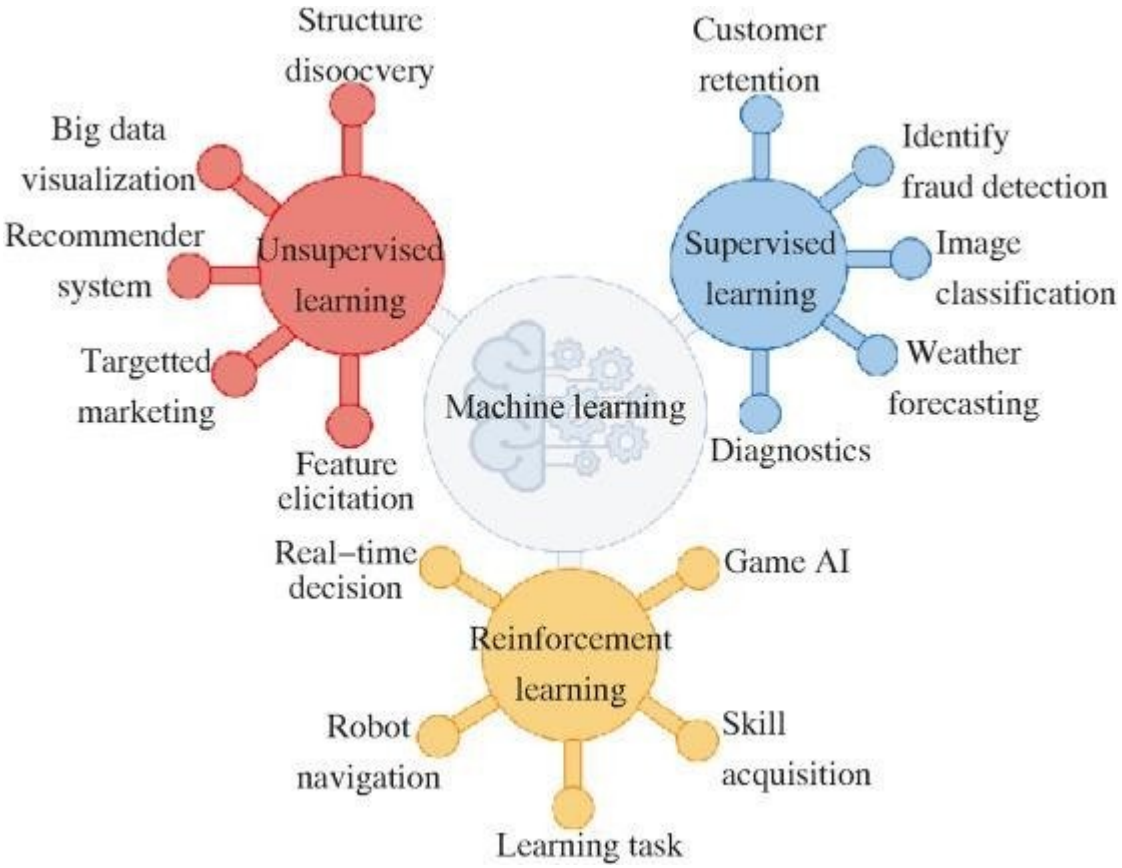
In ML, statistical techniques are used to give computers the ability to learn with data, providing a given system the ability to automatically improve from experience without being explicitly programmed to do so. The processes involved in ML require searching through data to look for patterns and adjusting program actions accordingly (ZHONG *et al.*, 2021). They are based on non-parametric algorithms that mimic the cognitive functions of the human brain, like learning and problem-solving. This way, ML models can enhance themselves by learning strategies that have worked well in the past (BAGHERI; AKBARI; MIRBAGHERI, 2019).

Using ML algorithms to build models that uncover connections and predict dynamic system, system operators can make intelligent decisions without human intervention. For example, ML enables a system to grasp the entire knowledge of social relationships between individuals and to recognize individuals' speech, face, and writing (CHEN *et al.*, 2019). For that purpose, ML models are exposed to training processes. In ML, training is the process that enables the ML framework to discover potentially relationships between input and output data and therefore teaches the model to achieve a specific goal. There are three main types of learning approaches: i) supervised, ii) unsupervised and iii) reinforcement learning (RUSSEL; NORVIG, 2022).

In supervised learning, the algorithm is provided with example inputs and their desired outputs. and learns a function that maps from input to output. For example, the inputs could be camera images, each one accompanied by an output saying "bus" or "pedestrian". The algorithm learns a function that, facing a new image, predicts the appropriate output. Supervised ML is commonly used to classification and regression problems. In unsupervised learning, no labels are given to the algorithm, leaving it on its own to find structure from the inputs. For example, when shown millions of images taken from the internet, an unsupervised ML model can identify a large cluster of similar images referring to cats. This type of ML is usually used to clustering and dimensionality reduction. In reinforcement learning, the algorithm learns from a series of reinforcements, rewards or punishments. For example, at the end of a chess game the algorithm is told that it has won (a reward) or lost (a punishment). It is up to the algorithm to decide which

of the actions prior to the reinforcement were most responsible for it and to alter its actions towards more rewards in the future. Reinforcement algorithms are used for robotics, gaming and navigation (RUSSEL; NORVIG, 2022; BAGHERI; AKBARI; MIRBAGHERI, 2019). Figure 8 displays the three types of learning in ML and common applications of each one.

Figure 8 - Three main types of learning in machine learning: supervised, unsupervised and reinforcement learning; and their most common applications.



Source: (PUGLIESE; REGONDI; MARINI, 2021).

Independently of the learning approach, the ML algorithm must thus be exposed to a dataset, from which it will map and learn relationships and patterns. However, to be a good ML model, it is desirable that the algorithm has the ability not only to fit well the input data, but more importantly, to generalize well for previously unseen data. Since testing the performance of the model in generalizing from samples already seen would be highly biased, the ideal is to divide the samples into two sets: a training set to train the model, and a test set to evaluate it. If the model has already predefined settings, this approach will be enough. If more settings are going to be tested in order to find the best model for one purpose though, it is recommended to divide the samples into three sets: a training set to train candidate models; a validation set, also known as a development set, to evaluate the candidate models and choose the best one; and a test set

to do a final unbiased evaluation of the best model performance (RUSSEL; NORVIG, 2022).

2.4.3 AI and ML models

Dues to their high performances solving real world problems, AI and ML techniques have been extensively researched and improved over the last years. There are many AI/ML techniques available nowadays and the most common include linear models, K-nearest neighbors (KNN), decision trees, like RF, SVM, FL, FA, ELM and GA. Among so many techniques though, PCA is considered to be the oldest and it is the most commonly applied (ABDI; WILLIAMS, 2010). It stands out specially for finding patterns in complex datasets and for reducing the number of dimensions without much loss of information. MSPC has also been standing out as a further development of the well stablished Statistical Process Control (SPC) methods. The application of control charts to monitor the quality of processes have found acceptance in many industrial typologies and the multivariate approach presents even higher performance rates in detecting and diagnosing operating faults in several industrial processes (HADIAN; RAHIMIFARD, 2019). Besides, currently, ANN are one of the most important techniques in ML, being pointed as the main reason for the sharp increase of the field in the last years (CHEN, MINGZHE *et al.*, 2019). They have been widely applied in many disciplines for solving many complex real-world problems, specially due to their high capacity of modelling complex relationships between inputs and outputs.

Therefore, since these techniques are consolidated in the literature and have demonstrated their ability to act as monitoring and control tools for various industrial processes, they were chosen to be applied in this work.

3 HYPOTHESES AND GOALS

3.1 Hypotheses

Based on what was observed in the literature, the following hypotheses were proposed and investigated in this work:

i. ANN and PCA modelling can reveal which variables influence the most on MBR ammonia removal capacity and membrane fouling occurrence, contributing to a better understanding of these complex mechanisms;
ii. ANN and PCA modelling can effectively predict the MBR behavior, estimating output conditions from input ones and, therefore, they can be used to forecast ammonia removal percentages and membrane permeability values;
iii. MSPC can detect and diagnose operation periods of low ammonia removal percentages and operating faults caused by membrane fouling occurrence on MBR, improving the decision-making regarding their control;
iv. The integrated assessment of ANN, PCA and MSPC models can be used to guide the definition of more efficient strategies for membrane fouling mitigation and for higher ammonia removal capacity, contributing thus for more efficient MBR operations.

3.2 Main goal

This work aims to support the definition of efficient strategies for better controlling ammonia removal and membrane fouling on MBR wastewater treatment systems from a data-driven approach.

3.3 Specific goals

i. To identify the most influential variables on membrane fouling occurrence on a pilot-scale MBR applied for the treatment of real oil refinery wastewater and to predict its membrane permeability;
ii. To identify the most influential variables on ammonia removal capacity of a pilot-scale MBR applied for the treatment of real oil refinery wastewater and to predict its values;
iii. To detect operations with low percentages of ammonia removal and operating faults caused by membrane fouling on a pilot-scale MBR treating real oil refinery wastewater

and identify their main causes;

iv. To propose efficient strategies to mitigate membrane fouling and to improve ammonia removal on MBR.

Table 1 summarizes and relates the main points of interest observed in the literature with the hypotheses and goals proposed for this work:

Table 1 - Points of interest observed in the literature, hypotheses and goals.

Literature	Hypotheses	Goals
PCA and ANN are effective in revealing relations between different variables (CHANG *et al.*, 2022; PANI, 2022);	(i) PCA and ANN can reveal which variables influence the most on ammonia removal and membrane fouling occurrence on MBR;	(i) and (ii) To identify the factors that impact the most the ammonia removal and the occurrence of membrane fouling on a pilot-scale MBR treating real oil refinery wastewater;
ANN and PCA can efficiently map a process behavior, being able to predict outputs from inputs (HONG *et al.*, 2019b; JAWAD; HAWARI; JAVAID, 2021);	(ii) PCA and ANN modelling can be used to predict the ammonia removal and the membrane permeability of MBR systems;	(i) and (ii) To forecast the ammonia removal percentages achieved and the membrane permeability values of a pilot-scale MBR treating real oil refinery wastewater;
MSPC is effective in detecting operating failures and in identifying their causes; (LE *et al.*, 2020; ZHAO *et al.*, 2020)	(iii) MSPC can detect and diagnose operating failures caused by membrane fouling on MBR and also detect and diagnose low ammonia removal capacity;	(iii) To detect low membrane permeability and ammonia removal points of operation on a pilot-scale MBR treating real oil refinery wastewater and identify their main causes;
Data analysis can lead to more well-informed decision-making, promoting more accurate control of processes and improving their performances (RABAN; GORDON, 2020).	(iv) The integrated assessment of ANN, PCA and MSPC models can guide the definition of more efficient strategies for improving ammonia removal and for membrane fouling mitigation.	(iv) To propose efficient strategies to mitigate membrane fouling and to improve ammonia removal on MBR wastewater treatment systems.

4 NOVELTY AND RELEVANCE

This work addresses AI/ML techniques that can meaningfully contribute to the technological advancement of MBR for municipal and industrial wastewater treatment. Despite the diversity of works involving membrane fouling on MBR, works that concern real industrial wastewater are still scarce, and therefore the matrix influence on membrane fouling occurrence and on ammonia removal is not accounted for. Besides, most of the works published are based on lab-scale units, investigate the relations between only a few variables and monitor the process performance for only a short period of time. In this work, both membrane fouling and ammonia removal were evaluated considering the treatment of a real industrial wastewater and a set of 14 analytic and operating variables related to both biodegrading and membrane separation of a pilot-scale MBR monitored over a long period of time was considered. Besides, although ANN, PCA and MSPC have been consolidated in several areas of knowledge, there are still few works in the literature about their application for better monitoring and controlling MBR performance. Therefore, the innovative character and the technological contribution of this work is evident.

Furthermore, ammonia removal control and stabilization is a great challenge for industrial water reuse and membrane fouling is one of the major drawbacks for MBR efficient operation, besides being the main cause for the higher operating costs still observed for MBR when compared to conventional systems. Many developed and developing countries, like the United States and China, have been expending efforts to improve advanced wastewater treatment technologies, like MBR. In Brazil, efforts are also needed in order to overcome the challenges and limitations that restrict its broader application on the country. Hence, controlling membrane fouling and allowing water reuse is critical for a more cost-effective MBR operation, contributing not only to a more widespread application of the technology, but also ensuring better performances for the already existing applications. Moreover, by contributing to a more extensive application of such an important wastewater treatment technology, this work contributes to reduce the impacts associated with inappropriate industrial wastewater disposal, which protects the environment and improves public health and welfare, highlighting the relevance of the work.

II. ARTIFICIAL NEURAL NETWORKS, PRINCIPAL COMPONENTS ANALYSIS AND MULTIVARIATE STATISTICAL PROCESS CONTROL

THEORETICAL FOUNDATION AND APPLICATIONS

1 PCA

According to Abdi and Williams (2010), PCA is probably the most popular multivariate statistical technique (CAMACHO *et al.*, 2016; JOLLIFE; CADIMA, 2016) and it is also likely to be the oldest one. Indeed, its origin can be traced back to Pearson (1901), but its modern formulation was formalized by Hotelling, who coined the term principal component (HOTELLING, 1933). This statistical technique allows us to summarize and to visualize the relevant information present in a complex dataset containing observations described by multiple correlated variables. PCA also represents the pattern of similarity between observations and variables by displaying them as points in maps, expressing the data in such a way that highlight their similarities and differences (JOLLIFFE, 2002). Since finding patterns can be hard in data of high dimension, PCA is a useful statistical technique that has found application in almost all scientific disciplines (ABDI; WILLIAMS, 2010).

1.1 Theoretical foundation

1.1.1 Notation

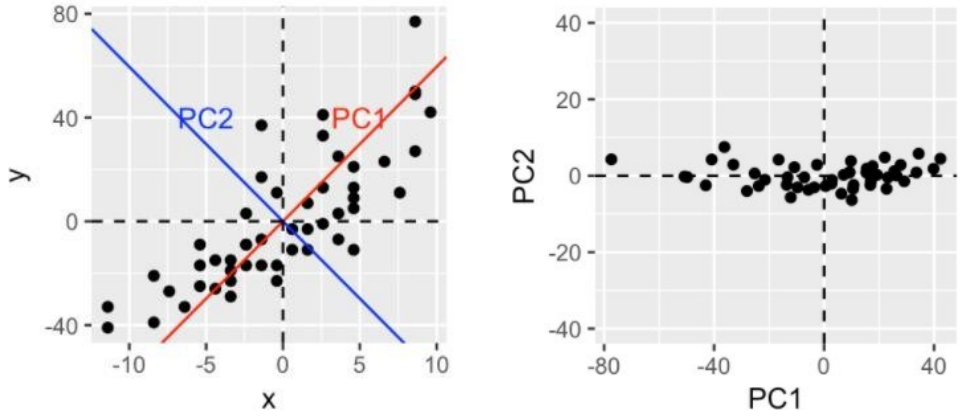
Matrices are denoted in upper case bold, vectors are denoted in lower case bold, and elements are denoted in lower case italic. Matrices, vectors, and elements from the same matrix all use the same letter (e.g., **A**, **a**, a). Index i represents the observations ($i = 1, \ldots, I$), index j represents the variables ($j = 1, \ldots, J$) and index k represents the components ($k = 1, \ldots, K$). The transpose operation is denoted by the superscript $^{\mathbf{T}}$ and the identity matrix is denoted by **I**.

1.1.2 Finding the PC

PCA is a multivariate procedure aimed at reducing the dimensionality of multivariate data while accounting for as much of the information in the original dataset as possible. The information in a given dataset corresponds to the total variation it contains. Therefore, PCA aims to identify the directions along which the data variation is maximal (EVERITT; HOTHORN, 2008). Figure 9 displays an example: in the first plot (Figure 9a), the data are represented in the X-Y coordinate system. The dimensionality reduction is achieved by identifying the PC, i.e., the directions in which the data mostly varies. Thus, in the second plot (Figure 9b), PC1 axis is the first principal direction along which the samples show the largest variation and PC2 axis, which

is orthogonal to PC1 axis, is the second most important direction. This way, PCA can be geometrically explained as the rotation of the original axes into the directions of most variation (JOHNSON; WICHERN, 2014).

Figure 9 - Principal components (PC) geometric representation.



A **B**

Source: (KASSAMBARA, 2017).

Algebraically, PCA seeks to transform the original variables into a new set of variables that are linear combinations of the original ones and uncorrelated to each other (EVERITT; HOTHOR, 2008). However, before these new variables can be determined, the data must be preprocessed, since PCA performed on raw data is often not very meaningful. Although there are many types of preprocessing methods available, mean-centering and scaling are the two most common preprocessing methods and they are often required (BRO; SMILDE, 2014).

Mean-centering and scaling are recommended procedures when the variables are measured in different scales, which can severely affect PCA outputs, since the variables are not comparable. This way, the data should be preprocessed to have i) standard deviation (sd) one and ii) mean zero, which is achieved by subtracting from every variable the corresponding mean and by dividing them by the corresponding sd, as shown in Eqn. 2. This procedure reduce the large difference between the orders of magnitude of the different variables and thus make them more comparable, giving equal importance to each one in the multivariate projection models (GONZÁLEZ-CAMEJO *et al.*, 2020).

$$x_{i,j} = \frac{y_{i,j} - \mu_j}{\sigma_j} \tag{2}$$

where:

$x_{i,j}$ = centered and scaled i^{th} value of variable j

$y_{i,j}$ = original i^{th} value of variable j

μ_j = mean of variable j

σ_j = sd of variable j

After being mean-centered and scaled thus, the data is collected in a matrix that comprises I observations described by J variables and it is represented by the $I \times J$ matrix $\mathbf{X}$, whose generic element is $x_{i,j}$ ($i = 1, \ldots, I$ and $j = 1, \ldots, J$). The individual variables of $\mathbf{X}$ are denoted by $\mathbf{x_j}$ ($j = 1, \ldots, J$) and are all vectors in the I-dimensional space. A linear combination of those $\mathbf{x}$ variables can be written as shown in Eqn. 3:

$$\mathbf{t_j} = w_1 * \mathbf{x_1} + \cdots + w_j * \mathbf{x_j} \tag{3}$$

where:

$\mathbf{t_j}$ = new vectors in the same space as $\mathbf{x}$

w_j = weight of each variable j

Since the aim is to lose the minimum of the information contained in matrix $\mathbf{X}$, it is needed to find the w_j weights that will get the new variables $\mathbf{t_j}$ that best explains the total variation of the original dataset. The variation in $\mathbf{t_j}$ can be measured by their variance, var($\mathbf{t_j}$). Thus, the problem can be translated as maximizing this variance by choosing optimal weight vectors $\mathbf{w_j}$ with elements w_j ($j = 1, \ldots, J$). As the matrix $\mathbf{X}$ is mean-centered and scaled to unit variance, this is actually a standard problem in linear algebra and the optimal $\mathbf{w}$ are the eigenvectors of the covariance matrix ($\boldsymbol{\varepsilon}$) of matrix $\mathbf{X}$ (for detailed mathematical proof, refer to ABDI; WILLIAMS, 2010; BRO; SMILDE, 2014).

Eigenvectors and eigenvalues are vectors and numbers associated to square matrices. Together they provide the eigen-decomposition of a matrix, which analyzes the structure of this matrix. There are several ways to define eigenvectors and eigenvalues, but the most common approach defines an eigenvector of the matrix $\mathbf{A}$ as a vector $\mathbf{u}$ that satisfies Eqn. 4:

$$\mathbf{Au} = \lambda\mathbf{u} \tag{4}$$

where:

λ = eigenvalue associated to the eigenvector $\mathbf{u}$

This means that a vector **u** is an eigenvector of a matrix **A** if the length of the vector (but not its direction) is changed when it is multiplied by **A**. Traditionally, the set of eigenvectors of **A** are stored in a matrix **U**. Each column of **U** is an eigenvector of **A**. The eigenvalues are stored in a diagonal matrix **Λ**, where the diagonal elements give the eigenvalues. Therefore, Eqn. 4 can be rewritten as Eqn. 5 or Eqn. 6:

$$\mathbf{AU} = \mathbf{\Lambda U} \tag{5}$$

or:

$$\mathbf{A} = \mathbf{U\Lambda U}^{-1} \tag{6}$$

As eigenvectors corresponding to different eigenvalues are orthogonal, the matrix **U** is also orthogonal and therefore $\mathbf{U^{-1}} = \mathbf{U^T}$. Then:

$$\mathbf{A} = \mathbf{U\Lambda U^T} \tag{7}$$

Positive semi-definite matrices are used very often in statistics. A matrix is classified as positive semi-definite when it can be obtained as the product of a matrix by its transpose, which implies that a positive semi-definite matrix is always symmetric. In particular, covariance matrices are always positive semi-definite matrices, which is convenient because the eigen-decomposition of these matrices always exists (ABDI; WILLIAMS, 2010). Thus, when the covariance matrix **ε** is eigen-decomposed, it can be written as Eqn. 8:

$$\boldsymbol{\varepsilon} = \mathbf{U\Lambda U^T} \tag{8}$$

The trace of the covariance matrix, i.e. the sum of the diagonal elements of a square matrix, can thus be calculated as Eqn. 9:

$$\text{trace}(\boldsymbol{\varepsilon}) = \text{trace}(\mathbf{U\Lambda U^T}) = \text{trace}(\mathbf{\Lambda UU^T}) = \text{trace}(\mathbf{\Lambda I}) = \text{trace}(\mathbf{\Lambda}) = \sum_{j=1}^{J} \lambda_j \tag{9}$$

As the covariance matrix has on its main diagonal the variances of each variable *j*, it can also be written as Eqn. 10:

$$\text{trace}(\boldsymbol{\varepsilon}) = \sum_{j=1}^{J} \sigma_j^2 \tag{10}$$

And so:

$$\sum_{j=1}^{J} \sigma_j^2 = \sum_{j=1}^{J} \lambda_j \tag{11}$$

This means that the total variation contained in the original variables is equal to the sum of the eigenvalues of the covariance matrix and that is why the best weight vectors **w** are the eigenvectors of the covariance matrix. Therefore, PCA transforms the dataset into new variables **t** using the eigenvectors of the covariance matrix as weights of the linear combinations. In matrix notation, this becomes Eqn. 12:

$$\mathbf{T}_{[i,k]} = \mathbf{X}_{[i,j]} \mathbf{U}_{[j,k]} \tag{12}$$

In order to the first PC explain most of the data variation, the eigenvectors are ranked according to their corresponding eigenvalues, in descending order. The importance of a PC is thus reflected by its eigenvalue (λ_k) and the percentage of total variation explained by each k component is calculated as Eqn. 13 (JOLLIFFE, 2002):

$$\%\text{explained}_k = \frac{\lambda_k}{\sum_{k=1}^{J} \lambda_k} \tag{13}$$

Therefore, it is possible to replace the original J variables with only the first K ($K < J$) components without losing much information. This reduction on data dimensionality provides several benefits: the influence of noise is minimized, the interpretation and visualization of the data is greatly improved and further modelling with the data is favored (BRO; SMILDE, 2014).

The values of the new variables computed by PCA for the observations are called scores and they can be geometrically interpreted as the projections of the original observations onto the PC. The correlation between a PC and an original variable is in turn called loading and estimates the information they share. The variables can also be plotted in the component space using their loadings as coordinates, however their representation differs from the scores plot: whereas observations are represented by their projections, variables are represented by their correlations. When the data are perfectly represented by only two components, the sum of the squared loadings is equal to one, and therefore the loadings will be positioned on a circle that is called the circle of correlations. When more than two components are needed to represent the data

perfectly, the variables will be positioned inside the circle of correlations. The closer a variable is to the circle of correlations, the better this variable can be reconstructed from that two PC (and the more important it is to interpret these components); the closer to the center of the circle of correlations a variable is, the less important it is for that two PC (ABDI; WILLIAMS, 2010). It is yet possible to visualize and interpret scores and loadings simultaneously through a graph name biplot. By plotting observations and variables together, it is possible to relate the behavior of observations to specific variables and, therefore, one can explain e.g. why a certain grouping is observed or why the behavior of the observations change throughout time (BRO; SMILDE, 2014).

1.1.3 PCA as a model

Another way of assessing the summarizing capability of the new variables **t** is evaluating how representative **t** is in terms of replacing **X**. This can be done by projecting the columns of **X** on **t** and calculating the residuals of that projection. So, we can regress all variables of **X** according to Eqn. 14, which derives from Eqn. 12:

$$\underset{[i,j]}{\hat{\mathbf{X}}} = \underset{[i,k]}{\mathbf{T}} \; \underset{[k,j]}{\mathbf{U}^{\mathrm{T}}} \tag{14}$$

where:

$\hat{\mathbf{X}}$ = matrix containing the estimated values of **X**

If all components were kept (i.e. $K = J$), $\hat{\mathbf{X}}$ would be equal to **X**. However, when the last PC are eliminated from the model, $\hat{\mathbf{X}}$ deviates from **X** and the difference between them is called residuals. Mathematically:

$$\underset{[i,j]}{\mathbf{X}} = \underset{[i,j]}{\hat{\mathbf{X}}} + \underset{[i,j]}{\mathbf{E}} = \underset{[i,k]}{\mathbf{T}} \; \underset{[k,j]}{\mathbf{U}^{\mathrm{T}}} + \underset{[i,j]}{\mathbf{E}} \tag{15}$$

where:

E = matrix of residuals, which elements are calculated according to Eqn. 16:

$$e_{i,j} = x_{i,j} - \hat{x}_{i,j} \tag{16}$$

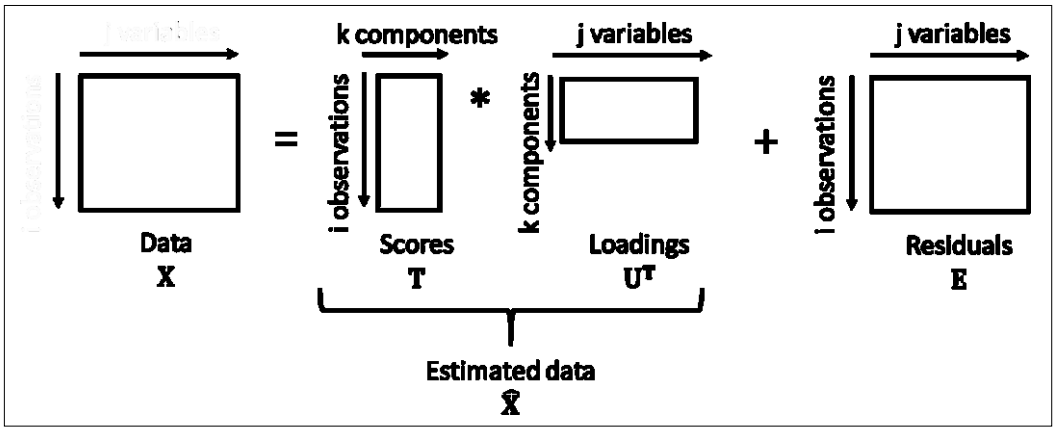
where:

$e_{i,j}$ = i^{th} residual value of variable j

$\hat{x}_{i,j}$ = estimated i^{th} value of variable j

Eqn. 15 denotes that PCA analysis can also be applied as a modelling activity since the product $\mathbf{TU^T}$ serves as a model of **X**, highlighting this important PCA feature. From the same equation, it is possible to derive all four parts of PCA models: the data (**X**), the scores (**T**), the loadings (**T**) and the residuals (**E**) (BRO; SMILDE, 2014). Figure 10 displays a PCA model structure. It is noticeable that the model approximation of the data ($\mathbf{TU^T} = \hat{\mathbf{X}}$) and the residuals have the same structure as the data.

Figure 10 - Structure of a PCA model.



The overall quality of the PCA model after K components is evaluated as the similarity between $\hat{\mathbf{X}}$ and **X**. Several coefficients can be used for this evaluation, but the most popular one is the degree of fit or coefficient of determination (R^2), calculated as shown in Eqn. 17 (ERIKSSON *et al.*, 2013):

$$R^2 = 1 - \frac{RESS}{SSX} \tag{17}$$

where:

RESS = residual sum of squares (Eqn. 18)

SSX = total variation of the mean-centered and scaled data matrix (**X**) (Eqn. 19).

$$RESS = \sum_{i=1}^{I}\sum_{j=1}^{J} e_{i,j}^2 \tag{18}$$

The smaller the value of RESS, the higher the value of R^2 and the better the PCA model.

$$SSX = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j}^2 \tag{19}$$

The PCA model can then be applied for data prediction. For this purpose, a dataset is used to build the model; and a second dataset (usually new observations) is left out to be predicted by the PCA model, as described above. The quality of the predictive model is commonly evaluated from its predictive ability (Q^2), calculated according to Eqn. 20 (ERIKSSON *et al.*, 2013):

$$Q^2 = 1 - \frac{PRESS}{SSX} \tag{20}$$

where:

PRESS = predicted residual sum of squares. It is calculated the same way as RESS, but for the data which was predicted by the model and not used for its development.

The same way, the smaller the value of PRESS, the higher the value of Q^2 and the better the predictive PCA model.

1.1.4 Choosing the number of PC to keep

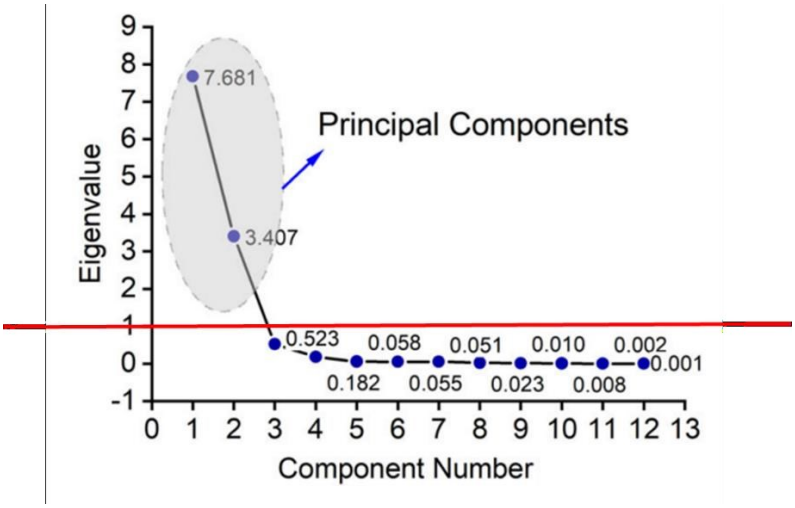
In exploratory studies, where the aim is generally just to have an overall look of the data, it is not urgent to fix the number of components very accurately. Often, the interest is only in looking at the main variation and per definition the first few PC provide information on that. For more complex purposes though, it is important to establish the number of PC to keep more precisely, since the residuals will change depending on how many PC will be used (BRO; SMILDE, 2014). Unfortunately, there is no well-accepted global and objective way to decide how many PC are enough. This will depend on the specific field of application and the specific dataset. Nevertheless, several methods, usually applied in a combined manner, have been proposed (JACKSON, 1993). The three most common methods are the scree test (CATTELL, 1966), the Kaiser criterion (KAISER, 1960) and the fraction of variation explained (JOLLIFFE, 2002).

According to Kaiser criterion, only components with eigenvalues greater than the unit should be kept, since these axes summarize more information than any single original variable, i.e. PC that have $\lambda > 1$ account for more variance than accounted by one of the original variables. This

happens because when data is mean-centered and scaled the eigenvalues sum is equal to the total number of original variables (and hence the total number of eigenvalues). This means that the mean of the eigenvalues is equal to one and therefore eigenvalues greater than one are greater than the mean (BRO; SMILDE, 2014; JOLLIFFE, 2002).

The scree plot is the plot of the eigenvalues ordered from the largest to the smallest. The number of components to keep is determined at the point beyond which the remaining eigenvalues are all relatively small and of comparable size. Visually, one must observe if there is a point in the scree plot (often called an 'elbow') such that the slope of the plot goes from 'steep' to ''flat'' and to keep only the PC which are before the elbow (ABDI; WILLIAMS, 2010; EVERITT; HOTHORN, 2008). Figure 11 displays an example of a scree test and demonstrates both Kaiser and the scree test criteria to keep only the first two PC.

Figure 11 - Scree plot example demonstrating the scree test and Kaiser criteria for deciding to keep the first two principal components.



Source: Adapted from Ajjur; Al-Ghamdi (2021).

The fraction of variation explained method consists in including all components up to some predetermined fraction of total variation explained. For example, if 70% of the total variance explained is enough for a certain application, then the number of PC necessary to achieve that fraction should be retained. It is important to note though that this is a heavily application and field of knowledge dependent method (JACKSON, 1993) and therefore should be used wisely. In general, fractions of variation explained greater than 60% are considered enough (JACQUIN *et al.*, 2018; YU *et al.*, 2017).

1.2 PCA for investigating membrane fouling on MBR

PCA has been successfully applied as a data-driven approach for investigating membrane fouling on MBR. Although there are still few works on this subject in the literature, the existing ones demonstrate that PCA is effective in revealing data trends and patterns, allowing the identification of the most influential variables on the occurrence of membrane fouling and supporting the decision-making concerning fouling mitigation.

Maere *et al.* (2012), for example, applied PCA to evaluate the membrane state of a lab-scale MBR treating a synthetic municipal wastewater based solely in transmembrane pressure (TMP) data monitored during five months. The number of PC to be retained was based on both the Kaiser criterion and the cumulative explained variation. The first two PC were kept, corresponding respectively to 96.1 and 3.4% of the data variation and making a combined explained variation of 99.5%. Two major trends were observed in the data and they were linked to reversible and irreversible fouling, which were represented by PC1 and PC2, respectively. According to the authors, TMP data appear to contain the necessary information for membrane management and PCA was able to efficiently extract the information in an automated way and therefore it could be used for membrane fouling control purposes.

Choi *et al.* (2013) also applied PCA aiming to evaluate membrane fouling on MBR. The authors monitored a lab-scale MBR treating a synthetic wastewater for one year in order to investigate the correlations between effluent organic matter (EfOM) parameters and membrane fouling. The variables considered were i) operating conditions: cake resistance, SRT, temperature, TMP, pH and total resistance; ii) biomass: viscosity, SMP, EPS, COD, mixed liquor suspended solids (MLSS), and MLVSS. The first PC embodied the characteristics of EfOM and explained 59% of the data variation, expressing its importance for membrane fouling. The second PC explained 8% of the data variation and therefore the two PC kept accounted for 67% of the data variation. PCA was a useful tool to evaluate the correlations among dissolved organic matter (DOM) and membrane fouling, and also to determine the group of parameters that most influence the MBR performance.

Yu *et al.* (2017) in turn assessed membrane fouling on a MBR applied for the treatment of an industrial wastewater, obtained from an antibiotic industry. The lab-scale MBR were monitored for six months. Different categories of variables, named i) feed characteristics: total organic

carbon (TOC), and ammonia; ii) biomass: viscosity, capillary suction time (CST), MLVSS, EPS, and SMP; (iii) operating conditions: OLR and pressure were evaluated. The first two PC were extracted based on the scree plot and the Kaiser criterion. They explained 53.2 and 9.6%, respectively, of the data variation, corresponding to 62.8% of total variation explained. Results from PCA highlighted both proteins and carbohydrates in EPS as the primary foulants. Membrane fouling associated with the first PC was positively related to EPS whereas PC2 was primarily related to influent proteins. Other important categories affecting membrane fouling were ranked as biomass characteristics, operating conditions, and feed characteristics.

Jacquin *et al.* (2018) also worked with real wastewater, but with municipal wastewater rather than industrial. Their work also stands out because a full-scale MBR monitored for 16 months was assessed. PCA was applied to establish the link between DOM, operating conditions, active biomass concentration and membrane fouling. The variables considered were also divided into categories: i) operating conditions: SRT, temperature, concentration factor (CF = SRT/HRT), MLVSS, nitrogen load and organic load; ii) biomass: different microbial species. The first two PC were kept based on the fraction of total variation explained, equal to 66.2%. PCA results showed that operating parameters did not have the same impact on active biomass populations. SRT and temperature were identified as the variables with the most influence on active biomass concentrations and consequently on its associated MSP production. Therefore, these parameters play an important role on the occurrence of membrane fouling and were considered as major foulants. Besides, MLVSS did not correlated to heterotrophic bacteria concentration, so the authors concluded that it was not appropriate to quantify active biomass.

Hong *et al.* (2019) also used PCA analysis to assess the behavior of the microbial community in a MBR, evaluating the concentration of different species. The authors, however, assessed a lab-scale MBR treating a synthetic municipal wastewater. The MBR was operating under two different modes: constant TMP and constant permeate flux and it was monitored during three months for each filtration mode. The authors decided to keep three PC to explain 77% of the total variation and meet the Kaiser criterion. The microbial species that characterized each operation mode were depicted well by PCA. PC1 distinguished operation modes of constant TMP and constant flux, PC2 distinguished early stage in constant TMP mode from all others; and PC3 distinguished cake sludge from bulk sludge. The authors concluded that membrane fouling was more intense in constant TMP mode due to the higher SMP and EPS release and to the higher abundance of biofilm-forming bacterial group.

Also aiming to understand the microbial community behavior in MBR systems, Rodriguez-Sanchez *et al.* (2019) investigated the performance, biomass kinetics and microbial community structure in biofouling and suspended biomass of a hybrid moving bed biofilm reactor-membrane reactor (MBBR-MBR) system subjected to four different scenarios of salinity. For this purpose, the authors worked with a lab-scale MBBR-MBR applied for the treatment of a salinity-amended municipal wastewater. The period of monitoring was not informed, as well as the percentage of total variation. The PCA model showed higher community similarity between biofouling and suspended biomass under variable salinity conditions than for constant salinity, which could be attributed to low adaptability of bacteria to variable salinity regimes. Also, differences were observed in the relative abundance of dominant bacteria between biofouling and suspended biomass at all salinity scenarios, indicating that some groups of species are more influential on membrane fouling occurrence.

The works presented demonstrate the great potential of PCA to be used as a tool for monitoring membrane fouling on MBR, with fractions of total variation explained above 60% being usually considered satisfactory to well represent the system. Despite the great results presented in these papers though, they also reflect a gap in the literature about assessing membrane fouling on MBR through PCA models, since most of the existing works focus on few variables, during a short period of monitoring and/or using synthetic wastewater. The investigation of membrane fouling in the treatment of real industrial wastewater is critical since the matrix strongly impacts on its occurrence and severity and thus must be accounted for. Besides, evaluating the MBR operation over a long period of time is also really important, since the membrane has its characteristics changed throughout its lifetime.

1.3 PCA for monitoring MBR general performance

PCA has been also applied to monitor other aspects of MBR operation and performance, beyond membrane fouling. Following are presented some of the most recent ones.

Qin *et al*. (2021) investigated the influence of distinct HRT and OLR on fungal dynamics during synthetic food waste anaerobic digestion in two immersed MBR. The first one was fed with a solution of COD concentration of 4 $gL^{-1}d^{-1}$ for 34 days, when the feed solution COD concentration was increased to 8 $gL^{-1}d^{-1}$. The second MBR was first fed with a 6 $gL^{-1}d^{-1}$ solution and thereafter elevated to 10 $gL^{-1}d^{-1}$. Samples were collected from both MBR after 1 (T1), 15

(T2) and 34 days (T3) to analyze fungal community by using 18s rDNA. PCA indicated that fungal diversity was varied among all three phases (T1, T2, and T3) for both MBR and the results showed that different OLR and HRT values have significantly influenced the fungal community.

Miwa *et al*. (2021) also used PCA to assess the microbial community of two lab-scale MBR treating municipal wastewater, but the authors focused on biofilm formation and, consequently, on membrane fouling. PCA modelling based on 16s rRNA showed that the biofilm microbial community changed significantly from middle stage to mature biofilm when compared with that of activated sludge. Besides, the model indicated the abundance of specific bacteria, such as unclassified *Neisseriaceae*, increased in middle-stage biofilm and the diversity indexes of middle-stage biofilm were lower than those of mature biofilm and activated sludge. These results suggested that the presence of specific bacteria with colonization ability played a crucial role in biofilm formation and that strategies to reduce membrane fouling on MBR should be sought during early- and middle-stage biofilm formation.

Viet and Jang (2022), in turn, investigated the feasibility of applying a novel methodology for constructing a fertilizer draw solution (DS) used in an OMBR for simultaneous wastewater treatment and sustainable fertigation. The lab-scale OMBR was fed with synthetic wastewater. The results indicated that the system performance, expressed by water flux, reverse salt flux and contaminant removal, varied critically under different fertilizers. Besides, NH_4NO_3 and $NH_4H_2PO_4$fertilizers caused the highest and lowest membrane fouling resistances, respectively. PCA model kept two PC and accounted for 70% of the total variation. The model showed that fouling resistance played a pivotal role in the total variation of the system.

Gutiérrez *et al*. (2022) applied PCA to compare and discuss the different results published in the scientific papers approached in their previous review article (GUTIÉRREZ *et al.*, 2021), regarding enhanced micropollutant removal on MBR coupled with powdered activated carbon (PAC). All data included in their dataset refers to lab-scale plants, with the exception of nine observations that refer to a pilot-scale unit. All the experimental MBR were fed with synthetic wastewater. PCA model was applied then to identify the most influential factors from a set of operational parameters (PAC dosage and retention time and SRT) and compounds physico-chemical properties (octanol-water distribution coefficient (D_{ow}), molecular weight and charge). The PCA model reduced the dimensionality of the dataset to four PC that explained 87% of the

total variation. Its results demonstrated that, based on the collected dataset, micropollutant charge and $LogD_{ow}$ seem to play the most important role in the removal mechanisms occurring in MBR coupled with PAC.

Similarly to monitoring membrane fouling through PCA, the works discussed demonstrate the great potential of PCA to be used as a tool for monitoring MBR performance regarding distinct aspects of its operation and at the same time reflect the gap in the literature, since most of the works focus on lab-scale MBR treating synthetic wastewater. Besides, considering bigger MBR scales can also contribute to better understand and control real processes, since the relations on lab-scale can differ from the pilot and real-scale ones.

2 MSPC

MSPC has also been standing out in the literature as a multivariate statistical technique that can be used as a monitoring tool for several industrial processes. This technique applies multivariate control charts to detect any unusual events, as well as to identify their main causes (DAS, 2019). MSPC methods reduce the information contained within all process variables down to a few composite metrics through the application of statistical modeling. These composite metrics can then be easily monitored in order to benchmark process performance and highlight potential problems (BERSIMIS; PANARETOS; PSARAKIS, 2009). The basis of this method is to build an empirical model from a set of measurements obtained under normal operating conditions (NOC) and to calculate statistical confidence limits from this model. If the process is operating according to the expected, the new observations projected onto the model should thus be within the confidence limits. If they are not, some atypical event has caused the process to deviate from its normal behavior (WESTERHUIS; GURDEN; SMILDE, 2000).

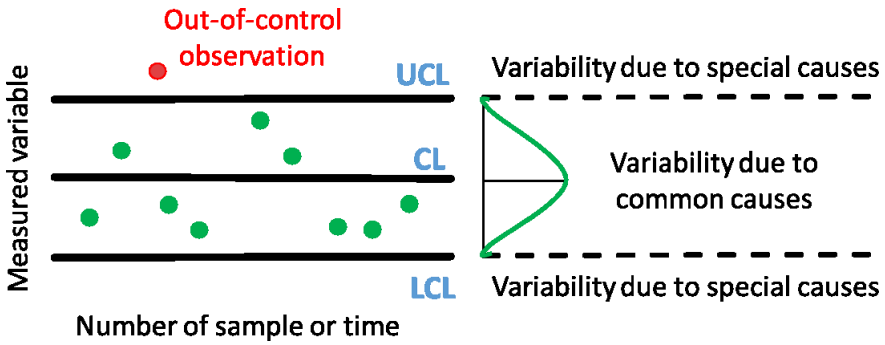
2.1 Theoretical foundation

2.1.1 SPC versus MSPC

SPC has been widely applied for quality control over the last two decades and the most common process control technique is control charting. Control charts can be applied to monitor process stability, to detect any assignable variations, and/or to forecast process movements (FERRER, 2007). They are graphical representations of the process variability and its natural and unnatural patterns. Common causes are related to the natural variability that always exists in any process and cannot be avoided, whereas special causes are not inherent to the process and, therefore, can be identified and eliminated (MONTGOMERY, 2016).

In order to achieve this goal thus, control charts display a value of the quality characteristic that has been measured versus the sample number or time (Figure 12). The central line represents the average value of the quality characteristic and the control limit lines are set so that when the process is statistically stable, nearly all the points in the control chart fall between them (NOSKIEVIČOVÁ, 2013). Thereby, the upper control limit (UCL) and the lower control limit (LCL) separate common and special causes of variation and, therefore, a process is said to be out of statistical control when an observation falls outside the control limits. In this case, it is assumed that an assignable cause of the abnormal process variability is present.

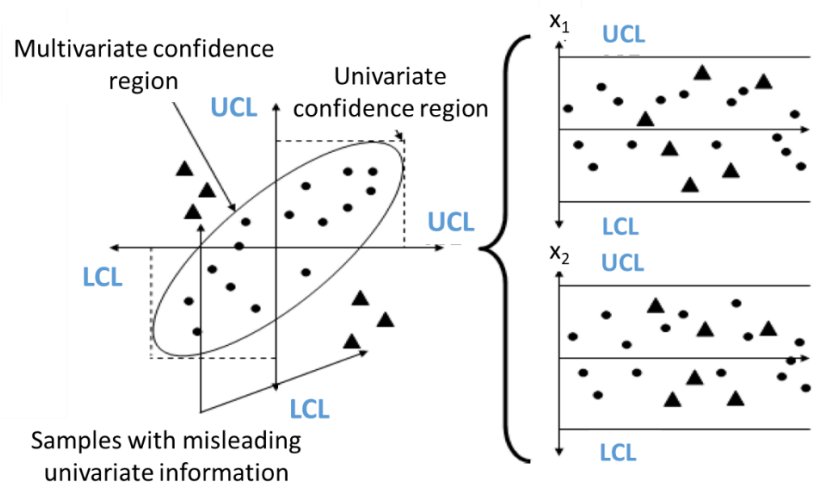
Figure 12 - Overview of a control chart with the upper (UCL) and lower control limits (LCL).



CL: central line

However, SPC methods are based on charting only one or a few product quality variables in a univariate way. This approach, although effective in past times when data were scarce, is totally inadequate for modern process, where massive amounts of highly correlated variables are being collected (HADIAN; RAHIMIFARD, 2019). Relying on univariate control charts when more than one variable is involved is unpractical, as it forces the operator to inspect a large number of control charts; and risky, since they may lead to unsatisfactory results, such as increasing the rate of false alarms (BERSIMIS; PANARETOS; PSARAKIS, 2009) and decreasing the rate of fault detection. This happens because when there is correlation between the variables the correct multivariate in-control region is considerably different from the in-control region determined via individual charts (HADIAN; RAHIMIFARD, 2019), as shown in Figure 13.

Figure 13 - Multivariate vs. univariate approach and comparison of the in-control regions.



Source: Adapted from Ordóñez (2008).

Therefore, multivariate methods that treat all the variables simultaneously are required in these

current data-rich environments. Process monitoring in which multiple correlated variables are of interest is known as MSPC, a method that was first introduced by Hotelling (1947). Since then, several approaches have been proposed for MSPC, such as multivariate Shewhart control charts, multivariate cumulative sum (MCUSUM) control charts and multivariate exponential weighted moving average (MEWMA) control charts (BERSIMIS; PANARETOS; PSARAKIS, 2009). Aiming to consolidate and improve the monitoring of multivariate processes, Jackson (1991) stated that an MSPC procedure should fulfill four conditions: a) the relationships among the variables should be taken into account; b) the question: "Is the process in control?" must be answered; c) an overall probability for the event "Procedure diagnoses an out-of-control state erroneously" must be specified; and d) the question: "If the process is out-of-control, what is the problem?" should be answered. However, the traditional MSPC control charts do not take into account correlation among variables, which may lead to accuracy problems (CAMACHO *et al.*, 2016; KOURTI; MACGREGOR, 1996), and identifying the variable that is causing the process to be out-of-control is not simple in these methods, preventing Jackson's first and last conditions from being fulfilled. Furthermore, these control charts may be impractical for high-dimensional systems with collinearities (BERSIMIS; PANARETOS; PSARAKIS, 2009).

A common procedure for reducing the data dimensionality thus is to use projection methods (also called latent variable methods), like PCA. Latent variable methodologies exploit the correlation structure of the original variables and reveal the few independent underlying events that are driving the process at any time (FERRER, 2014).

2.1.2 MSPC based on PCA (MSPC-PCA)

Latent variables-based MSPC was firstly proposed by Kourti and Macgregor (1996) and has revolutionized the idea of MSPC. For the last 25 years, this approach has been increasingly applied and recommended for monitoring complex industrial processes (CAMACHO *et al.*, 2016), since the performance of an entire unit can be monitored by the operator looking at only a few multivariate control charts, that can be thought of as process performance indices. These charts are simple, easy to understand and have found quick acceptance in the control rooms (KOURTI, 2005). More importantly, latent variables-based charts are able to detect problems that manifest themselves as changes in the covariance structure of the process variables, which traditional control charts will miss if the variables remain within their expected control limits. Besides, the methodology based on latent variables also provides diagnostic tools that help the

operators to determine quickly and efficiently the source of the problem (KOURTI, 2002).

Due to its plentiful benefits (as discussed in the previous item), PCA is the most common multivariate statistical projection method used to reduce the dimensionality of the monitoring space. The process is then monitored in the reduced dimensional space obtained with the first few PC. The MSPC-PCA monitoring method, as any SPC method, is carried out in two phases: Phase I, model building; and Phase II, model exploitation. This distinction is highly relevant and must be done carefully, as the method performance depends strongly on that (CAMACHO *et al.*, 2016). The main goal in Phase I is to model the in-control process performance based on a set of historical in-control data. This dataset must contain only observations in which the process had been operating consistently in an acceptable manner and any periods containing variations arising from special events that one would like to detect in the future must be omitted at this stage. A PCA model and the control charts are thus built according to this in-control dataset. In Phase II then, the PCA model and the control charts built from in-control data are used to monitor the process using on-line data (FERRER, 2014). The limits of the multivariate control charts are calculated according to the Phase I reference dataset and the limit values are defined according to what are good operating conditions for a particular process. On Phase II, values of future measurements are compared against these limits (KOURTI, 2005).

At least two complementary multivariate control charts are required for process monitoring using projection methods: one related to the scores and therefore to the portion of data explained by the model; and one related to the residuals and therefore to the portion of data left out by the model (KOURTI, 2005). Among the multivariate control charts available for MSPC based on projection methods, the most commonly applied are the Hotelling's T^2 statistic (HOTELLING, 1947) and the Q-statistic or sum of squared prediction errors (SPE). T^2 statistic is computed from the scores and represents the estimated Mahalanobis distance from the center of the latent subspace to the projection of an observation onto this subspace. Q statistic in turn is related to the residuals and represents the squared Euclidean distance of an observation from this subspace (CAMACHO *et al.*, 2016; FERRER, 2007). These statistics are calculated for each observation according to Eqn. 21 and Eqn. 22, respectively:

$$T_i^2 = \sum_{k=1}^{K} \frac{t_{i,k}^2}{\lambda_k} \tag{21}$$

$$Q_i = \sum_{k=1}^{K} e_{i,k}^2 \tag{22}$$

where:

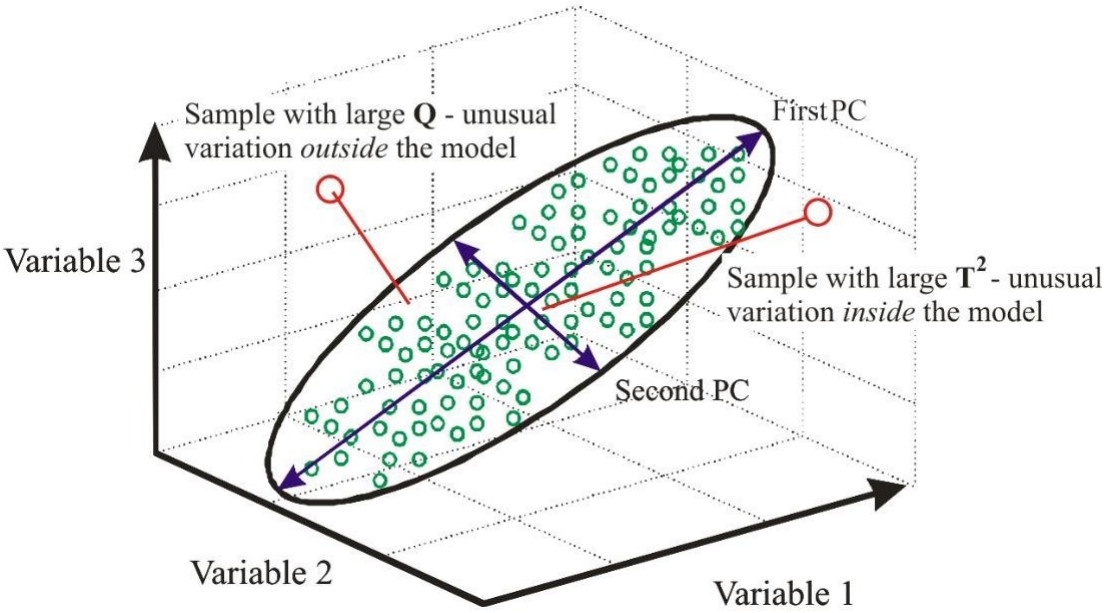
$t_{i,k}$ = i^{th} score value in component k

λ_k = eigenvalue of component k

$e_{i,k}$ = i^{th} residual value of component k

From these two statistics that summarizes all the information contained in the original variables, the respective multivariate control charts are built. The control charts have different conceptual meanings. T^2 control chart essentially checks if a new observation projects on the principal component hyperplane within the limits determined by the reference data. Thereby, a value of this statistic exceeding the control limits indicate that the corresponding observation presents abnormal extreme values in some of its original J variables, even though it maintains the correlation structure between the variables in the model. This observation can be tagged as an abnormal outlier inside the PCA model. Q control chart, in turn, checks the occurrence of any new events that cause the process to move away from the hyperplane defined by the reference model. So, values exceeding the control limits are related to observations that do not behave in the same way as the in-control data because there is a breakage of the correlation structure of the model. This observations can be tagged as outliers outside the model (FERRER, 2014; KOURTI, 2005). For illustration, the measurements of three variables of a process and a model described by two PC are shown in Figure 14. There are two atypical observations: one with unusually large Q value and the other with unusually high T^2 value.

Figure 14 - PCA model of a three-dimensional dataset with emphasis on outlier values of T^2 and Q statistics.



Source: (ORDÓÑEZ, 2008).

When a fault is detected by one of the control charts, a diagnostic approach to isolate the original variables responsible for the out-of-control signal is needed. In MSPC-PCA, due to the nature of latent variable models, the most widely applied approach for fault diagnosis are the contribution plots (CAMACHO *et al.*, 2016; FERRER, 2014).

2.1.3 Diagnosis approach: contribution plots

Contribution plots show the contribution of the variables to the atypical values of the monitoring statistics. In general, they are bar plots where the contribution of the set of variables to a statistic (T^2 or Q) can be inspected. When an out-of-control observation is detected on the Q control chart, the contribution of each variable of the original dataset is simply given by its respective squared residual, as shown in Eqn. 23. Variables with high contributions in this plot should be investigated (FERRER, 2007).

$$\mathrm{contQ_{j,i}} = \frac{e_{i,j}^2}{\sum_{j=1}^{J} e_{i,j}^2} \tag{23}$$

where:

$\mathrm{contQ_{i,j}}$ = Q contribution of j^{th} variable to the i^{th} atypical observation

$e_{i,j}$ = i^{th} residual value of variable j

If the abnormal observation is detected by the T^2 control chart, there are different ways to compute each variable contribution, proposed by several authors. According to Kourti (2005), the diagnosis procedure should be carried out in two steps: (i) a bar plot of the normalized scores (Eqn. 24) for that observation is plotted and the k^{th} score with the highest normalized value is selected; (ii) the contribution of each j^{th} original variable to this k^{th} score at this new abnormal observation is calculated (Eqn. 25) and a plot of these contributions is created. Variables on this plot with high contributions and the same sign as the score should be investigated (contributions of the opposite sign will only make the score smaller).

$$t^*_{i,k} = \left(\frac{t}{\lambda_k} \right)^2 \tag{24}$$

where:

$t^*_{i,k}$ = i^{th} normalized score value in component k

$$\text{cont}T^2_{\text{j,i}} = w_{j,k} * x_{i,j} \tag{25}$$

where:
$\text{cont}T^2_{\text{i,j}}$ = T^2 contribution of j^{th} variable to the i^{th} atypical observation
$w_{j,k}$ = weight of variable j in the k^{th} component
$x_{i,j}$ = i^{th} atypical value of variable j

According to Ballabio (2015), the contribution of each variable to the T^2 statistic is calculated as Eqn. 26:

$$\text{cont}T^2_{\text{j,i}} = \sum_{k=1}^{K} \frac{t_{i,k} * w_{j,k}}{\sqrt{\lambda_k}} \tag{26}$$

Contribution plots are a powerful tool for fault diagnosis; however, the user should be careful with their interpretation. In general, contribution plots will provide a list of the process variables that contribute numerically to the out-of-control condition. The role of the contribution plots to fault diagnosis thus is to indicate which of the variables are related to the fault rather than to reveal the actual cause of it. Those variables and any variables highly correlated with them must

be investigated and the incorporation of technical process knowledge is crucial to accurately diagnose the problem and discover the root causes of the fault (FERRER, 2014; KOURTI, 2005).

2.2 MSPC for monitoring and controlling industrial processes

The application of MSPC for monitoring membrane separation processes (MSP) is still very scarce in the literature. However, MSPC has been extensively and successfully applied for monitoring and controlling industrial processes in several other areas.

Sales *et al.* (2016) applied MSPC for monitoring the soybean oil transesterification in biodiesel production in order to detect and diagnose faults on the reactor operation, in both offline and online monitoring. The authors worked with MSPC based on projection methods, namely PCA. Distance to the Model (DModX), Q and T^2 control charts were constructed from in-control near-infrared (NIR) spectra collected in-line during soybean oil methanolysis. Q and DModX control charts showed high performance regarding offline fault detection, since most of the failures related to the reaction temperature, catalyst content and stirring speed were properly highlighted. T^2 control chart in turn was barely able to identify the failures due to modifications in the agitation speed and to the change in the catalyst content. On online monitoring, most of the failures were also properly highlighted by all three control charts, however, they proved to be much more sensitive to changes in the catalyst concentration and in the temperature conditions, probably owing to the fact that they were sufficiently strong to cause a significant modification in the reaction rate and, thus, in the composition of the reaction mixture. Moreover, contribution plots enabled a clear identification of the spectral region mostly affected by the faults when both the approaches were resorted to.

Liu *et al.* (2017) also developed a MSPC-PCA model to monitor cell cultures aiming to detect contamination using in-line Raman spectroscopy, an analytical instrument that provides large multivariate databases in the biopharmaceutical industry. T^2 and Q multivariate control charts were built and the control limits were calculated based on the confident degree of 99% of the NOC batches. Both control charts were able to identify abnormal operation conditions, including the early detection of contaminated batches, which is of prime importance in cell culture monitoring since it can only be visually detected when the foam due to contamination is visible (which usually takes several hours) and it cannot be detected by the traditional

diagnosis approaches. The application of MSPC thus could save time and money for the biopharmaceutical industry.

Leite *et al.* (2018) in turn applied MSPC to offline monitor single and two stage thermophilic sludge digestion for fault and/or abnormal schemes detection. The authors also worked with MSPC-PCA model and with the T^2 control chart, and they also adopted the Shewhart control chart. Confidence levels of 99% and 95% were used for UCL and upper warning limit (UWL) calculations, respectively, for the T^2 control chart, and the two and three-sigma control limits (two or three sd around the mean) were considered as UCL and UWL, respectively, for the Shewhart control chart. The multivariate control charts applied revealed a transition period in which the stability pattern of the single stage anaerobic digestion changed strongly. Besides, out-of-control samples were detected, indicating an unstable dynamic behavior along time for this digester, which the contribution plot developed to diagnosis the out-of-control signals depicted proved to be caused by an accumulation of volatile fatty acids (VFA). For the two-stage digester, no faults were detected and the operation was considered stable, in accordance with reality.

Catelani *et al.* (2018) expanded the application of MSPC for real-time monitoring to control the end product quality on a coffee roasting process through NIR spectroscopy. Again, the authors worked with MSPC-PCA and with T^2 and Q control charts. However, the control charts were represented in terms of the reduced values of T^2 and Q statistics, dividing each one by the corresponding 95% confidence limit. Therefore, the 95% confidence level control limit was equal to one in both charts. The multivariate control charts effectively detected all the disturbances of different nature imposed to simulate real faulty situations (different roasting conditions and coffee species and origins) and the time region where the deviation was observed was compatible with the type of disturbance imposed. Additionally, the statistics' values of NOC batches were all below the control limits, indicating that this methodology was able not only to signal abnormal batches but also to prevent false alarms for NOC batches. The authors concluded that the application of MSPC with real-time monitoring through NIR spectroscopy is a step forward toward a deeper control over roasting processes since it can decrease the operating costs by avoiding production of faulty roasted coffee batches.

Grassi *et al.* (2019) also combined NIR spectroscopy with MSPC for improve process control in the dairy industry. The authors developed MSPC-PCA T^2 and Q multivariate control charts

aiming the detection of failures in the milk coagulation during cheese manufacturing. For this purpose, fifteen cheese manufacturing batches were set up varying temperature, milk pH, and fat content (NOC batches) and three failure batches were also considered. The 99% confidence interval was considered as the control chart limits. T^2 and Q control charts detected the failure batches at different points, but with the combination of these two control charts, it was possible to detect the in-control tested batch and to distinguish the failure batches just from the first minutes of the process. The authors stated thus that the combination of T^2 and Q charts gave specificity and sensitivity results more reliable than their single check. The authors also discussed that this kind of industrial control systems perfectly fit with the Industry 4.0 roadmap towards a fully digital enterprise.

Le *et al.* (2020) monitored and evaluated the quality of process water in mining industry in order to improve the performance of concentrators for water recycling. MSPC-PCA with T^2 and Q control charts was applied to detect shifts in water quality and to identify associated causes. The authors also highlight that both T^2 and Q control charts should be considered for process monitoring since they take into account both the control observation range (T^2-statistic) and also the structure of the variable correlation (Q-statistic). The 95% confidence limit defined the control region. The model built was able to detect changes in water quality due to modifications of the water circuits and it also identified all four major changes that were implemented in the mine water circuits. Among them, the modification from the long water cycle to the short water cycle completely changed the water matrix and it was assertively indicated by the multivariate control charts. Therefore, MSPC can be an extremely useful tool for mineral engineers and operators to control the water quality in the plant and to make well-informed decisions on process/water circuit modifications. The investigation of the contribution to the T^2-statistic showed that the out-of-control observations had abnormally high concentrations of sulphate and low concentration of thiosalts and calcium.

Zhao *et al.* (2020) also applied NIR spectra and MSPC-PCA for the real-time monitoring of a fluid bed granulation (FBG) process. Three types of control charts were developed: PCA scores, Hotelling's T^2, and DModX. The multivariate control charts were validated on NOC batches and were tested on four batches abnormal operating conditions (AOC) samples to simulate real-time fault analysis. Their results revealed that for NOC batches, the multivariate control charts did not detect abnormalities, whereas for AOC batches, they successfully detected the abnormal situation, identifying all of the imposed disturbances: inlet air temperature failure, atomization

pressure failure, and spray mode failure. Therefore, the applied control charts presented good sensitivity and specificity, and can be used to monitor abnormal batches in the FBG process. Like Catelani *et al.* (2018) and Grassi *et al.* (2019) thus, the authors concluded that the application of MSPC combined with NIR spectra is an attractive tool for real-time process monitoring.

Moreira *et al.* (2021) positively contributed to this topic by applying MSPC for the monitoring of a full-scale wood pellets production for biofuel production. The authors applied T^2 and MEWMA control charts to ensure that the international and regulatory agencies standards regarding product quality were met. The two-sigma control limits were considered as UCL and UWL, respectively. The MSPC model accurately tracked the non-random errors and it was also used to predict the operationally finest scenario for developing high-quality pellets: pre-heating the compressing channel of the machine regularizes the feeding and thus produces pellets with excellent bulk density, durability and hydrophobicity, without points outside the specific critical ranges.

The works presented demonstrate how the application of MSPC for monitoring and controlling industrial processes is standing out in the international literature, due to its high potential for operating faults detection and diagnosis. Most of the works have been applying the T^2 and Q control charts. In addition, the application of contribution plots to identify the main factors that caused the failures has also been highlighted. Therefore, based on the results reported for MSPC for monitoring and controlling different industrial processes, its application on MBR intending to detect operating failures caused by membrane fouling and to detect low ammonia removal percentages is highly promising and should be investigated, since it can support the decision-making and meaningfully contribute to improve MBR performance.

3 ANN

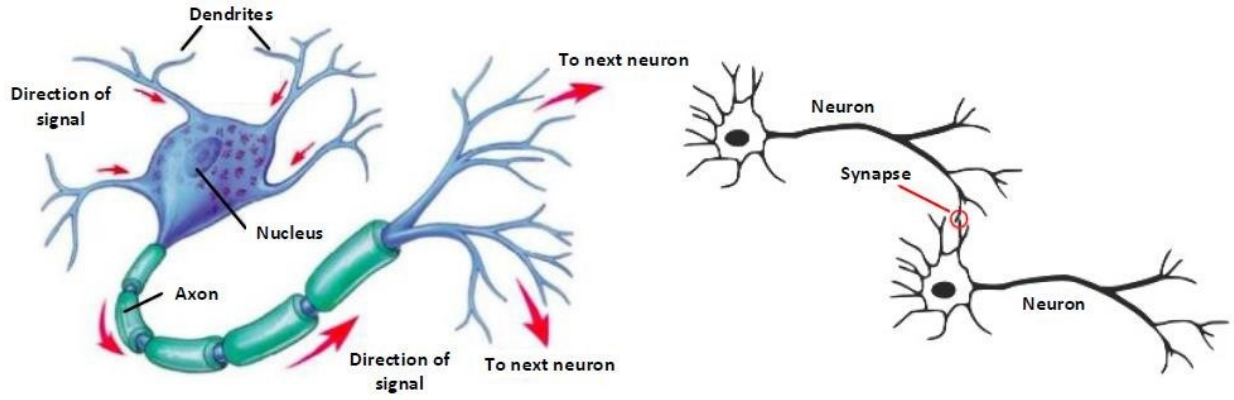
Currently, ANN constitute one of the most important pillars of ML, as they are able to mimic human intelligence, to model complex relationships between inputs and outputs, to find patterns in data and to extract statistical structure from the observed data (CHEN *et al.*, 2019). Although they are a relatively new technology, they have rapidly found extensive acceptance in many disciplines for modelling and solving many complex real-world problems. Technically, ANN are data-processing systems based on and inspired by the neurological networks found in brains. They are mainly used for pattern identification and processing, and are able to progressively improve performance based on results from previous tasks (BASHEER; HAJMEER, 2000). The attractiveness of ANN comes from their remarkable information processing characteristics such as high parallelism, nonlinearity, fault and noise tolerance, and learning and generalization capabilities (HAGLIN; JIMENEZ; ELTORAI, 2019).

3.1 Theoretical foundation

3.1.1 General architecture of ANN

ANN are essentially an artificial model of a human nervous system. They are built from simple but highly interconnected processing elements known as neurons, which are used to mimic how the human brain learns. Biological neurons consist of nucleus, dendrites and axons (the last two connect neurons to each other) and the connection point between two neurons is called synapse, as shown in Figure 15. The information signal received by a neuron changes its membrane potential and, if it exceeds a certain value, the neuron will send a signal to all of its connected neurons. This is how signals propagate through the human nervous system.

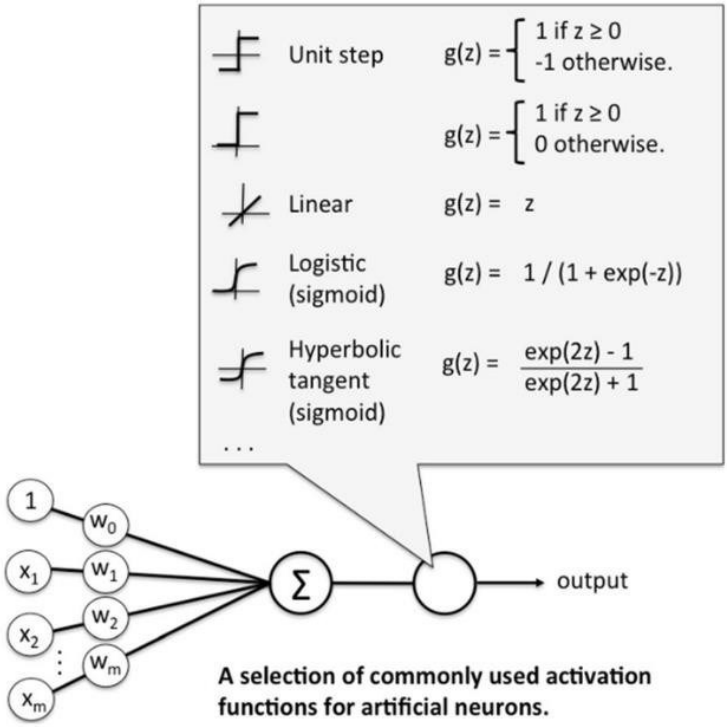
Figure 15 - Schematic representation of a biological neuron.



Source: (CHEN *et al.*, 2019).

ANN use artificial neurons to replicate this operation of the human nervous system, enabling AI. Mathematically, an artificial neuron consists of the following components: i) a number of incoming connections, analogous to synapses on the dendrites; ii) a number of outcoming connections, analogous to synapses on the axon; and iii) an activation value assigned to each neuron, analogous to a biological neuron's membrane potential. Each connection between two neurons has a strength captured by a weight value. The magnitude of the weight controls the strength of the influence of that input on the receiving neuron whereas its sign controls whether the influence is stimulating or inhibiting the signal to the next layer. The basic model for an artificial neuron *j* is shown in Figure 16 and mathematically given by Eqn. 27 (CHEN *et al.*, 2019):

Figure 16 - Schematic representation of an artificial neuron.



Source: (HAGLIN; JIMENEZ; ELTORAI, 2019).

$$o_j\ (\mathbf{w_j}, b_j, \mathbf{x_j}) = \mathrm{f}\left(b_j + \sum_{i=k}^{N} x_{j,k}\ w_{j,k} \right) \tag{27}$$

Where:

o_j = is the output signal of neuron *j*;

$\mathbf{x_j} = [x_{j1};\ x_{j2};\ \ldots;\ x_{jN}]$ is the vector of input signals of neuron *j*;

$x_{j,k}$ = is each input signal of neuron *j;*

$\mathbf{w_j} = [w_{j1}; w_{j2}; \ldots; w_{jN}]$ is the vector of input weights of neuron *j*;
$w_{j,k}$ = is the corresponding input weight value;
b_j = is the bias of neuron *j;*
f = is a nonlinear activation function;

The nonlinear activation function (e.g. a logistic function) will determine neurons output values based upon the values of their inputs. The selection of the activation function depends on the sought objectives, available computational power and the type of the desired output signal (logistic or continuous) (CHEN *et al.*, 2019). They also represent the rate of action triggering a neuron, since only when o_j exceeds (i.e., is stronger than) the neuron's threshold limit (also called bias, b_j), will the neuron becomes activated (BASHEER; HAJMEER, 2000).

This way, ANN consist thus of combining multiple neurons connected in structured layers. The neurons in a given layer are independent of each other, but each of them connect to all neurons in the next layer. In general, ANN will include the following layers (CHEN *et al.*, 2019):

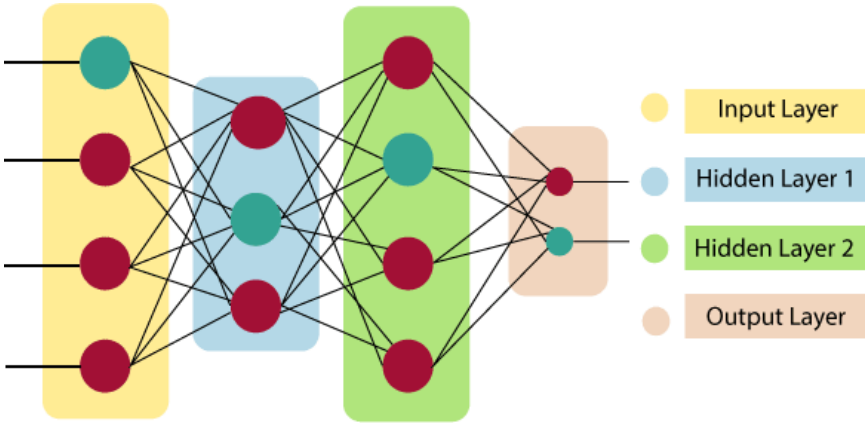
i. One input layer that contains a neuron for each predictor variable and represent the input signals that will be transmitted through the neurons;
ii. One or more hidden layers containing a user defined number of neurons. Each neuron in the first hidden layer receives an input from each neuron in the input layer. If there is a second hidden layer, each neuron in this layer receives an input from each neuron in the first hidden layer, and so on with additional layers. The hidden layer is used to analyze the relationship between the input signal in the input layer and the output signal in the output layer;
iii. One output layer that contains one neuron for each response variable (usually it consists of only one neuron, but in multivariate response situations there can be more neurons). Each output neuron receives an input from each neuron in the final hidden layer and they represent the ANN output signal.

The pattern of connection links between the different layers is called architecture (also known as type or structure of an ANN) and it plays an important role in the ANN performance. The choice of the type of network used is up to the user and is normally related to the type of data available and the purpose of the network.

3.1.2 Different types of ANN

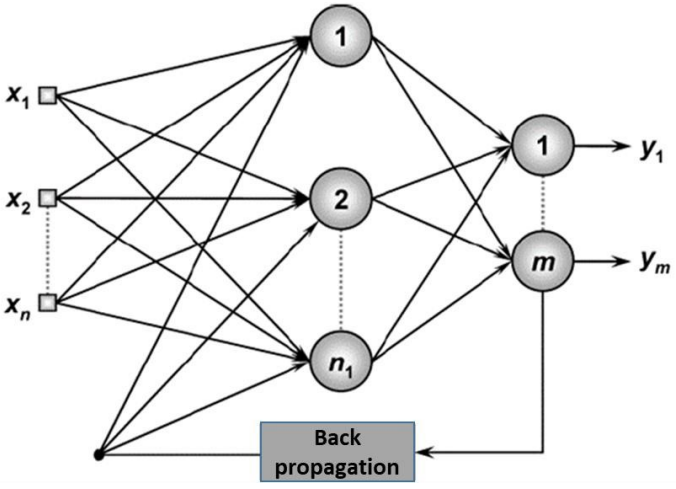
One of the simplest but at the same time most common types of ANN is the feed-forward neural network (FNN). In this architecture, the connection between the neurons is unidirectional and there is no connection between the neurons in a layer. Each neuron in the hidden layers has incoming connections only from the previous layer and outgoing connections only to the next layer, as shown in Figure 17 (WASZCZYSZYN, 1999).

Figure 17 - Feed-forward neural network (FNN) representation.



If the connections between neurons form a loop though, the network is called a recurrent neural network (RNN). This architecture allows connections from a neuron in one layer to neurons in previous layers (Figure 18). This seemingly simple change enables the output of an ANN to depend not only on the current input but also on the historical input, enabling the network to make use of sequential information and exploit dynamic temporal behaviors such as those faced in mobility prediction, handwriting recognition, or speech recognition (SILVA *et al.*, 2010).

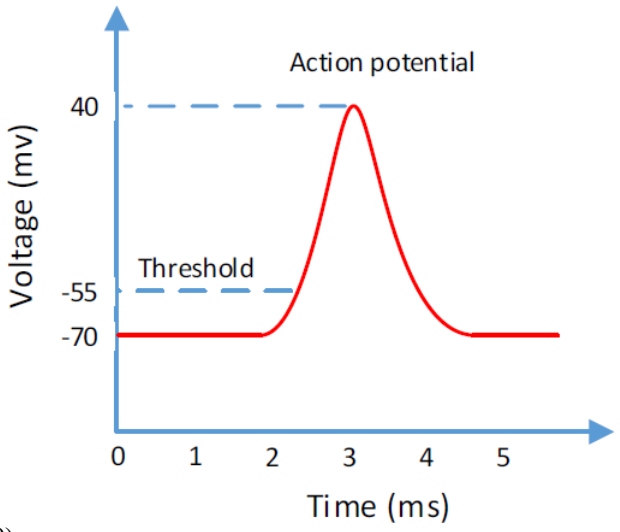
Figure 18 - Recurrent neural network (RNN) representation.



Source: (SILVA *et al.*, 2010).

Another important type of ANN is the so-called spiking neural networks (SNN). In contrast to other ANN such as FNN and RNN that simply use a single value to denote the activations of neurons, SNN use a more accurate model of biological neural networks to denote the activations of neurons. In biological neural networks, neurons use spikes to communicate with each other. Incoming signals alter the voltage of a neuron and when the voltage reaches above a threshold value, the neuron sends out a short and sudden increase in voltage as an action potential, which is referred to as a spike or a pulse, like represented in Figure 19. After sending out a spike, the neuron enters a short moment of rest, the refractory period, during which it cannot send out a spike again. In RNN and FNN, the synaptic weight values are always non-negative, which means that if two neurons are connected, then the activation value of each one will increase. In SNN, in turn, the weight values can be negative due to inhibitory neurons. Thereby, the use of spikes on SNN can meaningfully improve the dynamics of the network, leading to two major advantages over traditional neural networks: fast real-time decoding of signals and high information carriage capacity by adding a temporal dimension. However, the training of SNN will be more challenging and potentially more time consuming (CHEN *et al.*, 2019).

Figure 19 - Action potential of each spiking neuron.



Source: (CHEN *et al.*, 2019).

Finally, deep neural networks (DNN) is an ANN with multiple hidden layers between the input and output layers. The multiple layers enable high-level abstractions in data through multiple nonlinear transformations. All types of ANN previously presented can also be DNN, which improves their performance. The FNN represented in Figure 17, for example, is also a DNN. Several reasons contributed to move from conventional single layer ANN towards DNN, like: improved computational capacity, which has resulted in a faster and more parallelized

computation and thus decreased the processing time; availability of large amounts of data that has made the training of DNN possible; improved training algorithms, like the use of rectified activation function instead of sigmoid or tanh, which has made training faster (RUSSEL; NORVIG, 2022).

3.1.3 Learning process in ANN

To learn information from input data, ANN must adjust the weights of the connections between the neurons and the biases of each neuron in the system. The training process consists precisely of adjusting and updating the weights and biases aiming to fit the data in the best possible way. To do so, different learning tasks require different training algorithms. For example, to perform supervised learning tasks such as predictions, ANN must be trained using labeled data. For unsupervised learning tasks such as clustering, ANN is trained without labeled data. By all means, the main goal of training an ANN is to minimize the errors between obtained output signals and desired output signals. This error (**E**) can be typically defined as Eqn. 28 (CHEN *et al.*, 2019):

$$\mathbf{E}(\mathbf{W},\mathbf{b}) = \sum | o(\mathbf{W},\mathbf{b},\mathbf{x}) - o_D | \tag{28}$$

Where:
W = the weight matrix, which is a combination of input weight values, hidden weight values, and output weight values;
b = is the vector of bias factors;
x = is the vector of input signals;
$o(\mathbf{W}, \mathbf{b}, \mathbf{x})$ = is the obtained output signal, calculated by Eqn. 27;
o_D = is the desired output value.

In order to minimize **E**(**W**, **b**) thus, it is needed to update the synaptic weights and biases related to each neuron. One of the most common approaches to do this is using a gradient descent algorithm. In mathematics, gradient descent (often called steepest descent) is a first-order iterative optimization algorithm for finding the local minimum of a differentiable function. The idea is to take repeated steps in the opposite direction of the gradient of the function at the current point, since this is the direction of steepest descent (LEMARÉCHAL, 2012). This way, for each neuron *j*, the minimization of $E_j(\mathbf{w_j}, b_j)$ using gradient descent algorithms follows from the following Eqn. 29 and Eqn. 30 (CHEN *et al*, 2019; AMARI, 1993):

$$w_{j,k,new} = w_{j,k,old} - \gamma \frac{\partial \mathrm{E_j}(\mathbf{w_j}, b_j)}{\partial w_{j,k}} \tag{29}$$

$$b_{j,new} = w_{j,old} - \gamma \frac{\partial \mathrm{E_j}(\mathbf{w_j}, b_j)}{\partial b_j} \tag{30}$$

Where:

γ = is the learning rate, which has a strong impact on optimization performance. The smaller the learning rate, the longer the algorithm will take to converge and it may reach maximum iterations before reaching the optimum point. On the other hand, if the learning rate is too high, the algorithm may jump around the optimum point or even diverge completely. Therefore, its value must be defined carefully and there are different approaches for its definition (AMARI, 1993).

The first order derivative allows to determine whether the error is decreasing or increasing when the weight value is $w_{j,k}$ and the bias value is b_j. Based on Eqns. 29 and 30, thus, ANN can update the weight and bias values to find the optimal $\mathbf{w_j}$ and b_j that will minimize $\mathrm{E_j}(\mathbf{w_j}, b_j)$. This is accomplished by repeating the iterative steps until one of the criteria is met: i) the maximum number of iterations reached; or ii) step size is smaller than the tolerance (AMARI, 1993). After minimizing the error of neuron *j* then, backpropagation (which is basically a chain rule) is the most widely used algorithm to calculate the gradient of the error and to effectively minimize $\mathbf{E}(\mathbf{W}, \mathbf{b})$ for an ANN.

Backpropagation algorithm will thus progressively apply the gradient descent algorithm and compute the error between the desired outputs and actual outputs based on Eqn. 28 to derive an error propagation value δ_j for each neuron *j*, as Eqn. 31:

$$\delta_j = \frac{\partial \mathbf{E}(\mathbf{w_j}, b_j)}{\partial o_j} \frac{\partial o_j}{\partial \mathrm{n_{sum,j}}} \tag{31}$$

Where:

$\partial \mathrm{n_{sum,j}}$ = is the summation of all input signals of neuron *j* and its bias, calculated according to Eqn. 32:

$$\partial \mathrm{n_{sum,j}} = b_j + \sum_{i=k}^{N} x_{j,k} w_{j,k} \tag{32}$$

As the input signals reach the output layer after being transmitted from the input layer to each hidden layer, the error propagation δ_j of a neuron in layer L depends on the error propagation of a neuron at layer L+1. Therefore, each neuron must transmit its error propagation parameter to the neurons at the former layer. This is the central definition of backpropagation (AMARI, 1993).

One of the biggest advantages of gradient descent method relies on te fact that $w_{j,k}$ and b_j updates are easy to compute and, hence, the gradient descent algorithm is known to be computationally fast, even on large datasets. However, choosing a proper learning rate for the update of the weights and bias can be difficult and gradient descent algorithms can often converge to a sub-optimal local minimum rather than the global minimum. Aiming to overcome these challenges, several algorithms have been proposed, such as stochastic gradient descent (SGD) algorithm, nesterov accelerated gradient, Adagrad and AdaDelta and pruning techniques (CHEN *et al.*, 2019). Nevertheless, gradient descent algorithm is still the most widely used optimization method in ML and, therefore, the others algorithms will not be discussed in this work.

Two central problems in training ANN are overfitting and underfitting. Overfitting corresponds to the case in which the model learns the random fluctuations and noise in the training dataset to the extent that they negatively impact the model's ability to generalize when fed with new data. This occurs specially when the dataset is too small compared to the number of parameters that must be learned. On the other hand, underfitting occurs when a learning algorithm cannot capture the underlying trend of the data, i.e., the learning algorithm does not fit the data well enough. This occurs mainly due to insufficient amount of training data or too simple modelling.

To assess the model performance and understand if it is properly fitting the data thus, statistical metrics are fundamental. Some of the most applied metrics are the R^2, the Mean Average Error (MAE) and the Mean Squared Error (MSE) or the Root Mean Squared Error (RMSE). In short, MAE evaluates the absolute distance between the observations and predictions on a regression, taking the average over all observations, as Eqn. 33 (CHUGH, 2020). It represents the average of the absolute difference between the actual (x_i) and predicted values ($\hat{x}_i$) (residuals) in the dataset.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{x}_i| \tag{33}$$

MSE in turn represents the average of the squared difference between the original and predicted values in the dataset (Eqn. 34) (CHUGH, 2020) and therefore it measures the variance of the residuals. MSE is more sensitive to outliers than MAE, since higher errors weigh more in the metric than lower ones, due to the nature of the power function (TREVISAN, 2011).

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_i)^2 \tag{34}$$

A backlash in MSE is the fact that the metric's unit is also squared. RMSE is the square root of MSE and therefore it measures the sd of the residuals and returns the metric to the original unit, while maintaining the property of penalizing higher errors (TREVISAN, 2011).

Lower values of MAE, MSE and RMSE imply higher model accuracies, whereas a higher value of R^2 is desirable. In order to ensure a good quality model thus, it is needed to carefully choose the ANN architecture, along with proper training methods to avoid both over and underfitting.

3.2 ANN for monitoring MBR performance

Recent works regarding monitoring MBR performance through ANN modelling have showed good prediction of real data, despite the complexity of MBR systems.

Yusuf *et al.* (2015) applied ANN to assess membrane fouling on submerged MBR treating palm oil mill effluent. The authors employed feed-forward and radial basis neural networks, fed with three input filtration parameters (aeration airflow, suction pump voltage and TMP) to predict the permeate flux in the filtration process under schedule relaxation condition. The performance of both ANN was evaluated from R^2 and MSE and the results showed that feed-forward architecture presented better performance compared with radial basis regarding accuracy and reliability. The work also showed that the permeate flux at low aeration airflows is faster to decline compared to high aeration airflows.

Giwa *et al.* (2016) introduced a new configuration of an electrically-enhanced MBR to treat medium strength wastewater. The investigated components in this study were COD, phosphates (PO_4^{3-}-P) and ammonium (NH_4^+- N). Variation in environmental compositions such as DO, MLVSS, pH, and electrical conductivity influenced the effluent concentration. A feed-forward with sigmoidal activation function ANN was then used to model the experimental findings of COD, PO_4^{3-}-P and NH_4^+- N removal given the initial compositions. Comparison between the

model results and experimental data set gave high correlation coefficients for COD (R^2=0.9942), PO_4^{3-}-P (R^2 = 0.9998) and NH_4^+- N (R^2 = 0.9955).

Schmitt *et al.* (2018) also applied ANN modelling to investigate membrane fouling on a pilot-scale anoxic-aerobic membrane bioreactor (AO-MBR) treating domestic wastewater. The goal was to select the most relevant input parameters to predict the evolution of TMP. However, MLSS, COD, pH and DO, common parameters of wastewater treatment processes could not be linked to TMP as the performances obtained by the ANN taking them as input variables were not satisfying (from R^2 = 0.169 for DO to less than 0.70 for COD). The poor ANN performance can be explained by an inappropriate setting of the network. ANN modelling performance is highly dependent on the initial set of weights and biases created at the beginning of the training phase, along with the dataset division for training, validation and testing. For next models thus, the authors recommend that the input dataset is constituted of representative data and that particular attention is payed to choosing the proper settings for training the network, such as the initial set of weights and biases, the division of the database and the definition of the best architecture, i.e. the number of hidden layers and neurons.

Hosseinzadeh *et al.* (2020) in turn assessed membrane fouling on OMBR through Adaptive Network-based Fuzzy Inference System (ANFIS) and ANN models, developed for simulating and predicting water flux in OMBR systems. MLSS, electrical conductivity (EC) and DO were used as inputs to a feed-forward ANN. The number of neurons in input and output layers were determined according to the number of input and output variables. The number of neurons in hidden layers in turn was defined based on the model that led to the lowest R^2 and MSE. Data was divided into two sections of 80% and 20%; the first section was used for training, validation and test with portions of 70%, 15% and 15%, respectively; and the rest was applied for an additional test. Good prediction was demonstrated by ANFIS, with R^2 of 0.9755 and 0.9861, and ANN models, with R^2 of 0.9404 and 0.9817, for thin film composite (TFC) and cellulose triacetate (CTA) membranes, respectively. Sensitivity analysis showed that EC was the most important factor for both TFC and CTA membranes in ANN model, while EC (TFC) and MLSS (CTA) were key parameters in ANFIS model.

Banerjee *et al*. (2022) used ANN modelling to investigate the performance of an indigenously developed ceramic UF membrane in a lab-scale MBR treating real tannery wastewater with varying OLR. ANN was used to analyze the influence of HRT (4–10 h), MLSS (2–8 g/L) and

influent COD (1500–6000 mg/L) on COD removal percentages. A feed-forward single-layer (only one hidden layer) ANN architecture was used, accounting from 1-30 neurons, and tan-sigmoid and linear activation functions were applied. MSE and R^2 were used to evaluate the network performance. ANN modeling revealed that COD removal efficiency was influenced mostly by MLSS among other input variables. Influent COD is another important input parameter and HRT has the least impact among the three input variables. Satisfactory accuracy was found between the experimental and the predicted removal efficiency, with R^2 as 0.9974.

Kovacs *et al*. (2022) used data-driven ML techniques consisting of RF, ANN and long-short term memory network (LSTM) to predict TMP at various stages of the MBR production cycle, again aiming to model membrane fouling on these systems. The models were built from a four-years monitoring dataset from a full-scale municipal WWTP and their performances were examined using the statistical measures R^2, RMSE, MAE and MSE. For the ANN model, a feed-forward multi-layer architecture was used with a sigmoidal activation function. The results showed that all models provided reliable predictions and the difference in performance metrics between RF and ANN were low in magnitude, with RF models being more accurate in training and testing. However, ANN model made a higher number of proper extreme predictions, which is interesting from a control point of view, since the predictive performance for all models decreases when aiming to predict extreme values. The authors concluded that the proposed models can be useful tools in providing decision support to WWTP operators employing fouling mitigating strategies, leading to reduced operational costs.

The works presented illustrate the growing application of ANN models to investigate different aspects of the operation of a MBR. It is worth noticing the predominance of studies referring to membrane fouling, as previously discussed in this work, and again demonstrating how this is a matter of high importance for MBR further development. Moreover, the works presented reveal the great potential of ANN to model and predict MBR behavior with high accuracy. However, they also demonstrate how important it is for the network to be well configured and trained to make this possible. It is also noticed the predominance of feed-forward multilayer networks, due to their good results, and the lack of a more well-defined approach to determine the network settings, as number of hidden layers, number of neurons in each of them and activation function, being the trial-and-error approach the most common alternative.

III. IMPROVING MEMBRANE FOULING CONTROL

1 INTRODUCTION

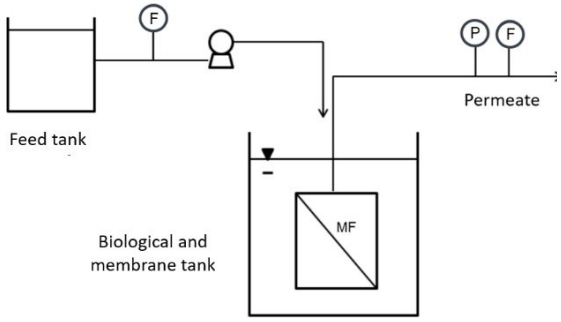
In this chapter, the relations between different operating and analytics variables of a pilot-scale MBR treating a real oil refinery wastewater were investigated by applying PCA and MSPC, aiming to comprehend the main causes of membrane fouling and to propose efficient strategies for its control. The MBR was monitored for five years and the variables were chosen so valuable information could be obtained on both MBR treatment mechanisms, i.e., biological degradation and membrane separation. MSPC and PCA models were developed in software R and have proven to be suitable for monitoring MBR wastewater treatment systems aiming membrane fouling control. PCA modelling was effective in identifying the most important variables for membrane fouling occurrence and in predicting the MBR behavior, which allows to distinguish samples with atypical behavior and enables the detection of operating problems. T^2 and Q control charts were able to preventively detect membrane permeability reduction and thus can be used to guide proper fouling mitigation strategies. Although these results were obtained from offline data, the models applied here have proven to be powerful tools for the assessment of the system state in real-time, as long as online monitoring data is available, which would promote better-informed decision-making regarding membrane fouling control. It is important to mention, though, that membrane fouling was evaluated herein by membrane permeability, which is also reduced by clogging. This way, the distinction between the two phenomena is not possible with the adopted methodology.

2 METHODOLOGY

2.1 MBR configuration and performance

The pilot-scale MBR used as a case study in this research is installed in a Brazilian oil refinery and it was monitored during five years. The unit consists of a biological tank equipped with a submerged flat sheet PES (polyether sulfone) membrane module (Kubota) (Figure 20). The oil refinery wastewater was fed to the MBR pilot unit after a series of pretreatment steps including water and oil separation, flotation, sand filtering and hydrogen peroxide dosing for sulphides concentration control. Table 2 shows the characteristics and the operating conditions of the unit.

Figure 20 - MBR pilot unit scheme.



MF: microfiltration

Table 2 - MBR design conditions.

Parameters	Value	Unit
Biological tank volume	8	m^3
Aeration flow	45	Nm^3h^{-1}
Hydraulic retention time (HRT)	5.6	h
Sludge retention time (SRT)	40	d
Organic load	13	$kgCODd^{-1}$
Permeate flow	0.5	m^3h^{-1}
Membrane area	70	m^2
Pore size	0.4	µm
Membrane configuration	MF/flat sheet	

MF: microfiltration

The driving force for permeation was the hydrostatic pressure of the water column and the unit had an aeration system to ensure oxygen to the biological process requirements and to ensure fouling control by the shear stress caused by the ascending flow of air bubbles. The permeate flow was maintained by allowing one min of relaxation after every nine min of permeation. Regarding chemical cleaning, the membrane was periodically submitted to a recovery cleaning with a 5,000 mg L^{-1} sodium hypochlorite (NaClO) solution.

The MBR performance was assessed by periodically monitoring: i) operating conditions –TMP, temperature, permeate flow, membrane permeability, SRT and pH; ii) sludge characteristics - MLVSS, MLSS, sludge filterability, sedimentability, EPS and SMP; and iii) feed and effluent characteristics – biological oxygen demand (BOD), COD, TOC, alkalinity, ammonia, chloride, sulphides, phosphorous, oil and grease (OG), turbidity, color and conductivity. Physicochemical parameters were measured according to the Standard Methods for Examination of Water and Wastewater (APHA; AWWA; WEF, 2012); sludge filterability was measured according to the Kubota recommended method (Filter Test - FT), in which 50 mL of

sludge are filtered in a filter paper (Whatman 42 185 mm) folded with pleats with the aid of a simple funnel and the volume filtered during the initial five min is recorded as the sludge filterability (mL $5min^{-1}$); and membrane permeability was calculated by dividing the measured permeate flow by the membrane area and by the pressure applied to the system. Table 3 presents the MBR performance regarding pollutants removal.

Table 3 - MBR performance in pollutants removal.

		Alkalinity ($mg\ L^{-1}$)	Ammonia ($mg\ L^{-1}$)	BOD ($mgO_2\ L^{-1}$)	COD ($mgO_2\ L^{-1}$)	TOC (ppm)	Turbidity (NTU)
MBR feed	n	225	560	10	755	705	155
	Min	107	9.50	44.0	91.0	30.6	1.02
	Max	502	82.8	330	7070	850	236
	Mean	250	29.5	238	539	160	15.2
	Median	245	27.3	256	463	144	8.51
	sd	61.3	9.70	92.5	350	76.8	24.1
MBR permeate (Removal)	n	225	561	10	756	706	156
	Min	0.00 (-74%)	0.14 (-72%)	2.00 (95%)	5.96 (-3%)	2.63 (-43%)	0.28 (42%)
	Max	378 (100%)	38.6 (99%)	8.00 (99%)	499 (98%)	643 (99%)	4.64 (100%)
	Mean	68.1 (73%)	2.89 (90%)	4.50 (98%)	106 (79%)	30.2 (80%)	0.90 (87%)
	Median	41.2 (81%)	1.39 (95%)	4.50 (98%)	97.0 (82%)	26.9 (82%)	0.65 (91%)
	sd	70.4 (27%)	4.39 (17%)	2.32 (1%)	64.6 (14%)	26.4 (10%)	0.71 (11%)

n: number of samples; min: minimum; max: maximum; sd: standard deviation. Values in parentheses refer to percentage of removal. Negative removal values are due to sporadic samples in which concentration in the permeate was higher than the concentration in the feed.

It is noticeable the high performance of the unit in the treatment of the wastewater, with mean removal efficiencies above 70% for all the pollutants. For COD, TOC and turbidity the MBR presented mean removal efficiencies around 80% and above 90% for ammonia and BOD. The removal efficiencies observed for the MBR are in accordance with which is reported in the literature for the treatment of oil refinery wastewater using pilot-scale MBR (KARRAY *et al.*, 2020; SAMBUSITI *et al.*, 2020).

2.2 Database and preliminary statistical analysis

The monitoring data was provided by the oil refinery and consisted originally of 1,131 samples, collected during the five years. Each sample refers to a day in the monitoring and covers a set of 30 variables, specifically: TMP, temperature, permeate flow, membrane permeability, SRT, sludge filterability, sedimentability, MLVSS, MLSS, EPS, SMP, pH, influent and effluent TOC, BOD, COD, ammonia, turbidity and alkalinity, effluent chloride, color and conductivity, and influent sulphides, phosphorous and OG. However, these parameters were not monitored with the same frequency and so the dataset had a large number of missing data. For the development

of the statistical models thus, some of the variables were selected so that valuable information could be obtained on both biodegradation and membrane filtration treatment mechanisms and that the greatest possible amount of samples could be kept, since it was decided not to work with missing data. The variables COD of the feed, sludge filterability, MLVSS, pH, temperature and membrane permeability were then selected. A variable named 'sequential days without cleaning (SDWC)', which represents how many days the membrane had been working without having a chemical cleaning, was also added to the variable set.

Since even these variables were not measured with the same frequency though, the dataset still presented several missing data and, therefore, samples that did not contain measurements for all seven selected variables were removed from the database, reducing it to 728 samples. Some of these samples presented censored data (values below the method quantification limit - MQL) for the variable sludge filterability. For these samples, the censored data were replaced by the method's threshold value (5 mL $5min^{-1}$). Furthermore, to ensure data consistency, the presence of outliers was investigated using the interquartile range (IQR) (SCHWERTMAN *et al*., 2004) and the Hampel identifier methods (DAVIES; GATHER, 1993). IQR is a measure of statistical dispersion equal to the difference between the 75^{th} and 25^{th} percentiles (Q3 and Q1, respectively), i.e. IQR = Q3−Q1. In the IQR method thus, observations that fall below Q1−1.5*IQR or above Q3+1.5*IQR are defined as outliers. Hampel in turn suggested that outliers were identified from more robust statistical estimates such as the median and the median absolute deviation (MAD), calculated according to Eqn. 35:

$$\text{MAD} = \text{median}(x_{i,j} - m_j) \tag{35}$$

where:

$x_{i,j} = i^{th}$ value of variable j

m_j = median of variable j

So, according to Hampel identifier method, any observation that lies outside the moving interval (MI) (Eqn. 36) should be considered an outlier. The factor 1.4826 is used so that the expected MAD is equivalent to the sd of normally distributed data (YAO *et al.*, 2019).

$$\text{MI} = m_j \pm 3*(1.4826 * \text{MAD}) \tag{36}$$

To avoid distortion in the analyses and results though, all observations classified as outliers by

any of the methods were individually examined to verify if there was proven inconsistency in the data. For this evaluation, a report of occurrences provided by the oil refinery along with the monitoring dataset was consulted to check if any operating problem could justify the extreme values. Nevertheless, after the individual and careful verification of all extreme observations, it was concluded that all of them should be kept in the dataset, since they were all possible to occur and especially because they referred to atypical operating conditions and therefore added important information about the process control. This way, the final database consisted of 728 samples covering seven variables that approach biomass and feed characteristics and operating conditions. Table 4 presents the descriptive statistics for the final database.

Table 4 - Descriptive statistics of the seven selected variables in the final database.

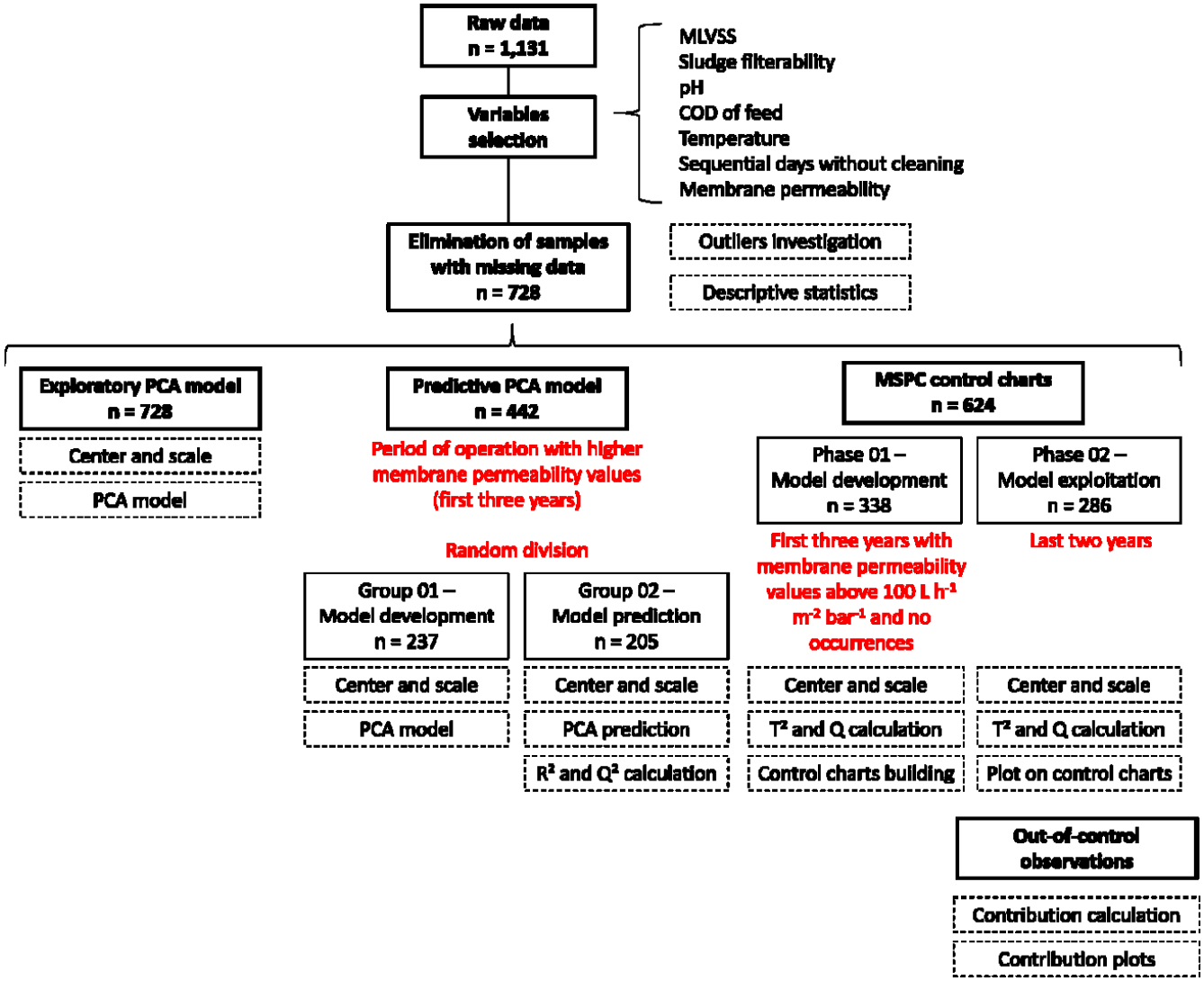
	Sludge filterability (mL 5min^{-1})	**MLVSS (mg L^{-1})**	**pH (-)**	**COD (mg L^{-1})**	**Temperature (°C)**	**Sequential days without cleaning (d)**	**Membrane permeability (L h^{-1} m^{-2} bar^{-1})**
n	728	728	728	728	728	728	728
Min	5.00	2150	5.61	91.0	22.1	0.00	46.3
Max	33.0	14350	10.5	1617	46.9	370	1463
Median	11.0	7506	7.53	452	28.5	67.0	166
Mean	12.4	7551	7.87	517	29.4	94.0	261
sd	6.05	2126	0.89	244	5.12	92.7	230
Variance	36.6	4521435	0.80	59543	26.2	8590	53004
Coef. of variation	0.49	0.28	0.11	0.47	0.17	0.99	0.88
Coef. of skewness	1.07	0.25	1.10	1.23	1.37	1.22	2.08

n: number of samples; min: minimum; max: maximum; sd: standard deviation; coef: coefficient.

2.3 Multivariate statistical analyses

Figure 21 displays a schematic plot of data processing for the pilot-scale MBR. All multivariate statistical analyses were performed using software R version 4.0.2 (R CORE TEAM, 2020). For data importation from Microsoft Excel® 2016, the readxl package was used (WICKHAM, 2019). The R scripts developed to execute the statistical analyses are presented in Appendix A (PCA) and Appendix B (MSPC). More details are described in the following items.

Figure 21 - Schematic plot of data processing for membrane fouling assessment.



2.3.1 PCA

PCA was firstly performed considering the whole database (n = 728) to obtain an overview of the data and to explore the strongest relations between the variables and the observations. The data was previously mean-centered and scaled to unit variance according to Eqn. 2.

PCA was then computed using the PCA function found in the FactoMineR package (LÊ *et al.*, 2008). Loading plots and biplots were built using the factoextra (KASSAMBARA; MUNDT, 2020), ggpubr (KASSAMBARA, 2020), ggplot2 (WICKHAM, 2016) and scales (WICKHAM; SEIDEL, 2020) R packages. In order to further explore the data and the PCA results, boxplots were also plotted and the nonparametric Kruskal-Wallis statistical test (KRUSKAL; WALLIS, 1952) followed by the multiple comparison test of Dunn (ZAR, 1999) at a significance level of 1% were applied to compare the different years of monitoring. For running the nonparametric

statistical tests, the rstatix package (KASSAMBARA, 2021) was used.

From these plots, it was possible to identify a period of higher membrane permeability values, i.e. milder membrane fouling occurrence and consequently more stable operation (first three years), and a period of lower membrane permeability values, i.e. more severe membrane fouling occurrence and thus a more instable operation (last two years). So, in a second moment, only the data from the period during which the membrane permeability was higher was considered and PCA was performed again in order to evaluate its quality as a predictive model. For this purpose, the data from the first three years of monitoring (n = 442) was randomly divided into two groups: Group 01 was used to develop the model (n = 237), whereas Group 02 was left out to be predicted by the model (n = 205). To randomly divide the dataset into the two subsets and at the same time grant reproducibility, the R functions rbinom and set.seed(37645) were used.

Both groups of data were previously mean-centered and scaled to unit variance (Eqn. 2). Group 01 was used to build the PCA model and Group 02 was then projected on the built model. The number of components to keep was determined based on the Kaiser criterion (KAISER, 1960), the scree test (CATTELL, 1966) and the fraction of variation explained (JOLLIFFE, 2002), as often adopted in the literature (FORKMAN *et al*, 2019; RODRIGUEZ-SANCHEZ *et al.*, 2019). For engineering purposes, 60% of total variation explained is usually considered enough to well represent the system (HONG *et al.*, 2019a; JACQUIN *et al.*, 2018) and therefore this fraction was considered as the cut-off.

The quality of the model was thus evaluated from R^2 and Q^2 parameters, calculated as shown in Eqn. 17 and Eqn. 20, respectively (ERIKSSON *et al.*, 2013).

2.3.2 MSPC

To generate the multivariate control charts and thus monitor the MBR performance regarding membrane fouling, a latent-variable based approach for MSPC was adopted, using PCA as the statistical projection method, as presented by Ferrer (2007) and also adopted by Liu *et al*. (2017) and Taghezouit *et al*. (2020). The number of components to keep was again determined based on the Kaiser criterion, the scree plot and the fraction of variation explained.

As any SPC scheme, the methodology was carried out in two phases: Phase I, model building with in-control data; and Phase II, model exploitation with both in-control and out-of-control

data. The out-of-control criterion applied was the minimum acceptable value for membrane permeability, equal to 100 L h^{-1} m^{-2} bar^{-1}, according to the scale of the unit. A well-defined in-control dataset is critical for the success of control charts, so besides the value of membrane permeability, the report of occurrences made available by the oil refinery was also checked to identify samples in which none operating problem occurred. Therefore, the database was divided into two groups: the first one (Phase I) enclosed only samples from the first three years of monitoring in which membrane permeability was greater than 100 L h^{-1} m^{-2} bar^{-1} and for which no operating problem had occurred (n = 338); and the second one (Phase II) contained all samples from the last two years of monitoring (n = 286).

Phase I dataset was then mean-centered and scaled to unit variance according to Eqn.2 and used to develop a PCA model from which the two complementary statistics Hotelling's T^2 and Q were derived for each observation. T^2 and Q statistics were calculated according to Eqn. 21 and Eqn. 22, respectively (FERRER, 2007). From T^2 and Q statistics, the respective multivariate control charts were built, assuming as control limits the respective 95% percentiles.

Phase II dataset was also mean-centered and scaled, but now considering the mean and the sd values of the Phase I dataset. Then, the data was projected onto the developed PCA model. The same way, T^2 and Q statistics were calculated for each observation (Eqn. 21 and Eqn. 22, respectively) and plotted on the multivariate control charts to be checked against the control limits. For the observations detected to be beyond the control limits on the Q control chart, a diagnostic approach based on contribution plots was performed. The contribution of each variable *j* for each atypical observation *i* was calculated according to Eqn. 23 (FERRER, 2007).

To confirm the role played by the variables pointed as the major factors that caused operating faults by the contribution plots, boxplots were plotted and the nonparametric Wilcoxon-Mann-Whitney statistical test (MANN; WHITNEY, 1947) was applied at a significance level of 1% to compare the in-control data with the data identified as out-of-control.

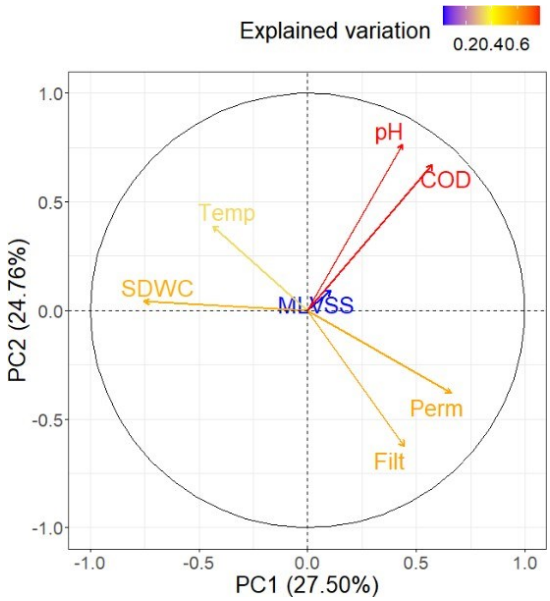
For computing these methodologic steps, the factoextra (KASSAMBARA; MUNDT, 2020), FactoMineR (LÊ *et al.*, 2008), rstatix (KASSAMBARA, 2021), scales (WICKHAM, 2020), ggplot2 (WICKHAM, 2016) and ggpubr (KASSAMBARA, 2020) R packages were used.

3 RESULTS AND DISCUSSION

3.1 Identification of the most influential variables on membrane permeability on MBR

A PCA model was developed from the final pilot-scale MBR monitoring database (n = 728). The first two PC explained over 50% of the data variation and for this first moment, in which we attempted to explore only the strongest relations between the variables and the observations, they were the only ones selected since the interest is in looking at the main variation and per definition, the first components provide information on that (BRO; SMILDE, 2014). Figure 22 displays the loadings of the variables. From this plot, we can infer how much of the variation of the original variables PC1 and PC2 explain (indicated by the arrow size and color scale) (ERIKSSON *et al.*, 2013) and how the variables are correlated to each other (indicated by the angles between the arrows: 90° for uncorrelated, 0° for complete positively correlated, and 180° for complete negatively correlated variables) (NAESSENS *et al.*, 2017).

Figure 22 - Loading plot indicating the correlation between the variables and their explained variation by the first two PC.



Perm: membrane permeability; Filt: sludge filterability; Temp: temperature.

All variables, except MLVSS, had a reasonable contribution to PC1 and PC2. COD, pH and MLVSS presented an approximated angle of 90° with membrane permeability, meaning weak correlation between these variables and membrane permeability. Conversely, sludge filterability and membrane permeability had a positive correlation while SDWC and temperature had a negative correlation with membrane permeability. Thus, the variables that

influence the most on membrane permeability are sludge filterability, temperature and SDWC. Similar results were observed by Jacquin *et al*. (2018), who applied PCA to evaluate the relations between dissolved organic matter, active biomass concentration, operating conditions, and membrane fouling. The authors also observed that MLVSS was badly represented by the PCA model and since it was not correlated with heterotrophic bacteria concentration, they concluded that MLVSS was not appropriate to quantify active biomass. Furthermore, their results also indicated temperature as a major factor on membrane fouling, together with SRT.

The strong impact of temperature on membrane permeability may be explained by its effect on the microbial community and on its metabolism. EPS and SMP are microbial metabolites secreted by cells or generated during cellular lysis that significantly influence fouling on membrane surface (AMARAL *et al.*, 2015; MENG *et al.*, 2017). Yu *et al*. (2017) applied PCA to assess membrane fouling on a MBR treating wastewater from the production of antibiotics in order to identify the major foulants and found out that EPS concentration was the primary factor affecting membrane fouling. Temperature in turn affects the microbial community development and therefore impacts on EPS and SMP release and concentration (DING *et al.*, 2020; GIL *et al.*, 2010). Gao *et al*. (2012) investigated the effects of temperature (ranging from 37 to 55 °C) on membrane fouling in MBR and found that EPS, SMP and colloidal particles content increased with an increase in temperature, since it accelerates the metabolic activity of microbes, which lead to an increasing on the extent of membrane fouling. Although in the present study temperature values were slightly lower (ranging from 22 to 47 °C), the same phenomena could explain why higher temperature values contributed to lower membrane permeability values. Moreover, Jacquin *et al*. (2018) demonstrated through their PCA model that temperature strongly influenced on the microbial activity and consequently in the production of SMP, with higher temperature leading to higher SMP release. On the other hand, since the MBR presents a high solids content (median mixed liquor suspended solids concentration of 7,800 mg L^{-1} and maximum of 18,650 mg L^{-1}), temperature can also influence on sludge viscosity. Baroutian *et al*. (2013) and Cheng and Li (2015) assessed the effects of operating temperature on the rheological behavior of sludge with high solid content (total solids content from 4 to 10% and from 7 to 15%, respectively). The first authors concluded that the yield stress decreased exponentially when temperature increased from 25 to 55 °C and the latter found that sludge viscosity decreased by up to 65% when temperature rose from 9 to 55 °C. Studies also show that less viscous sludges are associated with less severe membrane fouling (AZAMI; SARRAFZADEH; MEHRNIA, 2011; KOMESLI; GÖKÇAY, 2014). Therefore, it

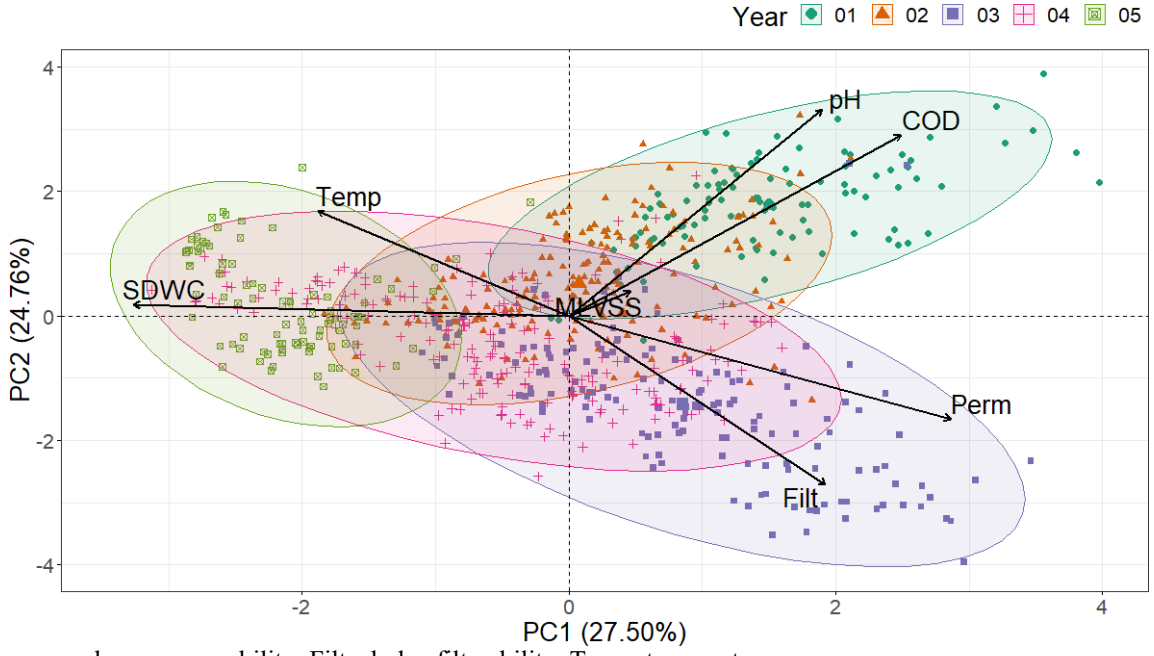
appears that for the MBR studied here, the effect of temperature on the biomass metabolism was more relevant than its effect on sludge viscosity, contributing to membrane fouling. This fact may be related to the high concentration of solids in the system, which has a greater effect on sludge viscosity than temperature (JIANG *et al.*, 2014). Besides, the MBR was operating under a high C/N ratio (20.0 ± 11.1), which also contributes to a higher production of EPS, as discussed by Feng *et al*. (2012) and Miqueleto *et al*. (2010). Unfortunately, sludge viscosity was not monitored and there was not enough data on EPS and SMP on the database provided by the oil refinery for us to further explore these relations in the assessed MBR.

Sludge filterability is an important parameter to evaluate sludge properties (CAI *et al.*, 2019) and thus it plays an important role as an indicative of membrane fouling in MBR, explaining the strong positive correlation observed, since high values of sludge filterability indicate good quality of the sludge, low tendency of fouling occurrence and therefore lead to high values of membrane permeability. Indeed, Alkmim *et al*. (2015) demonstrated that sludge filterability is directly related to membrane fouling potential and can be used as a tool for monitoring the fouling process in MBR. The strong negative correlation between membrane permeability and SDWC can also be easily justified. Chemical cleaning has been proven to be one of the most efficient ways to revert membrane fouling and to recover permeate flux (CHENG *et al.*, 2020; MENG *et al.*, 2017). Thus, if the membrane spends too much time without being cleaned, it is expected to favor fouling occurrence and severity. However, although chemical cleaning is essential to control fouling and avoid expressive reduction of permeate flux, it also causes damage to the membrane structure and thus it is directly related to the membrane lifetime and, consequently, to the operational cost of the process. Therefore, chemical cleanings must be carried out in a really well-planned manner and that is why the application of tools that can support this decision-making is so important.

To explore the relations between the observations, the scores were plotted together with the loadings on a biplot Figure 23. The observations were grouped by year of monitoring (with 95% confidence ellipses) so we can assess how the behavior of the system changes throughout monitoring time. In the first two years, samples presented a trend to higher values of pH and COD of the feed, in the third year they tended to higher values of sludge filterability and membrane permeability and finally in the last two years the observations tended to higher values of SDWC and temperature. To illustrate this, Figure 24 shows boxplots of SDWC, membrane permeability and COD. The graphs also indicate if there was significant difference between the

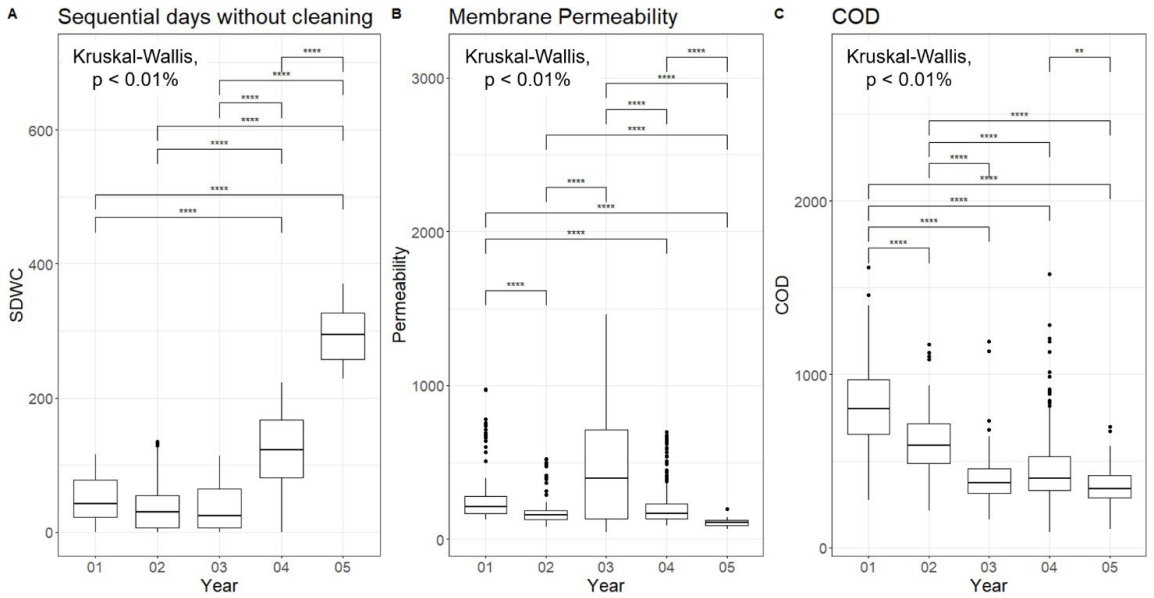
medians of each year according to the Kruskal-Wallis statistical test (KRUSKAL; WALLIS, 1952) (p-values smaller than 2.2E-16 for all variables) followed by the multiple comparison test of Dunn (ZAR, 1999) at a significance level of 1% (the pairs marked with an asterisk are significant different. The p-values are presented on Appendix D).

Figure 23 - System behavior over the years: relations between observations and variables.



Perm: membrane permeability; Filt: sludge filterability; Temp: temperature.

Figure 24 - Boxplots and nonparametric statistical tests of Kruskal-Wallis and Dunn for a) sequential days without cleaning; b) membrane permeability; and c) COD for all monitored years.



*Significant difference according to multiple comparison test of Dunn. The more asterisks, the smaller are the p-value.

The boxplots and the nonparametric statistical tests confirm what was detected by the PCA model. Indeed, in the first two years the median COD of the feed was higher than in the other years. These high values combined with reduced values of membrane permeability induced improvements in the pretreatment stages, which explains the lower values observed for COD of the feed in the subsequent years and contributes to the higher sludge filterability and membrane permeability values observed in the third year. Besides, the higher values of membrane permeability in third year can be related to low values of temperature and SDWC, and to high values of sludge filterability (Figure 23), which indicates good quality of the sludge and thus low tendency of fouling occurrence. The median number of SDWC is also proven to be greater in the last two years than in the first ones. In the last two years, decreasing the frequency of membrane chemical cleaning was investigated as a strategy for improving pollutants removal and fouling control on the system. The hypothesis was based on the concept of dynamic membranes (DM). DM are formed by the deposition of suspended particles on the original membrane surface when filtering a solution containing suspended particles through it (SALEEM; LAVAGNOLO; SPAGNI, 2018) and they have been widely studied for fouling mitigation in MBR (YANG *et al.*, 2020). Therefore, the expected was that the lower cleaning frequency would lead to the formation of a DM on the membrane surface, which would reduce the effective pore size and consequently increase the membrane retention efficiency and minimize the fouling. However, the tested hypothesis has not been proven, possibly because the DM was never formed or it was formed but got too dense (MOHAN; NAGALAKSHMI, 2020). Thus, the longer period without cleaning the membrane actually led to the lower membrane permeability values observed in the last two years of monitoring. Furthermore, in these last two years (and specially in the last one) the membrane permeability was generally low (below 150 $Lh^{-1}m^{-2}bar^{-1}$), characterizing an unstable operation, since the low membrane permeability decreases the process productivity and the fouling on the membrane surface reduces the treated effluent quality. In the first three years of monitoring, the operation of the MBR was more stable, with satisfactorily high membrane permeability values, indicating no or mild occurrence of membrane fouling (Figure 23).

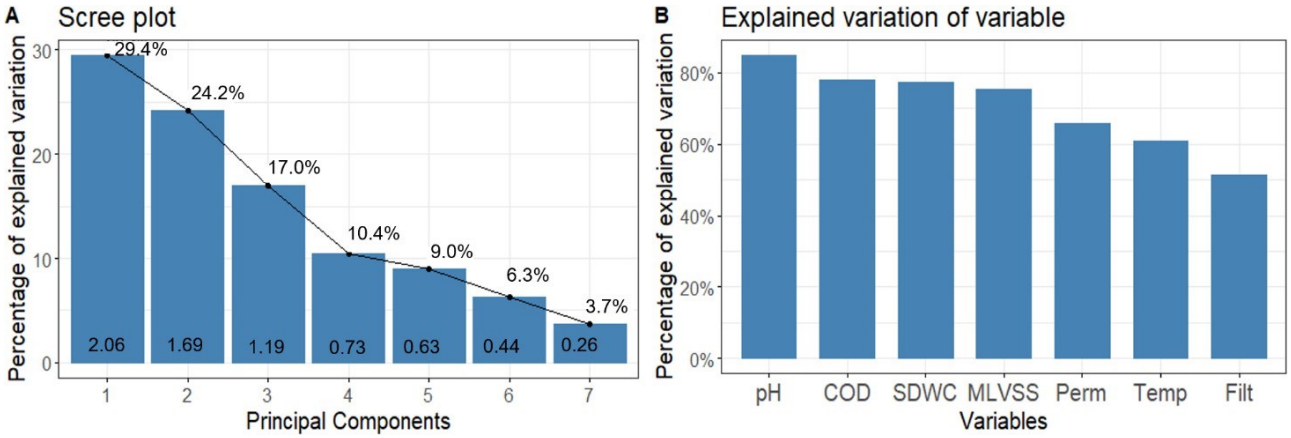
This demonstrates how PCA is effective in identifying patterns and expressing relations between different variables and observations. Therefore, its application for monitoring MBR wastewater treatment systems is of great interest, since it can reveal complex patterns of the system behavior.

3.2 PCA as a predictive model

Besides identifying patterns, PCA can also be used as a predictive model. Several papers in various distinct areas have been successfully applying PCA for this purpose, as in the food industry (ARSALANE *et al.*, 2018), metallurgical processes (HE; ZHANG, 2018) and even education (XU; LIANG; WU, 2017). As discussed in the previous section, the pilot-scale MBR presented a more stable operation regarding membrane permeability in the first three years and thus this period (n = 442) was considered to evaluate PCA capacity to model MBR behavior. As we have been working with offline data, the model was computed with a portion of the selected dataset (Group 01: n = 237) whereas the other portion was left out to be predicted, representing new observations (Group 02: n = 205).

Figure 25a displays the scree plot for the PCA model developed. The first three components have eigenvalues above one, are before the elbow point at the scree plot and explain 71% of the data variation, which was considered enough to well represent the system. Although there are no well-defined criteria to determine the minimum variation that must be explained since it is heavily application and field of knowledge dependent, other works involving MBR modelling have considered percentages above 60% as acceptable. Yu *et al*. (2017) and Jacquin *et al*. (2018), for example, worked with 62% and 65% of total variation explained. Figure 25b presents the percentage of the explained variation of each variable after three PC, demonstrating that all variables were well represented (all explained variations above 50% and most above 70%) (ERIKSSON *et al.*, 2013).

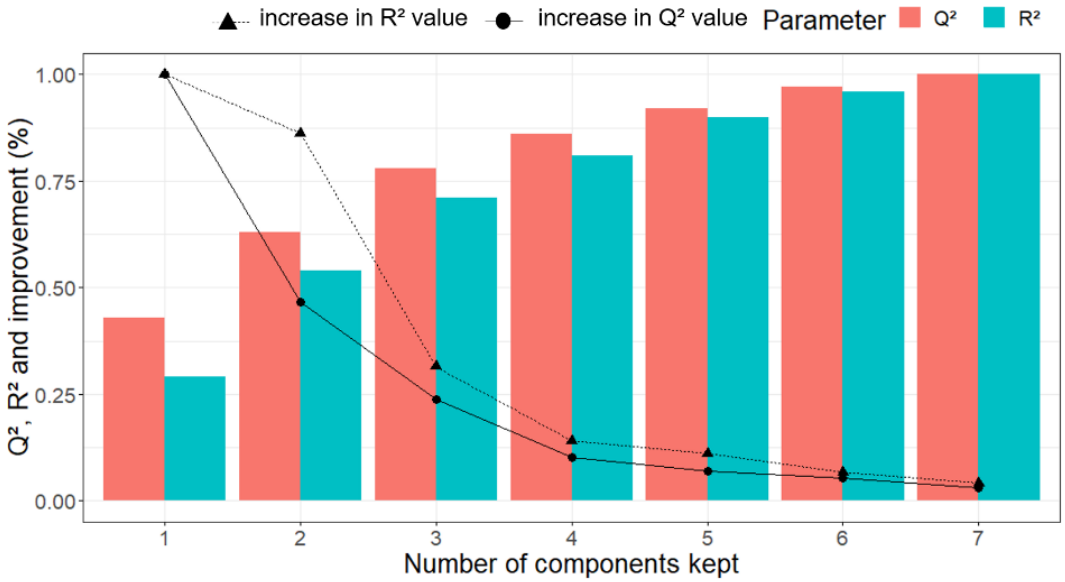
Figure 25 - a) Scree plot: eigenvalues and percentage of total variation explained by each PC and b) Explained variation of each variable after three PC.



Perm: membrane permeability; Filt: sludge filterability; Temp: temperature

Figure 26 displays the evolution of parameters R^2 and Q^2 with increasing model complexity, in an attempt to evaluate the model quality and also to ensure how many components to keep. As the complexity of the model increases, both R^2 and Q^2 approximate to one (perfectly fitting and predictive model). However, as this complexity increases, the gain in the quality of the model decreases. Adding the second PC to the model, for example, increases R^2 in 86% and Q^2 in 47%, an expressive gain. Adding the fifth PC, though, only increases R^2 in 11% and Q^2 in 7%. Therefore, it is not meaningful to keep more than four components, since both the degree of fitness and the predictive ability do not increase enough to justify the higher complexity of the model. Considering that PCA main goal is precisely to reduce data dimensionality, the fewer components kept, the better, as long as the loss of information is acceptable. Nevertheless, after three PC the model performance was already satisfactory, with both R^2 (0.71) and Q^2 (0.78) above 0.70 (according to Eriksson *et al*. (2013), generally values above 0.50 are regarded as good). Thus, in accordance with the other methods, it was concluded that the first three PC should be kept.

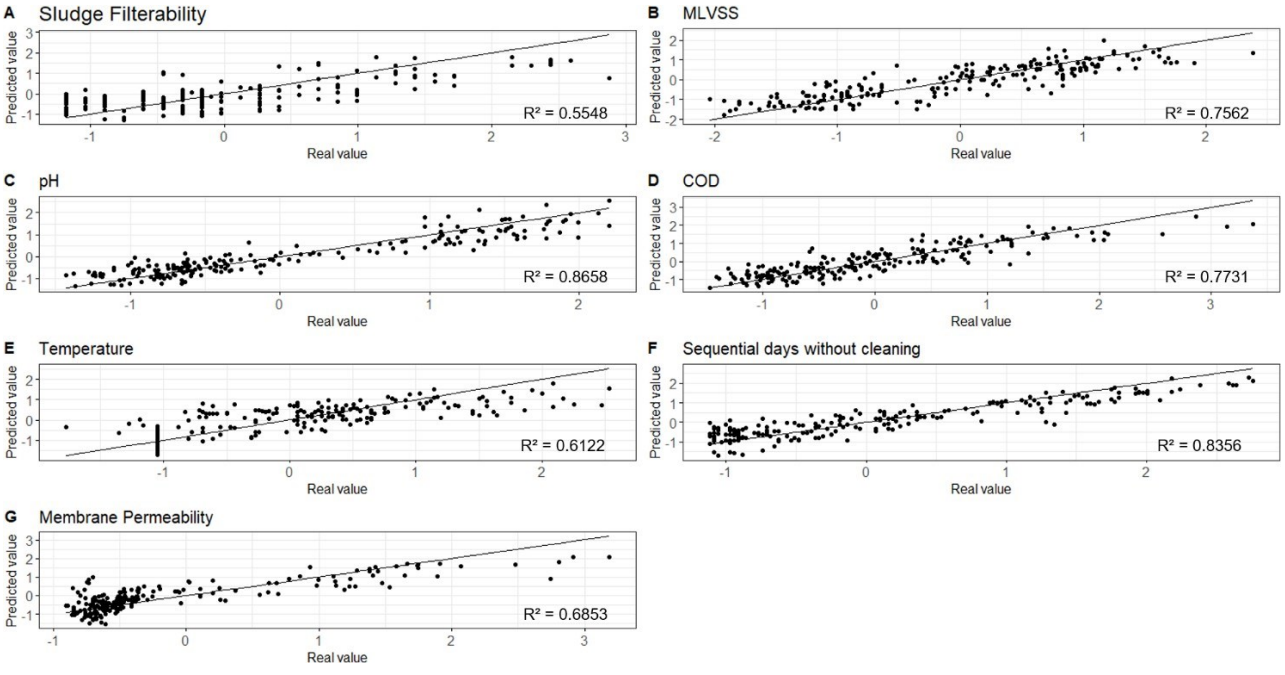
Figure 26 - Q^2 and R^2 values of PCA model and their increase for different numbers of components kept.



The PCA model developed from the data in Group 01 subset after three PC was then used to predict the behavior of the Group 02 subset. The correlations between the real (measured) and the predicted (by the model) values are shown in Figure 27. As expected, the provisions were adequately assertive, with R^2 values near or over 0.80 for four variables (pH, SDWC, COD and MLVSS) and above 0.60 for two variables (temperature and membrane permeability), indicating that the error between the values predicted by the model and the actual measured

values was low. This demonstrates that the PCA model was effective in predicting the behavior of the MBR. The only variable poorly predicted was sludge filterability (0.55). Many studies have demonstrated the application of filterability tests to monitor membrane fouling on MBR and they have employed different methods, such as Capillarity Suction Time (CST), Filter Test (FT) (used in the present study), Sludge Filtration Index (SFI) (THIEMIG, 2012), Time to Filter (TTF) (APHA; AWWA; WEF, 2012), and Delft Filtration Characterization (DFC) (EVENBLIJ *et al.*, 2005). Alkmim *et al.* (2015) compared the methods TTF, FT and SFI regarding their capability to sense sludge quality variation and their reproducibility and concluded that TTF method was the most effective. Therefore, it is possible that sludge filterability could be better predicted by the model if it had been measured by a different method, like TTF. As we have been working with data provided by the oil refinery, though, we could not test this hypothesis.

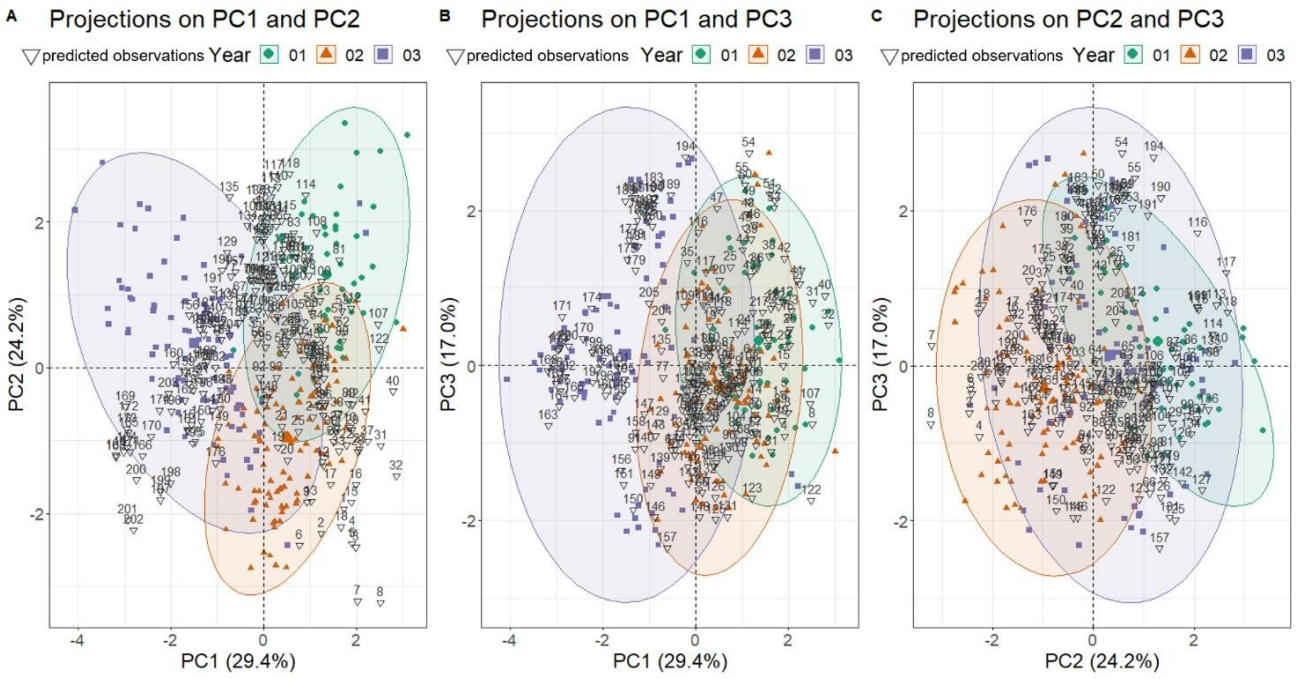
Figure 27 - Correlation plots between the real and the predicted values of each variable for Group 02 subset: a) sludge filterability; b) MLVSS; c) pH; d) COD of the feed; e) temperature; f) sequential days without cleaning; and g) membrane permeability.



The predicted observations were then projected onto the PCA reduced dimensional space. This projection is attention-grabbing because it allows us to evaluate if a certain sample is within the expected behavior or if it deviates from it. Figure 28 displays the score plot with the projections of the predicted samples for all three components kept. In general, the predicted observations were within the confidence ellipses (95%), indicating that these samples behaved as expected according to the PCA model built from Group 01 subset. However, some observations like 7, 8, 201 and 202 did not fit as expected, which indicates atypical behavior of the system in these

samples. Reviewing the original data, we found out that in samples 7 and 8, in fact, the values of COD of the feed were unusually high and checking the report of occurrences we found out that on those days the flotation equipment (one of the pretreatment steps) was not working properly, which justifies the higher values of COD observed. Observations 201 and 202, in turn, deviate from the group due to their higher temperature values (32 ºC, whereas the mean value for this subset was 28 °C). There is no occurrence reported for these samples that justify the high temperature and so we believe that it is most likely due to the discharge of a hot effluent that was fed to the MBR and raised its temperature. The PCA model was able to identify and reveal this atypical behavior and thus could be used to visually represent the process state and to detect extreme samples, as also stated by other authors (MAERE *et al.*, 2012; NAESSENS *et al.*, 2017).

Figure 28 - Projections of the predicted observations onto the PCA model: a) PC1 and PC2; b) PC1 and PC3; and c) PC2 and PC3.



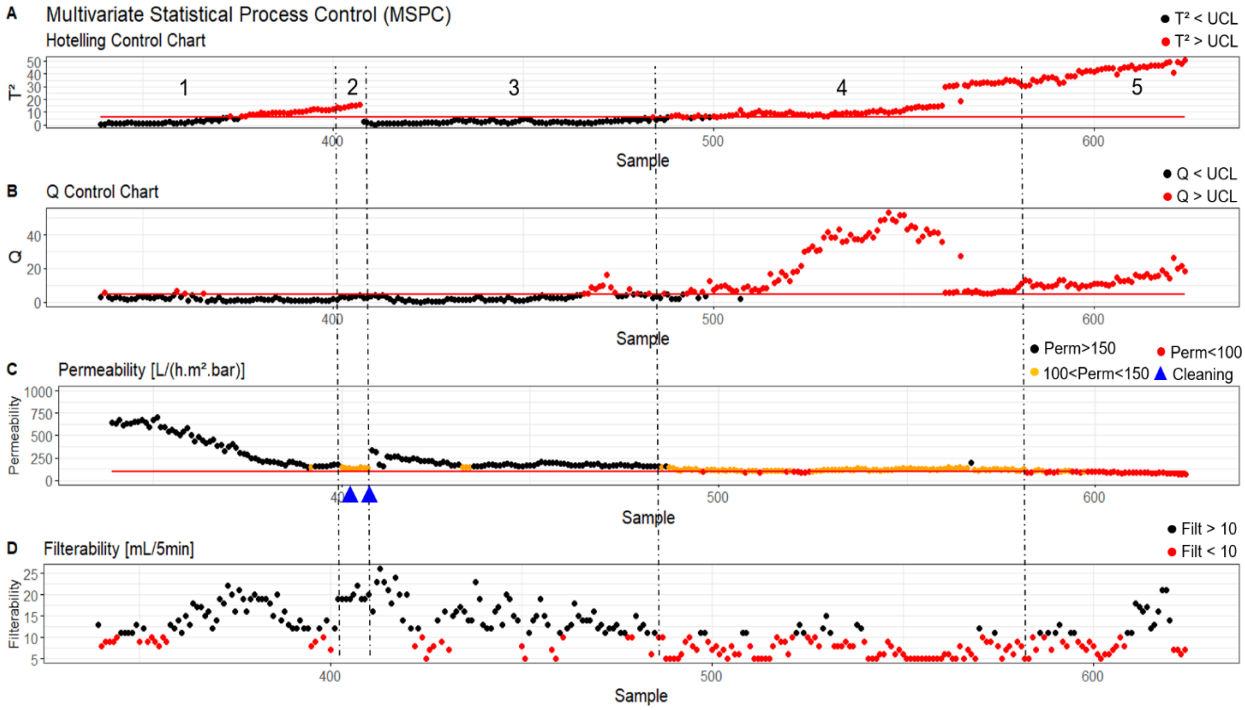
These results are of great interest because, despite being generated from offline monitoring data, they demonstrate the potential of PCA to be used for online monitoring of MBR wastewater treatment systems. Due to its high predictive ability, new samples can be predicted by the model and projected onto the reduced dimensional space to evaluate the if the system is performing as its expected to do or if there is some operating problem. By promoting an overview of the data in real time, the PCA model could support more assertive decision-making, improving the monitoring and control of the process.

3.3 Membrane fouling control on MBR systems

3.3.1 Detecting membrane fouling occurrence

In order to monitor the pilot-scale MBR performance along time and detect any unusual event that could be related to the occurrence of fouling, MSPC was applied to the database. The observations from the first three years with values of membrane permeability above 100 L h^{-1} m^{-2} bar^{-1} and no registered occurrences were considered as in-control operation and thus used in Phase I (n = 338) and the observations from the last two years (both in and out-of-control) were used for test in Phase II (n = 286). The PCA model used as projection method kept three PC, based on both Kaiser criterion and fraction of variation explained (75% of the total variation explained and all of the variables more than 50% explained). Figure 29 presents the Hotelling's T^2 and Q multivariate control charts, as well as the values of membrane permeability and sludge filterability of each Phase II sample.

Figure 29 - Detection of points with extreme low values of membrane permeability during MBR operation: a) Hotelling's T^2 control chart; b) Q control chart; c) membrane permeability; and d) sludge filterability.



UCL: Upper Limit Control; Perm: permeability; Filt: filterability

Aiming to make this discussion clearer, Figure 29 has been divided into five regions according to membrane permeability values. In region 1, the values of membrane permeability were generally greater than 150 L h^{-1} m^{-2} bar^{-1} and thus the MBR operation was stable (Figure 29c).

Despite that, both statistics pointed some observations as out-of-control (Figure 29b), configuring as false alarms. However, it can be noted that T^2-statistic false alarms are following a reduction in membrane permeability values until they reach region 2, in which membrane permeability values are between 100 and 150 L h^{-1} m^{-2} bar^{-1}. In that region, despite not being out-of-control, the MBR operation is alarming and should be closely watched. T^2 control chart was able to detect this tendency of reduction and to alert about the occurrence of low values of membrane permeability in this region. At the end of this region, the membrane was submitted to chemical cleanings that managed to increase membrane permeability and, as a consequence, T^2-statistic values decreased. In region 3, the values of membrane permeability were again generally greater than 150 L h^{-1} m^{-2} bar^{-1} and it is noticeable a similar pattern, since Q-statistic pointed some alarms at the end of this region following a decrease on membrane permeability values. In this region, the Q control chart was the one able to detect and alert about the membrane permeability reduction. This confirms that, as the two multivariate control charts differ in their conceptual meaning (KOURTI; MACGREGOR, 1996), they are complementary statistics that allow the overview of the goodness of the process (KOURTI, 2005) and thus they must be assessed in a combined manner since they can detect different anomalies. In region 4, the values of membrane permeability are generally between 100 and 150 $Lh^{-1}m^{-2}bar^{-1}$ (alarming operation). In this region, both statistics pointed the alarming situation out and could had been used as supportive tools for establishing fouling control strategies in an attempt to prevent the system from going out-of-control (membrane permeability under 100 L h^{-1} m^{-2} bar^{-1}). Besides, in the alarming operation, sludge filterability (Figure 29d) could had been used simultaneously with the control charts to support the definition of fouling control strategies and to avoid unnecessary chemical cleaning. Sludge filterability values below 10 mL $5min^{-1}$ indicates poor quality of the sludge and thus high propensity to fouling. In that case, substances that help to control fouling through coagulation/flocculation of the sludge, called permeability improvers, can be an effective option. Many studies have proved that their use increases membrane permeability (AMARAL *et al.*, 2015; ODRIOZOLA *et al.*, 2021). Therefore, high values on the control charts combined with high values of sludge filterability indicate the need of membrane chemical cleanings, but high values on the control charts combined with low values of sludge filterability can be an alert to dose permeability improvers in order to prevent membrane permeability from continuing to reduce. If the improver is not used in advance and membrane permeability reaches extremely low values, only chemical cleaning would be effective, highlighting the importance of the alarm, since membrane cleaning should be avoided

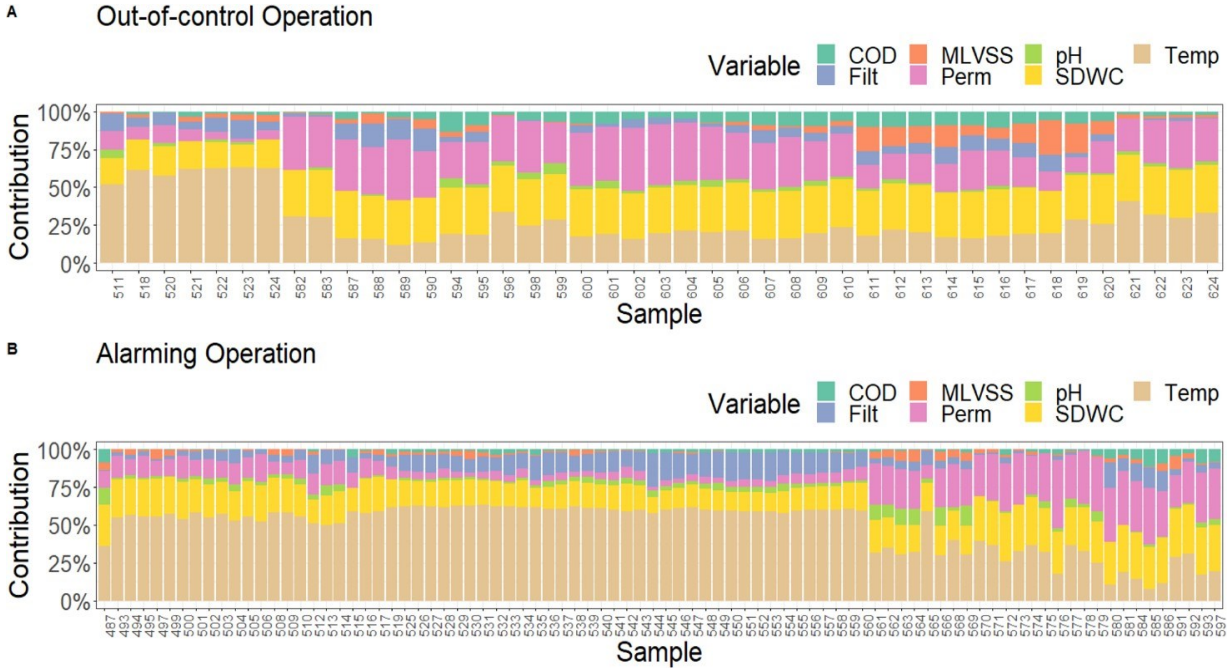
to ensure a longer lifetime for the membrane. As no fouling control measure was taken though, in region 5 the MBR operation went out-of-control, with membrane permeability values mostly below 100 $Lh^{-1}m^{-2}bar^{-1}$. Both statistics were able to detect all out-of-control observations in this region and again could had been used as tools to indicate the need of chemical cleaning.

In general, T^2 and Q control charts were able to detect 100% and 96%, respectively, of the out-of-control observations, proving their capability to identify irregular conditions of operation, and 91% and 86%, respectively, of the alarming observations, which is overwhelming from a control point of view, since it allows us to act preventively on the system. Both statistics also had a low percentage of false alarms (9% in T^2 chart and 6% in Q chart), considering that these errors are inevitable due to the very probability of the statistics (5%). Therefore, the multivariate control charts have proven to be a handy tool for the monitoring and control of membrane fouling as they can detect membrane permeability reductions and thus can be used to support the definition of efficient fouling mitigation strategies, guiding e.g. when to dose permeability improvers and/or perform chemical cleanings. The application of MSPC enables thus that chemical cleanings are performed at the most appropriate times, avoiding unnecessary costs, preserving the membrane lifetime and increasing the efficiency of the process.

3.3.2 Diagnosing membrane fouling occurrence

As the Q control chart presented a lower percentage of false alarms and all operating failures detected by this chart were also detected by the T^2 control chart, contribution plots based on the Q-statistics were created to assess what caused the MBR operation to go to an alarming or out-of-control state during the monitoring (Figure 30).

Figure 30 - Contribution plot based on the Q-statistic for a) out-of-control observations; and b) alarming observations.



Filt: sludge filterability; Perm: membrane permeability; Temp: temperature

It is clear from the plots that the variables with greater contributions to the exceeding values of Q-statistics were membrane permeability, temperature and SDWC. The high contribution of membrane permeability was completely expected, since the out-of-control condition regards precisely this variable. As for temperature and SDWC, this indicates that these two variables influenced the most on membrane permeability reduction, endorsing the results obtained with the PCA model. This also confirms the critical role that chemical cleaning plays on membrane fouling control, as well as demonstrates the importance of preventing the temperature of the system from increasing, since long intervals without chemical cleaning and high values of temperature were the most common causes of membrane fouling.

The definition of an effective chemical cleaning strategy has been extensively pursued in the last years and several papers have been published on this subject. Back in 2008, Brepols *et al*. (2008) were already evaluating different cleaning agents and procedures for a large-scale MBR. Cheng *et al*. (2020) evaluated four cleaning protocols for a high solid content MBR (like the one studied here): gas scouring, Milli-Q backwashing and rinsing, NaClO backwashing and desorption, and citric acid backwashing and desorption. The authors concluded that periodical NaClO cleaning and biogas scouring were recommended at low and high filtration to relaxation ratios, respectively. Guan *et al*. (2018) compared the cleaning effects of three typical chemical

cleaning reagents, i.e. NaOH, NaClO, and sodium dodecyl sulfate and all of them promoted membrane permeability recoveries over 75%. It is important to note that, besides the cleaning agent, defining the cleaning frequency is equally critical, especially because chemical cleaning also causes damage to the membrane structure and therefore reduces its lifetime, as discussed before. T^2 and Q multivariate control charts have proven to be useful in this regard since they can be used to guide when it is really necessary to proceed with the chemical cleaning of the membrane.

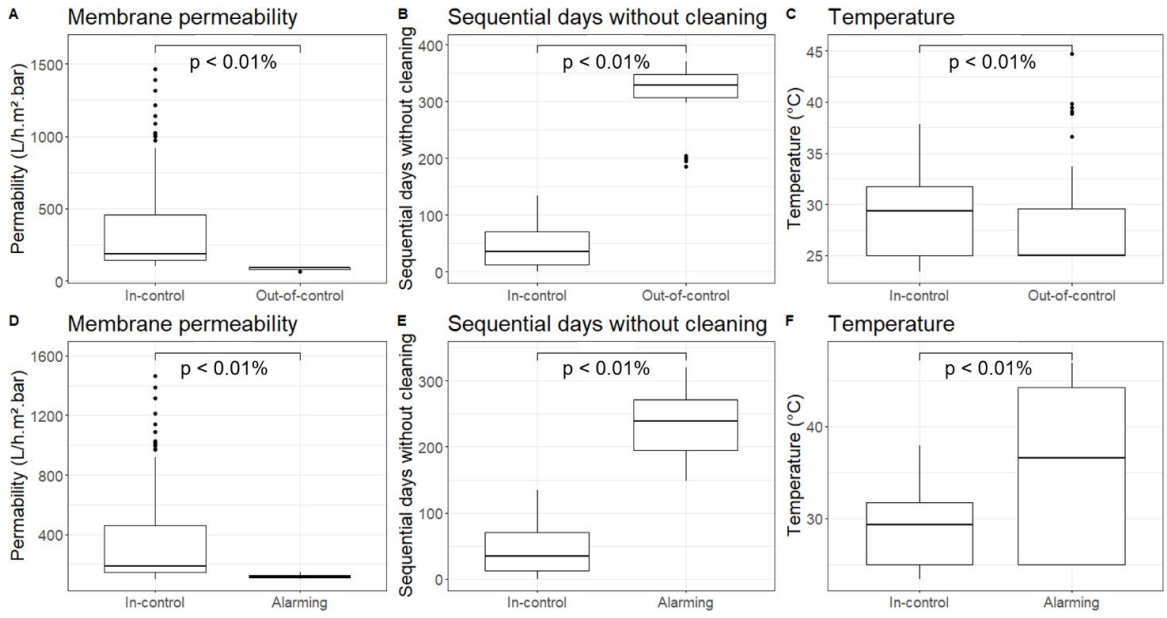
Although the absolute variation of temperature is not large (from 22 to 47 °C), its impact on the system conditions is considerable and can prejudice MBR performance. Establishing an ideal temperature value is intricate though, since its influence on the occurrence of fouling is closely related to biomass development and therefore depends on the specific microbial community and the other conditions of the environment of each biological system (GAO *et al.*, 2013). Analyzing the data from the MBR in-control operation, temperature was mostly below 30 °C, with a maximum value of 38 °C, whereas in the out-of-control operation temperature reached out 45 °C. Furthermore, as the effluent reaches the MBR after undergoing a series of industrial operations, controlling its temperature can be challenging. Adopting a proper planning for the discharge of effluents generated at high temperatures is, therefore, essential. In addition, although the refinery already has an equalization tank, its design may be inadequate and its revision may be another good strategy to control the temperature of the MBR feed. Besides controlling temperature, adopting an appropriate C/N ratio can also contribute to a better control of the microbial community and metabolism. A lower C/N ratio could favor nitrifier bacteria, which could decrease SMP and EPS release and significantly mitigate fouling in the MBR (SEPEHRI; SARRAFZADEH, 2018).

From the plot for the alarming operation (Figure 30b), it is also possible to observe a noteworthy contribution of sludge filterability between samples 525 and 560, period during which this variable was considerably low (median value equal to 8 mL $5min^{-1}$ and maximum value equal to 13 mL $5min^{-1}$), as can be noted from Figure 29d. This result endorses that sludge filterability can be used as a tool for membrane fouling control in MBR systems.

Finally, Figure 31 displays boxplots for membrane permeability, SDWC and temperature divided into in-control, out-of-control and alarming operation. The nonparametric Wilcoxon-Mann-Whitney statistical test (MANN; WHITNEY, 1947) at a significance level of 1% was

also applied to confirm that the medians of the variables were significantly different in the last two groups when compared to the in-control dataset (the p-values obtained are presented in Appendix E). It can be noted the clearly lower values of membrane permeability and the clearly higher values of SDWC in out-of-control and alarming operations. For temperature, in alarming operation the values are clearly higher than in the in-control operation, whereas in out-of-control operation, despite a lower median, there are some extreme observations who were responsible for the negative impact on the system. It is worth saying that the effect of these temperature extremes on biological systems is not punctual, since the sludge takes time to recover after these events. Therefore, once again the importance of a more adequate planning for the discharge of high temperature effluents becomes evident.

Figure 31 - Nonparametric statistical test of Wilcoxon-Mann-Whitney for membrane permeability, SDWC and temperature for in-control, out-of-control and alarming operation.



4 CONCLUSION

For a pilot-scale MBR applied for the treatment of an oil refinery wastewater, PCA and MSPC have proven to be suitable for monitoring the wastewater treatment system aiming membrane fouling control. PCA modelling was effective in mapping the MBR behavior and identified sludge filterability, temperature and SDWC as the variables that most influence on membrane permeability. The negative impact of temperature on membrane permeability on MBR systems is attention grabbing, since it is opposite to what is expected considering other MSP. In general, higher temperatures are related to higher permeabilities, but in MBR its effect on the biomass metabolism is predominant. The model was also applied to predict the MBR performance, with high values of R^2 and Q^2 (0.71 and 0.78, respectively). Therefore, the error between the values predicted by the model and the actual measured values was low, which demonstrates that PCA was effective in predicting the MBR behavior. Moreover, the model was able to distinguish atypical samples, enabling the detection of operating problems. T^2 and Q control charts, assessed in a combined manner, have also proven to be handy tools in the detection of membrane fouling and in the stablishment of more effective membrane fouling mitigation strategies. The multivariate control charts were able to preventively detect membrane permeability reductions and thus can be used to guide when to dose permeability improvers and/or perform chemical cleanings. Therefore, the application of MSPC enables that chemical cleanings are performed at the most appropriate times, avoiding unnecessary costs, preserving the membrane lifetime and increasing the overall efficiency of the process. Preventing the temperature of the system from increasing is also an important fouling mitigation measure, so an appropriate planning for the discharge of hot effluents should be assessed.

IV. IMPROVING AMMONIA REMOVAL

1 INTRODUCTION

In this chapter, ANN and PCA models were applied to identify the variables that contribute the most for ammonia removal on MBR and to predict its behavior, whereas MSPC modelling was used to detect and diagnose low removal conditions. The monitoring data from the pilot-scale MBR treating real oil refinery effluent was used, but now considering only the last four years of operation. All models were developed in R. ANN and PCA identified that influent COD and OG concentration, together with membrane permeability, contribute the most to lower ammonia removals, while influent ammonia concentration, temperature and SRT are the most related to greater removals. ANN modelling also effectively predicted the ammonia removal from a set of operating conditions, with $R^2 = 0.87$. From the MSPC model, Q control chart detected 100% of the operation with removals lower than 85%, which could enable to act more effectively on the system. Therefore, ANN and MSPC could be applied as tools for supporting and improving the decision-making regarding ammonia removal control, contributing to a more efficient process.

2 METHODOLOGY

2.1 MBR configuration and performance

The MBR used as case study in this work was described in details in Chapter ‘Improving membrane fouling control’, item 2.1.

2.2 Database and preliminary statistical analysis

The monitoring data provided by the oil refinery containing originally 1,131 samples was also used to assess and improve ammonia removal on MBR systems. However, only the last four years of monitoring were considered, due to the frequency of measurement of some important variables.

The variables selected were SRT, sludge filterability, MLVSS, pH, temperature, membrane permeability, influent concentrations of COD, phosphorous, ammonia, sulphides and OG, effluent concentration of COD, COD removal and ammonia removal. Alkalinity is an important variable for investigating ammonia removal, since it plays an important role in nitrification. However, it was not measured with a high enough frequency to feasible its use on the model.

Outliers were investigated using both IQR (SCHWERTMAN; OWENS; ADNAN, 2004) and Hampel identifier methods (DAVIES; GATHER, 1993) and the report of operation provided by the oil refinery along with the monitoring data was used to identify any operational problem that could justify the extreme values. It was concluded that all observations classified as outliers should be kept in the dataset though. Therefore, the final database consisted of 479 samples covering 14 variables. Table 5 presents the descriptive statistics for the final database.

Table 5 - Descriptive statistics of the 14 selected variables in the final database

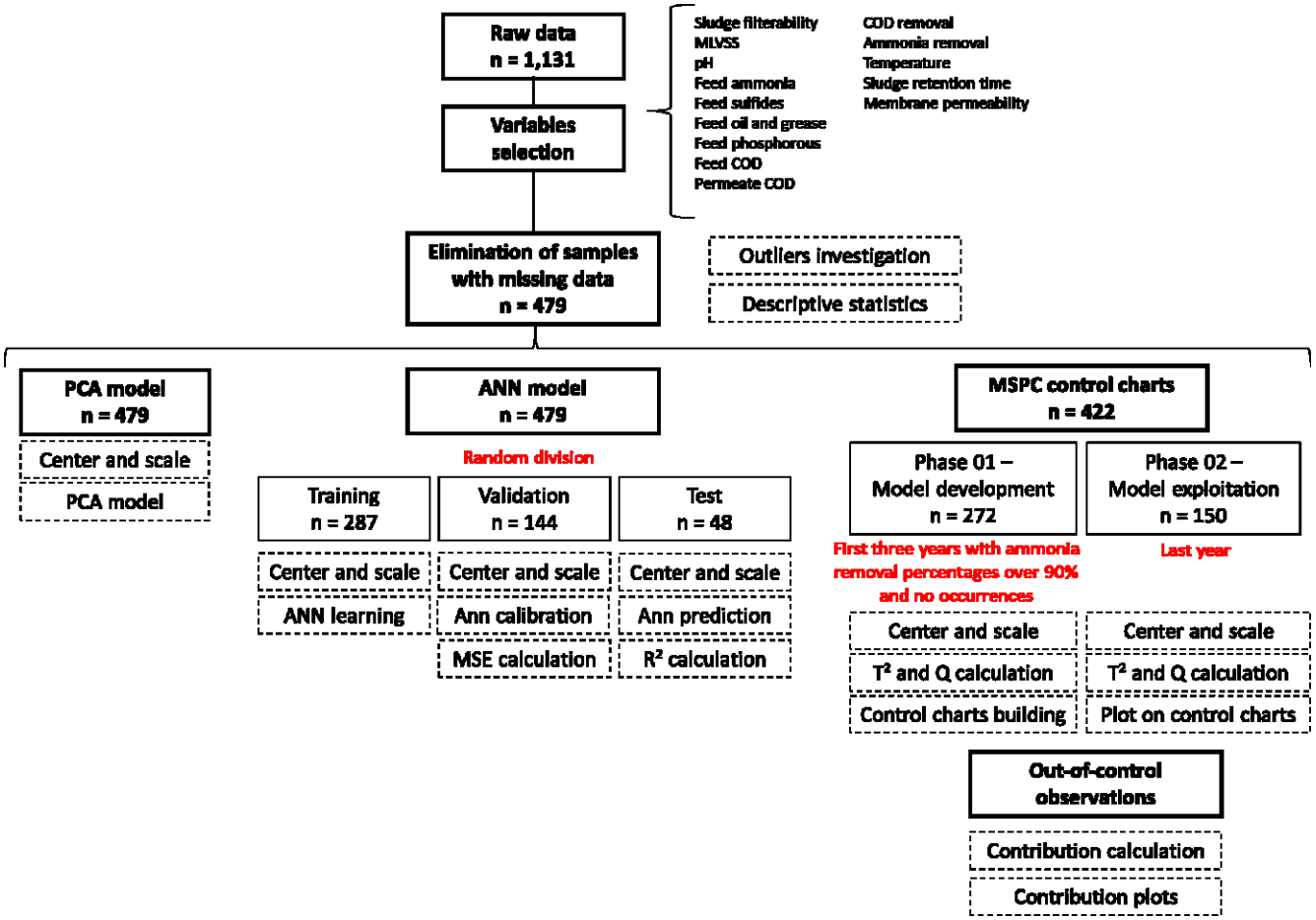
	Filt. (mL 5min^{-1})	MLVSS (mg L^{-1})	pH (-)	Influent ammonia (mg L^{-1})	Influent sulphides (mg L^{-1})	Influent OG (ppm)	Influent Phosp. (mg L^{-1})	Influent COD (mg L^{-1})	Permeate COD (mg L^{-1})	Ammonia removal (%)	COD removal (%)	Temp. (°C)	SRT (d)	Perm. (L h^{-1} m^{-2} bar^{-1})
n	479	479	479	479	479	479	479	479	479	479	479	479	479	479
Min	5.00	2150	5.61	11.2	1.00	0.60	0.01	144	17.0	15.8	19.3	23.4	30	87.9
Max	26.0	12500	10.5	82.8	21.5	48.4	1.50	1617	384	99.4	96.9	39.2	80	1463
Median	11.0	7200	7.70	29.1	6.40	11.6	0.20	563	102	94.2	81.0	29.0	30	179
Mean	11.6	7185	8.10	29.9	7.20	11.4	0.30	594	112	91.7	79.3	28.9	42	276
sd	4.44	2004	0.98	8.42	3.26	6.35	0.16	249	48.1	10.0	10.2	3.40	15	232
Variance	19.7	4016661	0.96	70.9	10.6	40.3	0.03	62188	2317	100	104	11.5	217	53783
Coef. of variation	0.38	0.28	0.12	0.28	0.45	0.56	0.61	0.42	0.43	0.11	0.13	0.12	0.36	0.84
Coef. of skewness	0.69	0.08	0.70	1.61	1.11	1.69	3.23	1.00	1.31	-4.09	-2.17	0.33	0.52	2.27

n: number of samples; min: minimum; max: maximum; sd: standard deviation; coef: coefficient. Filt: sludge filterability; MLVSS: mixed liquor volatile suspended solids; OG: oil and grease; Phosp: phosphorous; COD: chemical oxygen demand; Temp: temperature; SRT: sludge retention time; Perm: membrane permeability.

2.3 Multivariate statistical analyses

Figure 32 displays a schematic plot of the research methodology, regarding data processing. All models were developed using software R version 4.0.2 (R CORE TEAM, 2020). For data importation from Microsoft Excel® 2016, the readxl package was used (WICKHAM; BRYAN, 2019). More details are described in the following items.

Figure 32 - Schematic plot of the research methodology and data processing.



2.3.1 ANN model development

The final dataset was preprocessed through scaling and centering, according to Eqn. 2. This standardization reduces the large difference between the orders of magnitude of different variables and thus make them more comparable, giving equal importance to each one in the ANN model.

Once the data were organized and standardized, the ANN model was built in R using NeuralNet

(FRITSCH; FRAUKE; WRIGHT, 2019) and Keras (ALLAIRE; CHOLLET, 2022) packages. The variables sludge filterability, MLVSS, pH, influent concentrations of ammonia, OG, sulphides, phosphorous and COD, effluent COD, COD removal, temperature, SRT and membrane permeability were passed to the model as input variables and the variable ammonia removal was defined as the output, since the objective is to predict its value and assess its behavior. A multilayer feedforward architecture was applied and several tests were performed with different network configurations, combining different settings of activation function, number of hidden layers, number of neurons in each layer and number of epochs, in order to identify the configuration that presented the smallest MSE. Based on these tests, it was determined that the network configuration to be used for modelling the MBR would apply the rectifier with dropout activation function; contain two hidden layers, containing 12 and seven neurons, in that order; and 5,000 epochs.

The model was used for two purposes: sensitivity analysis of the MBR ammonia removal capacity to the input variables; and forecasting of the removal achieved from a set of input conditions. So, after importing the data, it was randomly divided into three sets: training, with 60% of data; validation, with 30% of data; and test, with 10% of data. Zhong *et al*. (2021) highlight at their review work the importance of proper model development and interpretation to deliver meaningful results when applying ML to ESE field. The authors stated that "more 'representative' data rather than 'big' data are more important for obtaining robust, powerful ML models". Aiming to ensure that the training set was truly representative of the actual behavior of the system and that different ammonia removal conditions were addressed thus, the division of the sets was verified and repeated until the training set covered a range of ammonia removal from 20% to 95%, so that the model could learn its behavior under different conditions. For the sensitivity analysis, the training and validation subsets were used to train and calibrate the network and for the ammonia removal prediction, the test subset was used as input.

The model performance was evaluated through the statistical parameters R^2, MAE and MSE, calculated according to Eqns. 17, 33 and 34, respectively.

2.3.2 PCA model development

PCA was performed in order to further explore the relations between the variables and the ammonia removal achieved, specially to analyze whether the correlations indicated by the ANN

model were positive or negative.

PCA model was computed using the PCA function found in the FactoMineR package (LÊ *et al.*, 2008). The data was previously centered and scaled and loading plots were built using the factoextra (KASSAMBARA; MUNDT, 2020), ggpubr (KASSAMBARA, 2020) and ggplot2 (WICKHAM, 2016) packages. The number of components to keep was determined based on Kaiser criterion (KAISER, 1960) and the scree test (CATTELL, 1966). The quality of the model was evaluated from its R^2 value.

2.3.3 MSPC model development

T^2 and Q multivariate control charts were built from a MSPC-PCA model. The number of components to keep in the PCA model was again determined based on Kaiser criterion and the scree test. The out-of-control criteria applied was a minimum of 90% of ammonia removal and therefore the dataset was split into two groups: Phase I contained only in-control (removal greater than 90%) samples from the first three years of monitoring (n = 272); and Phase II contained all samples from the last year of monitoring, both in and out-of-control (n = 150). Phase I dataset was centered and scaled (Eqn. 2) and then used to develop the PCA model from which T^2 (Eqn. 21) and Q (Eqn. 22) complementary statistics were derived for each observation. From the two statistics, the respective multivariate control charts were built assuming as UCL the respective 95% percentiles.

Next, phase II dataset was scaled and centered and projected onto the PCA model developed so that T^2 and Q statistics were calculated and plotted on the control charts to be checked against the UCL. For the observations detected to be beyond UCL on the Q-statistic control chart, contribution plots were built. For computing these methodologic steps, the scales (WICKHAM; SEIDEL, 2020), rstatix (KASSAMBARA, 2021), ggplot2 (WICKHAM, 2016) and ggpubr (KASSAMBARA, 2020) packages were used.

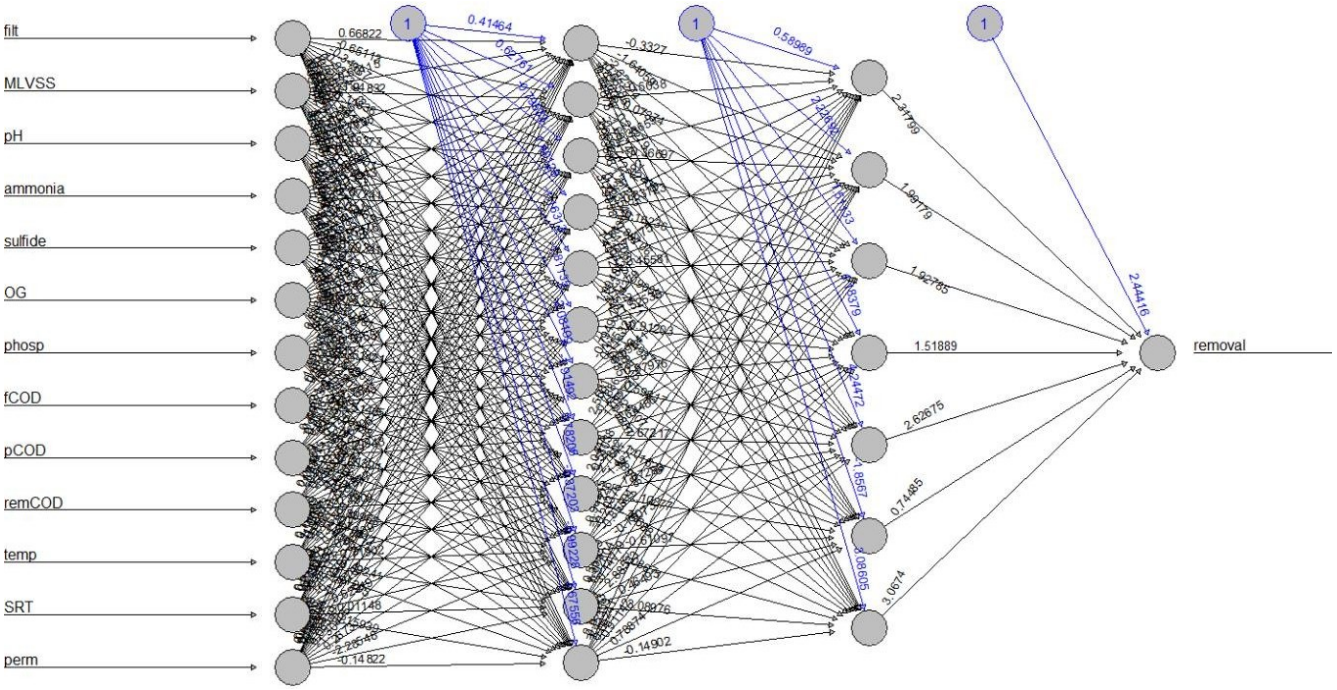
3 RESULTS AND DISCUSSION

3.1 Comprehending and predicting ammonia removal on MBR systems

3.1.1.1 Identification of the most influential variables on ammonia removal

Both ANN and PCA models developed were applied to identify the main factors that impact on MBR ammonia removal capacity. Figure 33 displays the ANN architecture, presenting its layers and synaptic weights. The black lines show the connections between each layer and the weights on each connection while the blue lines show the bias term added in each step. The model performed well modelling the MBR behavior, with reasonably low values of MAE (order of magnitude 10^{-2}) and MSE (order of magnitude 10^{-4}), as shown in Figure 34. The graphs show the learning process and evolution of the ANN model during the training and validation stages. It is possible to notice a considerable decline of both MAE and MSE, indicating that the model is adjusting to the data and learning its patterns.

Figure 33 - Artificial Neural Network developed for assessing ammonia removal by the studied MBR showing input, hidden and output layers and the respective synaptic weights



filt: sludge filterability; MLVSS: mixed liquor volatile suspended solids; ammonia: influent concentration of ammonia; sulphide: influent concentration of sulphide; OG: infleunt concentration of oil and grease; Phosp: influent concentration of phosphorous; fCOD: influent COD; pCOD: effluent COD; remCOD: removal of COD; temp: temperature; SRT: sludge retention time; perm: membrane permeability.

Figure 34 - ANN model Mean Absolute Error (MAE) and Mean Squared Error (MSE)

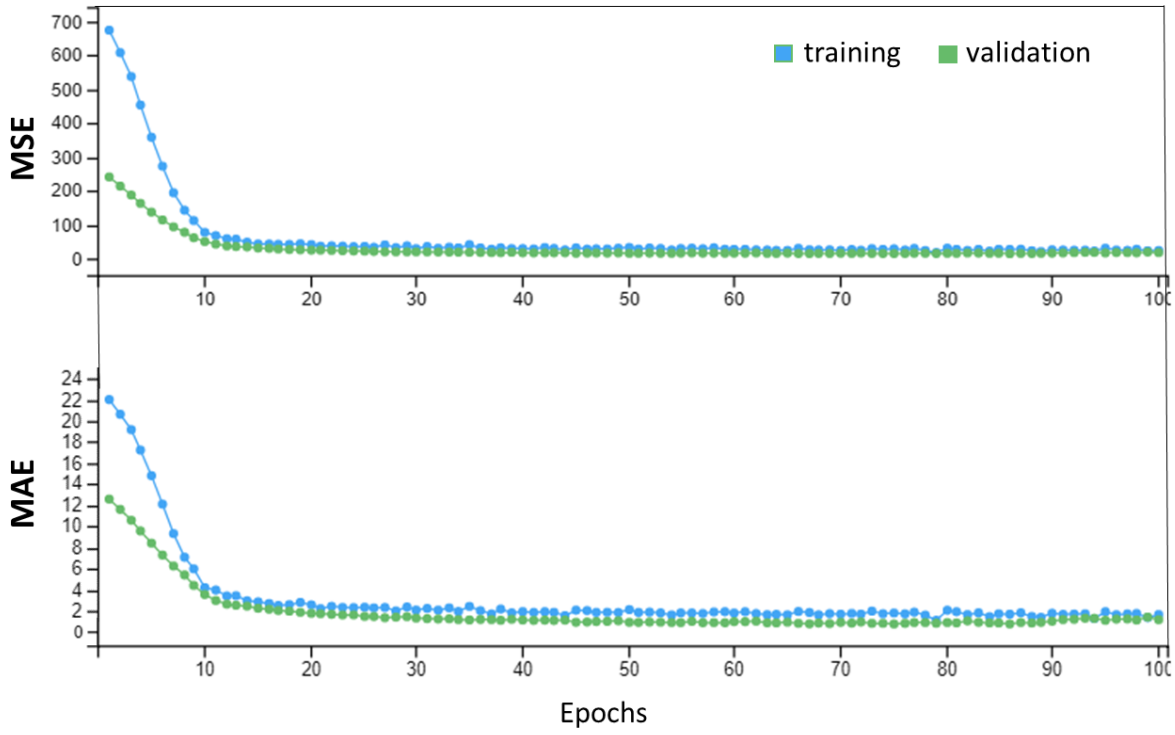


Figure 35a displays the scree plot for the PCA model. According to Kaiser criterion, PC with eigenvalues greater than one should be kept (KAISER, 1960) and according to the scree test one should keep the PC before the elbow point in the scree plot (CATTELL, 1966). The graph shows that the first three PC meet both criteria. Besides, it is noticeable that adding more PC after the third one is not meaningful, since R^2 values do not increase enough to justify the higher complexity of the model, as shown in Figure 35b. For simplicity purposes, only the first ten PC are displayed on the graphs. Considering that PCA main goal is precisely to reduce data dimensionality, the fewer components kept, the better, as long as the loss of information is acceptable. Nevertheless, after three PC the model explained 55% of the data variation, which is enough to explore the strongest relations between the different variables (BRO; SMILDE, 2014), and presented satisfactory performance, with R^2 equal to 0.772, which is regarded as good, according to Eriksson *et al*. (2013).

Figure 35 - PCA model a) Scree plot: eigenvalues and percentage of total variation explained of the first 10 principal components and b) R^2 values and its increase for different numbers of components kept

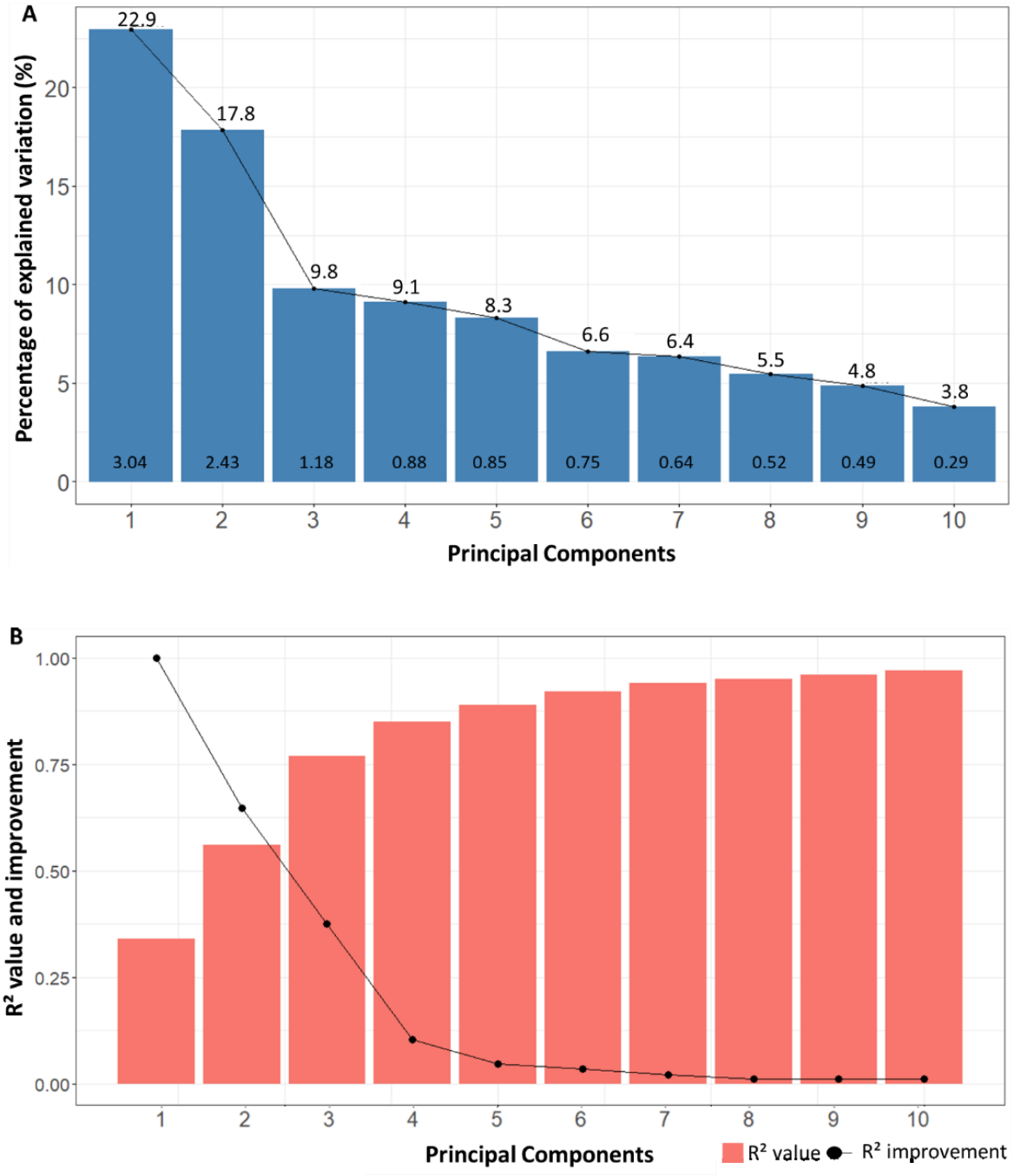
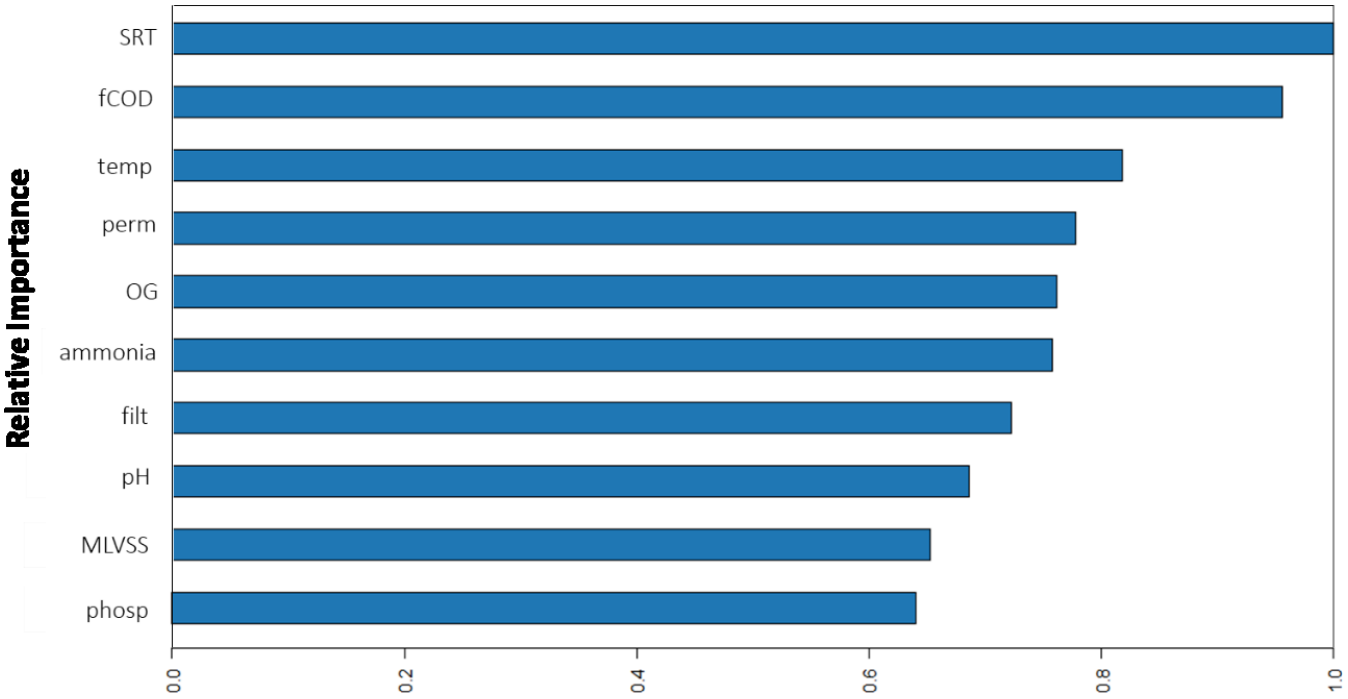


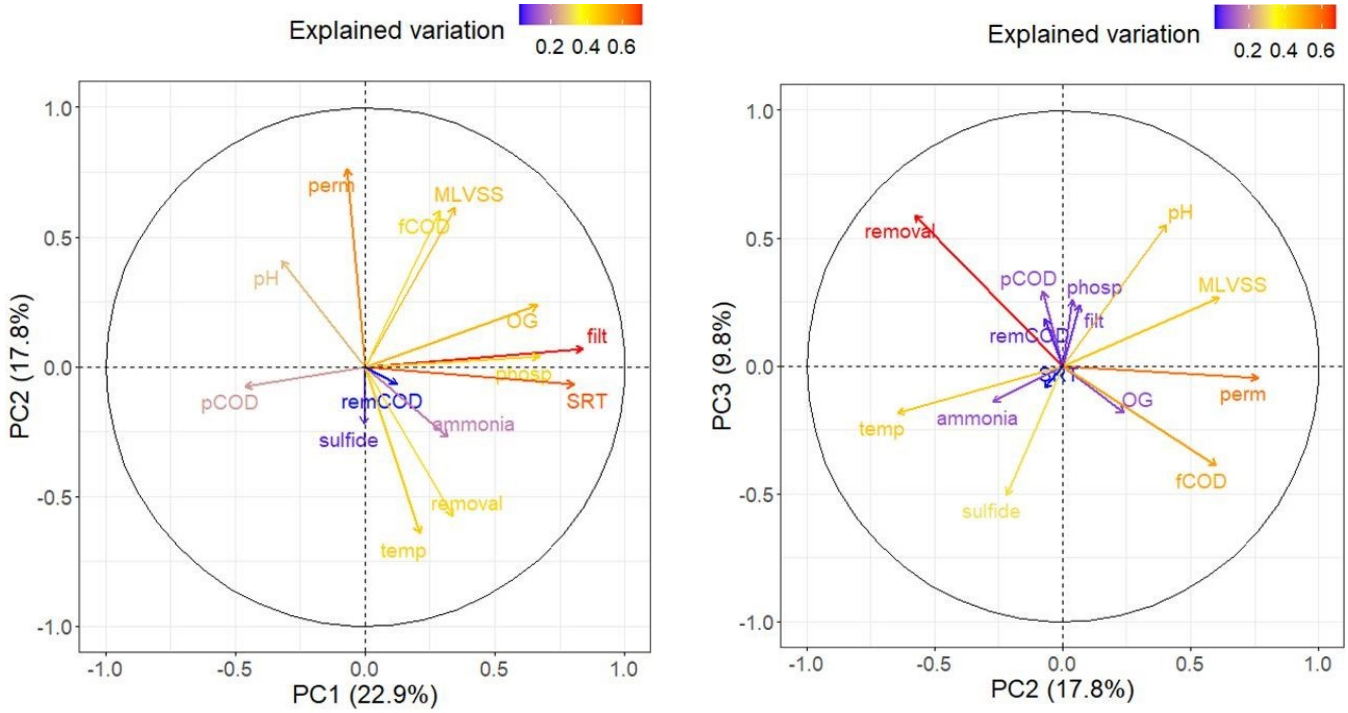
Figure 36 displays the sensitivity analysis accomplished with the ANN model, indicating the ranking of relative importance for ammonia removal. Again, only the ten most important variables are displayed, since the relative importance of the others was too low. Additionally, Figure 37 displays the loading plots obtained with PCA model, from which it is possible to understand how the variables are correlated to each other.

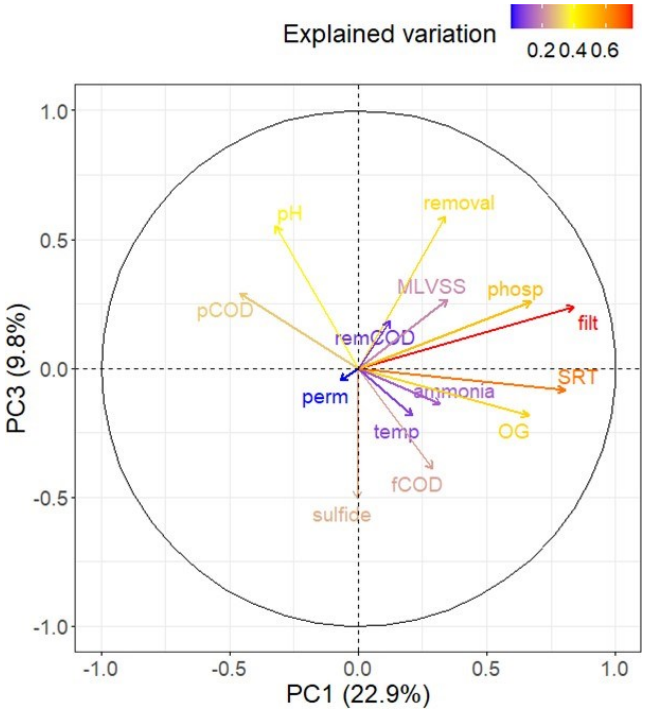
Figure 36 - Relative importance of input variables regarding ammonia removal according to the Artificial Neural Network model



filt: sludge filterability; MLVSS: mixed liquor volatile suspended solids; ammonia: influent concentration of ammonia; OG: influent concentration of oil and grease; phosp: influent concentration of phosphorous; fCOD: influent COD; temp: temperature; SRT: sludge retention time; perm: membrane permeability.

Figure 37 - Correlation between input variables and ammonia removal according to Principal Components Analysis model





filt: sludge filterability; MLVSS: mixed liquor volatile suspended solids; ammonia: influent concentration of ammonia; sulphide: influent concentration of sulphide; OG: influent concentration of oil and grease; Phosp: influent concentration of phosphorous; fCOD: influent COD; pCOD: COD in the permeate; remCOD: removal of COD; temp: temperature; SRT: sludge retention time; perm: membrane permeability.

According to both models, the variables with the greatest positive impact on ammonia removal are SRT, operating temperature and influent ammonia concentration, whereas the strongest negative impacts come from influent COD and OG concentration and membrane permeability. Zhang, Chen and Jiang (2022) also applied ANN modelling to assess the removal efficiency of ammonia in WWTP that apply different biological processes. The removal efficiencies of WWTP were successfully predicted by the model, which considered five variables: inlet flow rate, pH, influent ammonia concentration, COD and total phosphorus concentration. According to their results, influent COD concentration was the main influential factor, similar to the results found here; however, the influent ammonia concentration did not greatly affect its removal (which can be related to the alkalinity range of the system), conversely to what was observed in our work.

The strong influence of influent ammonia concentration observed can be explained though by the fact that high ammonia concentrations accelerate the kinetics of nitrification reaction (von SPERLING; CHERNICHARO, 2006), besides increasing the nitrifying bacteria growth rate, according to Monod's relation (Eqn. 1). For these reasons, higher concentrations of ammonia in the MBR feed contribute to increased oxidized ammonia load (SHARMA; AHLERT, 1977).

Besides, the growth rate of nitrifying bacteria also increases exponentially with temperature (BIAN *et al.*, 2017), which can explain the great influence of this variable on ammonia removal. Leite (2021) also investigated the main factors that influence ammonia removal on oil refinery wastewater treatment by different biological systems. The author applied multivariate statistical analyses to evaluate the impact of operating and analytic variables, like DO, pH, temperature and influent ammonia, phenol, sulphides, OG and COD concentrations. The results also showed a strong positive relationship between temperature and ammonia removal. Nitrification reaction also explains the observed negative correlation between ammonia removal and pH, since the reaction generates H^+ ions as a final product and, therefore, decreases the pH. Leite (2021) also observed this relation on her work. It is important to mention that the decrease of the pH is a function of its alkalinity (von SPERLING; CHERNICHARO, 2006). However, as previously mentioned, unfortunately there were not enough alkalinity data for us to further investigate this relation.

As for SRT, one of the main advantages of MBR over CAS technology is the independency of SRT from HRT, which allows to adopt higher SRT on MBR systems than usual in CAS. The higher SRT, in turns, allows a better sludge acclimatization, hence a greater capacity to remove pollutants (JUDD, 2016). Additionally, the complete retention of the biomass by the membrane ensures a greater diversity of microorganisms in the biological tank, which also contributes to the higher removal of different pollutants (LE-CLECH; CHEN; FANE, 2006). Furthermore, since the nitrifying bacteria's growth rate is considerably lower than the heterotrophic bacteria's one, higher SRT are needed to ensure the development of the nitrifying bacteria before they are washed out of the system (von SPERLING; CHERNICHARO, 2006). All these factors explain the strong positive correlation between SRT and ammonia removal, as previously observed in other works (ŻABCZYŃSK *et al.*, 2006).

The strong negative impact of influent COD is also related to the influence that this parameter has on the microbial community. High values of influent COD on biological treatment systems and, consequently, high C/N ratio values, provokes a greater growth of heterotrophic bacteria and result in competition for substrate and DO, prejudicing the growth of nitrifying bacteria (SHARMA; AHLERT, 1977). Sepehri and Sarrafzadeh (2018) assessed the effect of nitrifying community on nitrification efficiency on MBR. The authors obtained a nitrifying-enriched activated sludge (NAS) through particular ammonia feeding of CAS. NAS and CAS were then compared and the results demonstrated the higher nitrification efficiency of NAS, indicated by

both ammonia removal percentage (100% on NAS vs. 43% on CAS) and nitrate concentration produced (6.6 mgL^{-1} on CAS vs. 37.5 mgL^{-1} on NAS). Brasil *et al*. (2021) also demonstrated that lower C/N ratios (and higher temperatures) favor nitrifying bacteria and, consequently, improve ammonia removal. Thus, for the MBR assessed in this study, which operates under a high influent C/N ratio (20.0 $\pm$ 11.1), the ammonia removal is strongly affected by influent COD, as revealed by the models.

Higher influent OG concentrations also negatively influence on ammonia removal, due to its toxic effect on nitrifying bacteria metabolism. Studies show that even at low concentrations, toxic compounds can inhibit the bacteria activity and compromise the ammonia removal (NORIEGA-HEVIA et al., 2020). Back in 2007, Qin *et al*. were evaluating the feasibility of MBR to treat oil refinery wastewater using a testing system that ran continuously over two months. The authors observed that, in the first weeks of operation, the nitrification efficiency was low, as the ammonia removal was only 20–40%, and the high OG levels (53–153 mgL^{-1}) in the feed during this period were indicated as the principal reason. The authors also observed that nitrification was better when influent OG concentration was below 20 mgL^{-1}. OG concentrations on the assessed MBR were up to 48 mgL^{-1}, its inhibitory effect can be related to lower ammonia removals.

Since high membrane permeability decreases HRT, its negative impact on ammonia removal can be explained by the shorter time available for its degradation, which leads to insufficient time for nitrifier oxidize the available ammonia. Song *et al*. (2010) demonstrated that when HRT decreased below a certain level (6.5 days), insufficient nitrification caused a decrease in the overall nitrogen removal. Furthermore, recent works suggest that HRT is also an important factor for the development of the microbial community in biological treatment systems. Ni *et al*. (2022) studied the effects of shortened HRT on the microbial community of MBR equipped with different membrane pore size (0.40 or 0.05 μm), operated at 25 ˚C and fed with domestic wastewater. The changes observed in the microbial community in each reactor were consistent with HRT changes, indicating that it could be an important factor for maintaining microbial community structures.

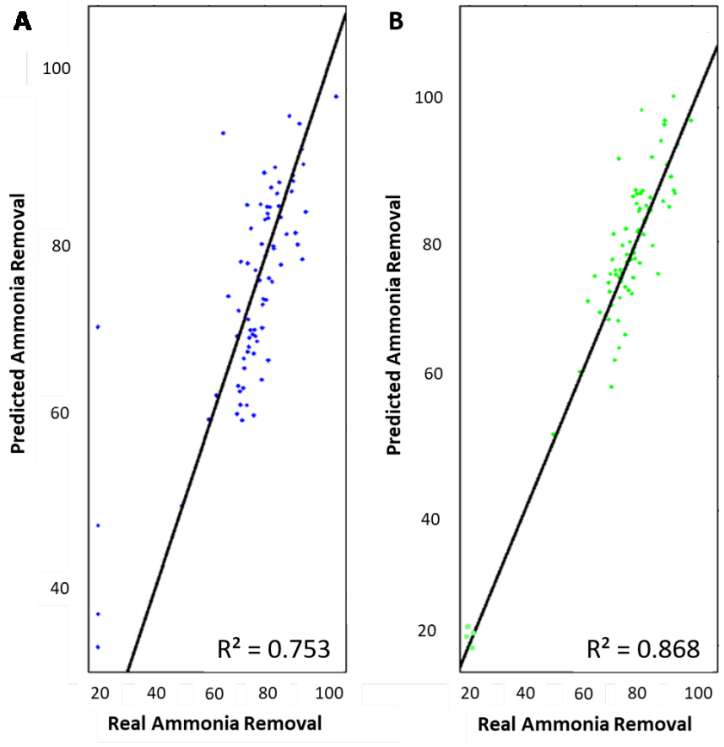
The less relevant impact of MLVSS also draws attention as it indicates that for greater ammonia removal, it is not enough to ensure a high concentration of sludge. Indeed, the most important factor is having enough ammonia-oxidizing microorganisms in the system. This result endorses

studies that state that MLVSS is not an appropriate indicator of biomass physiological activity (JACQUIN *et al.*, 2018; PAJOUM *et al.*, 2013). Hence, finding more suitable parameters to account for the concentration of nitrifying bacteria on MBR is of extreme interest. Sepehri and Sarrafzadeh (2018) also concluded in their previously presented work that NAS was twice as filterable compared to CAS. This means that higher sludge filterability values can be related to higher concentrations of nitrifying bacteria and, therefore, to higher ammonia removals. This effect can be observed on PCA and ANN models from the impact that sludge filterability exerts on ammonia removal.

3.1.2 Ammonia removal prediction

ANN model was also applied to predict the ammonia removal achieved by the MBR from a set of input conditions. As mentioned at the Methodology item, it is very important to properly know your data in order to ensure meaningful results when applying ML to ESE field. Figure 38 displays the correlation between the real (measured) and the predicted (by the model) values for ammonia removal a) without checking and improving the training dataset and b) checking and ensuring that the training dataset was representative of the real data, covering distinct removal ranges.

Figure 38 - Ammonia removal prediction by the Artificial Neural Network model a) without ensuring a representative training set and b) with a representative training set.



The graphs show that, despite apparently having predicted well the MBR behavior ($R^2 = 0.753$), the ANN initial model in which the training dataset had not been properly defined overestimated the ammonia removal in samples in which the removal was actually low. This illustrates the poor learning of the model, which was not able to correctly learn the relationships between the variables because it was not exposed to conditions of low ammonia removal. The final model though was able to adequately predict the MBR behavior ($R^2 = 0.868$), since it had the correct preparation of the training dataset. The samples with low ammonia removal percentages were thus assertively predicted and there was no further overestimation of removal, which would be a serious source of risk from an operational control point of view, since the model would disguise operational problems.

Therefore, the ANN model applied, developed with a representative training dataset, was able to effectively predict the MBR performance regarding ammonia removal, as its provisions were adequately assertive, with R^2 value equals to 0.868, indicating that the error between predicted by the model values and actual measured values was low, endorsing the high performance of ANN model (low values of MAE and MSE).

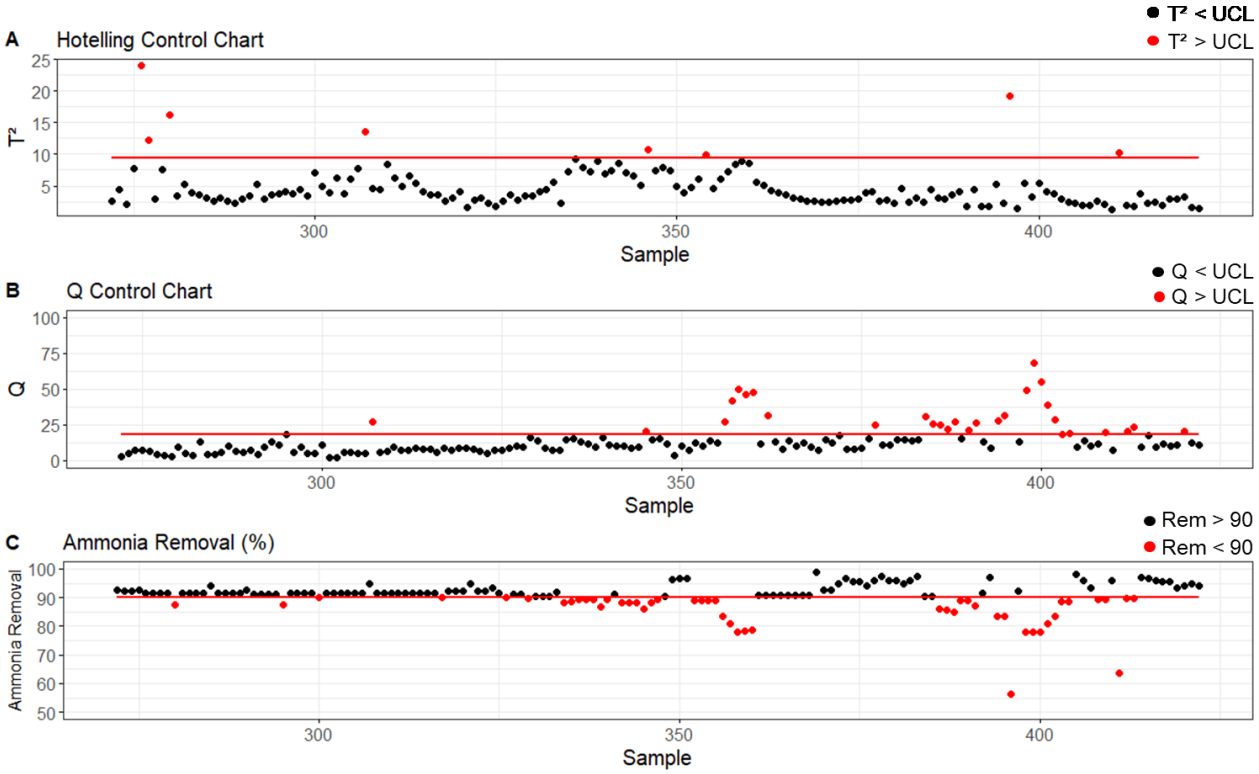
This result is of great interest because, despite being generated from offline monitoring data, it demonstrates the potential of ANN modelling to be used for online monitoring of MBR wastewater treatment systems. Due to its high predictive ability, new samples can be predicted by the model to evaluate if the system is performing as its expected to do or if there is some operating problem. By promoting a provision of data in real time, the ANN model could support more assertive decision-making, improving the monitoring and control of the process and contributing to more efficient operations.

3.2 Controlling ammonia removal on MBR systems

3.2.1 Detecting low ammonia removal conditions

In order to monitor the MBR performance regarding ammonia removal capacity along time and to detect any unusual event that lead to low percentages of removal, a MSPC-PCA model was applied to the database. The PCA model used as projection method kept three PC, according to both Kaiser criterion and the scree test. The three PC accounted for over 50% of data variation. Figure 39 presents the T^2 and Q multivariate control charts, as well as the values of ammonia removal percentage of each Phase II sample.

Figure 39 - Detection of MBR operation with low percentages of ammonia removal: a) Hotelling T² control chart; b) Q control chart; and c) ammonia removal.



UCL: Upper Limit Control; Rem: ammonia removal percentage.

T² control chart did not performed well in detecting low ammonia removal operation and presented a detection rate of samples with removal percentages below 90% of only 11%. T^2 control chart, as discussed in previous items, checks if a new observation projects on the PC hyperplane within the limits determined by the reference data. Thereby, a value of this statistic exceeding the control limits indicate that the corresponding observation presents abnormal values in some of its original variables, even though it maintains the correlation structure between the variables in the model (this observation is tagged as an outlier inside the PCA model) (FERRER, 2014; KOURTI, 2005). This way, the low detection rate of T² control chart can be explained by the fact that the out-of-control samples presented some breakage on the model correlation structure. This provides an interesting insight about failures in ammonia removal on MBR systems, since the problems are caused not by extreme values of the individual variables, but by the deviation of the correlation structure present among them.

Q-statistic control chart, in turn, checks the occurrence of any new events that cause the process to move away from the hyperplane defined by the reference model. This way, this chart can detect failures caused by the deviation of the correlation structure and, therefore, it was able to detect the operation with low ammonia removal percentages. Despite detecting only 54% of the

operation with less than 90% of ammonia removal, Q control chart detected all operating points with removals below 85%. This is meaningful for the control of operation point of view, since this early detection can prevent the level of ammonia removal from reaching extremely low values, guiding the most appropriate times and manners to act on the system. Furthermore, this result indicates that with better adjustments in the model or more representative input data, the detection rate of operation with ammonia removal percentages lower than 90% could be improved.

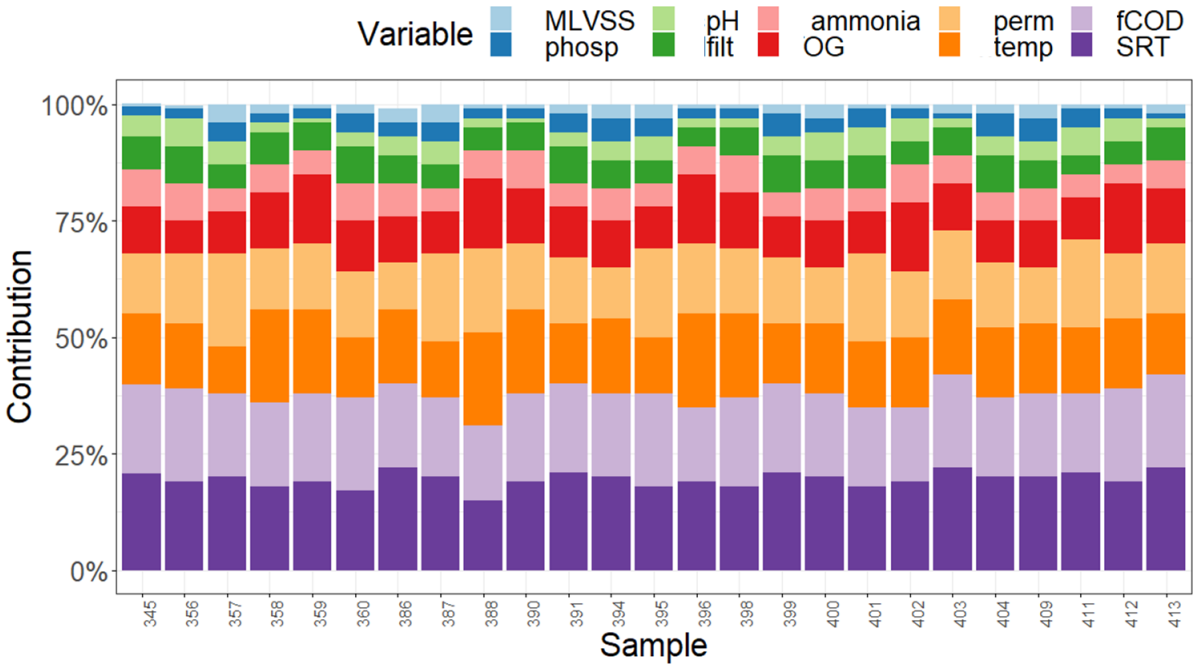
Nevertheless, both statistics had a low percentage of false alarms (3% in T^2 chart and 1% in Q chart), considering that these errors are inevitable due to the very probability of the statistics (5%). Therefore, Q control chart is potentially an interesting tool for the monitoring and control of ammonia removal on MBR, as it can detect its reductions. With further improvement in the building of the chart, MSPC application can thus enable preventive and more assertive acting on the system, avoiding unnecessary costs, increasing treated water quality and, consequently, improving MBR efficiency.

3.2.2 Diagnosing low ammonia removal conditions

Since Q-statistic performed better on the detection of low ammonia removal operation, this control chart was the one used to build the contribution plot aiming to investigate what caused the low ammonia removals observed during the MBR operation (

Figure 40). Alike to sensitivity analysis results, only the ten most important variables were displayed for simplicity purposes.

Figure 40 - Contribution plot based on Q-statistics for out-of-control (ammonia removal lower than 90%) observations.



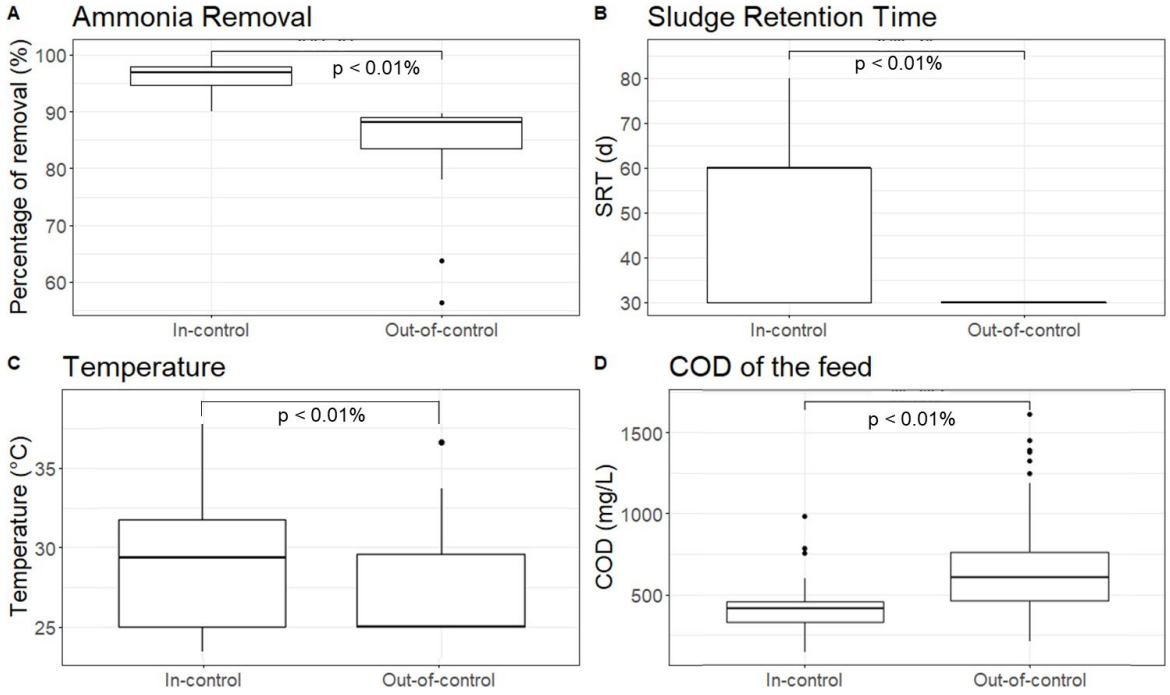
filt: sludge filterability; MLVSS: mixed liquor volatile suspended solids; ammonia: influent concentration of ammonia; OG: influent concentration of oil and grease; phosp: influent concentration of phosphorous; fCOD: influent COD; temp: temperature; SRT: sludge retention time; perm: membrane permeability.

From the plot it can be noticed that the variables with greater contributions to the exceeding values of Q-statistics were influent COD, SRT and temperature, which indicates that these three variables influenced the most on ammonia removal reduction, endorsing the results obtained with PCA and ANN models. This also confirms the important role that C/N ratio plays on ammonia removal control, as well as demonstrates the importance of keeping the system operating under proper values of temperature and SRT, since high values of influent COD and low SRT and temperature values were the most common causes of low percentages of ammonia removal by the MBR. Studying the association of these different factors in order to find a balance between them is also highly interesting, since, in practice, it is not always possible to keep all of them within the desirable range. Stewart *et al*. (2022), for example, assessed the performance of two pilot-scale biological nutrient removal (BNR) treatment trains and observed that increasing the SRT during the period of cold water temperatures helped maintain ammonia removal performance.

To confirm the results, the nonparametric Wilcoxon-Mann-Whitney statistical test (MANN; WHITNEY, 1947) at a significance level of 1% was applied to check whether the medians of the variables ammonia removal, temperature, SRT and influent COD were significantly

different between the in and out-of-control operation. Figure 41 displays the boxplots and the results of the hypothesis test (the p-values obtained are presented in Appendix F). It can be noted the significantly lower values of ammonia removal, temperature and SRT and the significantly higher values of influent COD in out-of-control operation, confirming the previous results.

Figure 41 - Boxplots and nonparametric statistical test of Wilcoxon-Mann-Whitney at a significance level of 1% for ammonia removal, sludge retention time, temperature and influent COD for in and out-of-control operation.



4 CONCLUSION

For a pilot-scale MBR applied for the treatment of a real oil refinery wastewater, ANN, PCA and MSPC have proven to be suitable for monitoring the wastewater treatment system aiming ammonia removal better understanding and control. PCA and ANN models were effective in mapping the MBR behavior regarding ammonia removal and to detect the most important factors for the pollutant removal, both positive and negatively. The models identified SRT, temperature and influent concentration of ammonia as the variables that improve the most ammonia removal, whereas influent concentration of COD and OG and membrane permeability are the variables that decrease it the most. ANN model was also successfully applied to predict the ammonia removal, with R^2 equals to 0.87. Therefore, the error between the values predicted by the model and the actual measured values was low, proving that ANN was effective in predicting the MBR behavior and thus could be used to forecast ammonia removals from a set of operating conditions. T^2 control chart did not perform well in detecting operating failures related to low ammonia removal percentages, which can indicate that these faults are caused by the deviation of the system from the correlation structure among the variables. Q control chart, in turn, was able to detect all of the operation with removals lower than 85% and, therefore, it is possible that, with further adjustments in the model, it would be able to detect operations below 90% of ammonia removal. Thus, MSPC can potentially be used to better control the MBR, properly adjusting the C/N ratio, the temperature or the SRT, for example, and preventing ammonia removal to go to severe low levels. Hence, the results demonstrate the potential of AI and ML techniques to be used for monitoring and controlling MBR operation, contributing for the improvement of the process efficiency.

V. FINAL CONSIDERATIONS

1 THESIS OVERVIEW AND INTEGRATED RESULTS DISCUSSION

Contamination of natural water by inappropriate disposal of industrial residues and wastewater is currently one of the greatest environmental harms and the presence of emerging and persistent pollutants in industrial wastewater is an important global concern that has been leading to increasingly stringent environmental regulations. Furthermore, due to the urgent need of better managing water resources, the demand for water reuse is growing fast, especially in industries. Therefore, the application of highly efficient wastewater treatment technologies that provide high quality treated water and allow its reuse has been increasingly sought. MBR are currently considered a highly efficient technology and stand out for the high effluent quality achieved. Other MBR advantages include high removal of micro- and persistent organic pollutants, small industrial area requirement, and low sludge production.

Oil refineries, in particular, have been increasingly applying MBR technology, since they can strongly benefit from water reuse due to the large amount of water needed. However, to enable water reuse, effluent ammonia concentration must be sufficiently low and thus understanding and controlling the factors that impact the most on its removal is essential, as well as being able to predict it. The application of AI and ML can greatly contribute to this matter, since they can successfully realize feature extraction, correlation analysis and patterns identification. Besides, membrane fouling is still a serious backdraw for the wider application of MBR, specially for treating complex industrial wastewater such as those from oil refineries, since it decreases the process performance and leads to permeate flux decline, which results in higher operating costs. As understanding fouling on MBR is a complex task though, AI and ML stand out once more as very promising alternatives for the investigation, monitoring and controlling of membrane fouling on MBR.

Therefore, this work aimed to improve membrane fouling control and ammonia removal on MBR through the application of different AI and ML models. As PCA, ANN and MSPC have been standing out for their high performances monitoring, mapping and predicting different complex systems, they emerged as an interesting option for better monitoring and controlling MBR operation and thus were the selected techniques to be applied in this work. A pilot-scale MBR treating a real oil refinery wastewater and monitored during five years was used as a case study. The models were applied to investigate the relations between different variables and to detect and diagnose operating faults related to the occurrence of membrane fouling and to low

percentages of ammonia removal, aiming to comprehend their main causes and to propose efficient strategies for their control.

PCA and MSPC have proven to be suitable for monitoring MBR aiming membrane fouling control. PCA was able to map the MBR behavior and identified sludge filterability, temperature and SDWC as the variables that influence the most on membrane permeability. It was also able to predict the MBR performance, with high values of R^2 and Q^2 (0.71 and 0.78, respectively), and to distinguish atypical samples, enabling the detection of operating problems. T^2 and Q control charts, assessed in a combined manner, have also proven to be effective in the detection of membrane fouling and allowed to preventively detect membrane permeability reductions. They can thus be used to guide when to dose permeability improvers and/or perform chemical cleaning, which ensures that preventive actions are performed at the most appropriate time, avoiding unnecessary costs and preserving the membrane lifetime. Besides controlling the sludge filterability and stablishing an efficient chemical cleaning strategy, MSPC modelling also revealed that preventing the temperature from increasing is an important fouling mitigation measure.

PCA and ANN were also effective in modelling the MBR behavior regarding ammonia removal and identified SRT, temperature and influent concentration of ammonia as the variables that improve the most the MBR ammonia removal capacity, whereas influent concentration of COD and OG and membrane permeability are the variables that decrease it the most. ANN model was also successfully applied to predict the ammonia removal from a set of operating and feed conditions, with R^2 equals to 0.87. Besides, MSPC showed its potential to be used as monitoring tool regarding the improvement of ammonia removal on MBR. Although Q control chart was not so effective in detecting ammonia removals below 90%, it detected all of the operation with removals lower than 85%. This demonstrates the potential of the model, that can possibly perform better whith further adjustments in input data and model settings. Furthermore, as T^2 control chart did not detect operating failures related to low ammonia removal percentages, it can be inferred that these operating problems are caused mostly by the deviation of the system from the expected correlation structure between the variables. Based on these insights and on the first results obtained with both control charts thus, MSPC can be improved and it could be then used to better control the MBR ammonia removal capacity, indicating the best times to act on the system, properly adjusting the C/N ratio, the temperature or the SRT, for example, and preventing ammonia removal to go to low levels.

Analyzing the results obtained with all models applied in an integrated way, thus, reveals some interesting insights about the MBR performance. Firstly, it is relevant to notice that well monitoring sludge filterability is highly important to increase MBR overall efficiency, since low values of this variable indicate greater propensity to membrane fouling occurrence and lower concentration of nitrifying bacteria. It is also worth commenting that the dosage of permeability improvers increases sludge filterability by coagulation/flocculation of the sludge and, consequently, although it contributes to mitigate membrane fouling, it doesn't help to increase ammonia removal. For that purpose, its necessary to keep the MBR under conditions that favor nitrifying bacteria growth, for example, keeping lower C/N ratios. By all means, investing in sludge filterability properly monitoring is strongly recommended as it can easily inform about the MBR biomass conditions and thus supports the decision-making.

Furthermore, keeping lower C/N ratios is, in fact, a great way to improve MBR performance as it contributes not only to higher ammonia removals, but also to reduce membrane fouling since favoring the nitrifying bacteria growth also leads to smaller releases of EPS and SMP. This way, investing in efficient pre-treatment steps that can meaningfully remove organic matter, to ensure reasonable concentrations of COD in the MBR feed, is also critical and should be pursuit. Nevertheless, an interesting investigation to be further carried out is verifying if, faced with an unavoidably high concentration of COD in the feed, higher values of temperature or SRT, for example, could overcome it and increase ammonia removal.

Increasing MBR temperature, however, should be careful evaluated, as it provokes more severe membrane fouling. As higher values of temperature accelerate the metabolism of both nitrifying and heterotrophic bacteria, they end up causing a greater release of SMP and EPS at the same time that they lead to higher nitrification rates. Therefore, finding an optimal point between greater ammonia removal and milder membrane fouling is a challenge to be better studied.

Another important point of attention is the relationship between membrane permeability and ammonia removal themselves. Higher membrane permeability values are desirable as they make the treatment process more productive and, consequently, increase its economic viability. However, they also reduce the HRT and thus compromise the ammonia removal achieved. This way, careful evaluating each case and finding an optimal point between productivity and quality of the treated effluent is essential.

2 CONCLUSIONS AND RECOMMENDATIONS

The hypotheses that motivated this work were: i) ANN and PCA would be able to reveal the most important variables for both ammonia removal and membrane fouling occurrence; ii) they would also be able to predict membrane permeability and ammonia removal percentage values from a set of input conditions; iii) MSPC would be able to detect and diagnose failures in MBR operation regarding both membrane fouling occurrence and low ammonia removal capacity; and iv) the integrated analysis of the models would support the definition of more efficient strategies for better membrane fouling control and greater ammonia removal. Based on the results presented and discussed in the present work thus, all hypotheses have proven to be true. Furthermore, all proposed goals, i.e., to further comprehend membrane fouling and ammonia removal on MBR systems; to predict removal percentages and membrane permeability values; to detect and diagnose operation points with low ammonia removal and membrane fouling occurrence; and to propose more effective strategies for a more efficient MBR operation; were accomplished.

ANN, PCA and MSPC proved to be highly efficient in monitoring, predicting and controlling MBR wastewater treatments. The results provided by the models were reliable (satisfactory R^2, Q^2, MAE and MSE values) and contributed to a better understanding of the process, as well as to the prediction of its behavior and to the definition of more proper control strategies, regarding both membrane fouling occurrence and ammonia removal capacity.

The main focus of action for more efficient MBR operations are lower COD concentrations in the reactor feed (to be guaranteed by an efficient pretreatment), higher SRT values and suitable temperature adjustment. Performing chemical cleanings in the most appropriate moments also ensure milder membrane fouling and, therefore, higher productivity and treated effluent quality. In addition, properly monitoring sludge filterability is of great importance, since this variable successfully performs as a monitoring tool for sludge quality and, consequently, to monitor the occurrence of both nitrification and membrane fouling.

Despite being generated from offline monitoring data, the results demonstrated the massive potential of AI/ML techniques to be used for monitoring and controlling MBR operations. They also indicate that AI/ML modelling can be used for the assessment of the system status in real-time, allowing better-informed decision-making, as long as online data is available.

Ensuring greater ammonia removals is essential, as it contributes to water reuse; and preventing membrane fouling is equally important, as it contributes to improve MBR cost-effectiveness. Hence, the work contributed to more efficient operations of MBR, contributing not only to its more widespread application, but also ensuring better performances for the already existing applications. Furthermore, by facilitating a more widespread application of such an important wastewater treatment technology, this work contributed to reduce the impacts associated with inappropriate industrial wastewater disposal, which protects the environment and improves public health and welfare.

There are some limitations for the results obtained here though and, for further improvements in the technology, some efforts are still needed. To further develop this research and improve MBR efficiency thus, some recommendations are:

i. Evaluation of the effects of other important variables that could not be included in this work, like SMP, EPS, HRT and alkalinity;
ii. Morphological characterization of the material deposited on the membrane surface and in its pores, in order to investigate the type of membrane fouling and possible distinct fouling from clogging;
iii. Investigation of the combined effect of distinct variables to check whether, keeping an important variable as it is (high values of influent COD, for instance), adjusting other high impact variables (increasing SRT or temperature, for example) is able to overcome the effects of the first one and improve the MBR performance;
iv. Simultaneous assessment of membrane fouling occurrence and ammonia removal capacity, to determine the best conditions for variables with opposite effects on ammonia removal and membrane fouling occurrence, like temperature and membrane permeability itself;
v. Further improvements on MSPC modelling, considering model settings and input data preprocessing, in order to increase its fault detection ability, speacily regarding low ammonia removal conditions;
vi. Application of other AI/ML models, like RF, GA, SVM to consolidate their potential as monitoring tools and to possibly reveal new insights about MBR operation;
vii. Model development with online data, to assess the MBR state at real time and promote better informed decision-making;
viii. Validation of the results obtained with a full-scale MBR, since the relations between the different variables in these systems can differ from the ones found in this work.

REFERENCES

ABDI, Hervé; WILLIAMS, Lynne J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, v. 2, n. 4, p. 433–459, 2010.

AESOY, Anette; ODEGAARD, Hallvard; BENTZEN, Greta. The effect of sulphide and organic matter on the nitrification activity in a biofilm process. *Water Science*, v. 1, n. 1, p. 1–15, 1998. Disponível em: <http://www.fao.org/3/I8739EN/i8739en.pdf%0Ahttp://dx.doi.org/10.1016/j.adolescence.2017.01.003%0Ahttp://dx.doi.org/10.1016/j.childyouth.2011.10.007%0Ahttps://www.tandfonline.com/doi/full/10.1080/23288604.2016.1224023%0Ahttp://pjx.sagepub.com/lookup/doi/10>.

AJJUR, Salah Basem; AL-GHAMDI, Sami G. Seventy-year disruption of seasons characteristics in the Arabian Peninsula. *International Journal of Climatology*, v. 41, n. 13, p. 5920–5937, 2021.

ALKMIM, Aline R. *et al.* Improving knowledge about permeability in membrane bioreactors through sensitivity analysis using artificial neural networks. *Environmental Technology (United Kingdom)*, v. 41, n. 19, p. 2424–2438, 2020. Disponível em: <https://doi.org/10.1080/09593330.2019.1567609>.

ALKMIM, Aline R. *et al.* Potential use of membrane bioreactor to treat petroleum refinery effluent: comprehension of dynamic of organic matter removal, fouling characteristics and membrane lifetime. *Bioprocess and Biosystems Engineering*, v. 40, n. 12, p. 1839–1850, 15 dez. 2017. Disponível em: <http://link.springer.com/10.1007/s00449-017-1837-4>.

ALKMIM, Aline Ribeiro *et al.* The application of filterability as a parameter to evaluate the biological sludge quality in an MBR treating refinery effluent. *Desalination and Water Treatment*, v. 53, n. 6, p. 1440–1449, 2015.

ALLAIRE, J; CHOLLET, François. *keras: R Interface to "Keras"*. . [S.l.]: CRAN.R-project.org. Disponível em: <https://cran.r-project.org/package=keras>. , 2022

AMARAL, Míriam C.S. *et al.* Pilot aerobic membrane bioreactor and nanofiltration for municipal landfill leachate treatment. *Journal of Environmental Science and Health - Part A Toxic/Hazardous Substances and Environmental Engineering*, v. 51, n. 8, p. 640–649, 2016.

AMARAL, Míriam Cristina Santos *et al.* Treatment of refinery effluents by pilot membrane bioreactors: pollutants removal and fouling mechanism investigation. *Desalination and Water Treatment*, v. 56, n. 3, p. 583–597, 16 out. 2015. Disponível em: <http://www.tandfonline.com/doi/full/10.1080/19443994.2014.953595>.

AMARI, Shun-ichi. *Backpropagation and stochastic gradient descent method*. *Neurocomputing*. [S.l: s.n.]. , 1993

APHA; AWWA; WEF. *Standard Methods for Examination of Water and Wastewater*. 22nd. ed. Washington: American Public Health Association, 2012. Disponível em: <https://www.standardmethods.org/>.

ARSALANE, Assia *et al.* An embedded system based on DSP platform and PCA-SVM algorithms for rapid beef meat freshness prediction and identification. *Computers and Electronics in Agriculture*, v. 152, n. January, p. 385–392, 2018.

AWOLUSI, O. O.; KUMARI, S. K.S.; BUX, F. Ecophysiology of nitrifying communities in membrane bioreactors. *International Journal of Environmental Science and Technology*, v. 12, n. 2, p. 747–762, 2015.

AZAMI, Hamed; SARRAFZADEH, Mohammad Hossein; MEHRNIA, Mohammad Reza. Influence of sludge rheological properties on the membrane fouling in submerged membrane bioreactor. *Desalination and Water Treatment*, v. 34, n. 1–3, p. 117–122, 2011.

BAGHERI, Majid; AKBARI, Ali; MIRBAGHERI, Sayed Ahmad. Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning techniques: A critical review. *Process Safety and Environmental Protection*, v. 123, p. 229–252, 2019. Disponível em: <https://doi.org/10.1016/j.psep.2019.01.013>.

BAGHERI, Majid; MIRBAGHERI, Sayed Ahmad. Critical review of fouling mitigation strategies in membrane bioreactors treating water and wastewater. *Bioresource Technology*, v. 258, n. March, p. 318–334, 2018. Disponível em: <https://doi.org/10.1016/j.biortech.2018.03.026>.

BALLABIO, Davide. A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemometrics and Intelligent Laboratory Systems*, v. 149, p. 1–9, 2015. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2015.10.003>.

BANERJEE, Srila *et al.* Performance assessment of the indigenous ceramic UF membrane in bioreactor process for highly polluted tannery wastewater treatment. *Environmental Science and Pollution Research*, p. 48620–48637, 2022. Disponível em: <https://doi.org/10.1007/s11356-022-19258-z>.

BAROUTIAN, Saeid; ESHTIAGHI, Nicky; GAPES, Daniel J. Rheology of a primary and secondary sewage sludge mixture: Dependency on temperature and solid concentration. *Bioresource Technology*, v. 140, p. 227–233, 2013. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2013.04.114>.

BASHEER, I.A.; HAJMEER, M. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, v. 43, p. 3–31, 2000.

BAYAT, Mitra *et al.* Petrochemical wastewater treatment and reuse by MBR: A pilot study for ethylene oxide/ethylene glycol and olefin units. *Journal of Industrial and Engineering Chemistry*, v. 25, p. 265–271, 2015. Disponível em: <http://dx.doi.org/10.1016/j.jiec.2014.11.003>.

BCC RESEARCH. *Membrane Bioreactors (MBRs) Market: Size, Share & Technology Report*. Disponível em: <https://www.bccresearch.com/market-research/membrane-and-separation-technology/membrane-bioreactors.html>. Acesso em: 28 maio 2021.

BERSIMIS, S.; PANARETOS, J.; PSARAKIS, S. Multivariate Statistical Process Control Charts and the Problem of Interpretation: A Short Overview and Some Applications in Industry. 2009. Disponível em: <http://arxiv.org/abs/0901.2880>.

BIAN, Wei *et al.* Achieving nitritation in a continuous moving bed biofilm reactor at different temperatures through ratio control. *Bioresource Technology*, v. 226, p. 73–79, 2017. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2016.12.014>.

BRASIL, Yara L. *et al.* Technical and economic evaluation of the integration of membrane bioreactor and air-stripping/absorption processes in the treatment of landfill leachate. *Waste Management*, v. 134, n. August, p. 110–119, 2021. Disponível em: <https://doi.org/10.1016/j.wasman.2021.08.013>.

BREPOLS, C. *et al.* Strategies for chemical cleaning in large scale membrane bioreactors. *Water Science and Technology*, v. 57, n. 3, p. 457–463, 2008.

BRO, Rasmus; SMILDE, Age K. Principal component analysis. *Anal. Methods*, v. 6, n. 9, p. 2812–2831, 2014. Disponível em: <http://xlink.rsc.org/?DOI=C3AY41907J>.

CAI, Meiqiang *et al.* Improving dewaterability and filterability of waste activated sludge by electrochemical Fenton pretreatment. *Chemical Engineering Journal*, v. 362, n. January, p. 525–536, 2019. Disponível em: <https://doi.org/10.1016/j.cej.2019.01.047>.

CAMACHO, José *et al.* PCA-based multivariate statistical network monitoring for anomaly detection. *Computers and Security*, v. 59, p. 118–137, 2016. Disponível em: <http://dx.doi.org/10.1016/j.cose.2016.02.008>.

CATELANI, Tiago A. *et al.* Real-time monitoring of a coffee roasting process with near infrared spectroscopy using multivariate statistical analysis: A feasibility study. *Talanta*, v. 179, n. November 2017, p. 292–299, 2018. Disponível em: <https://doi.org/10.1016/j.talanta.2017.11.010>.

CATTELL, R. The Scree Test for the number of factors. Multivariate Behavioral Research. *Multivariate Behavioral Research. 1*, v. 1, n. August, p. 116–141, 1966.

CHANG, Hau Ming *et al.* Enhanced understanding of osmotic membrane bioreactors through machine learning modeling of water flux and salinity. *Science of the Total Environment*, v. 838, n. May, p. 156009, 2022. Disponível em: <https://doi.org/10.1016/j.scitotenv.2022.156009>.

CHANG, In-Soung *et al.* Membrane Fouling in Membrane Bioreactors for Wastewater Treatment. *Journal of Environmental Engineering*, v. 128, n. 11, p. 1018–1029, 2002.

CHEN, Mingzhe *et al.* Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial. *IEEE Communications Surveys and Tutorials*, v. 21, n. 4, p. 3039–3071, 2019.

CHEN, Zhaobo *et al.* Performance of a novel multiple draft tubes airlift loop membrane bioreactor to treat ampicillin pharmaceutical wastewater under different temperatures. *Chemical Engineering Journal*, v. 380, n. August 2019, p. 122521, 2020. Disponível em: <https://doi.org/10.1016/j.cej.2019.122521>.

CHENG, Hui *et al.* Long-term operation performance and fouling behavior of a high-solid anaerobic membrane bioreactor in treating food waste. *Chemical Engineering Journal*, v. 394, n. March, 2020.

CHENG, Yingchao; LI, Huan. Rheological behavior of sewage sludge with high solid content. *Water Science and Technology*, v. 71, n. 11, p. 1686–1693, 2015.

CHOI, Byeong Gyu *et al.* Correlation between effluent organic matter characteristics and membrane fouling in a membrane bioreactor using advanced organic matter characterization tools. *Desalination*, v. 309, p. 74–83, 2013. Disponível em: <http://dx.doi.org/10.1016/j.desal.2012.09.018>.

CHUGH, Akshita. *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?* Disponível em: <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>.

COUTO, Carolina Fonseca *et al.* Hybrid MF and membrane bioreactor process applied towards water and indigo reuse from denim textile wastewater. *Environmental Technology (United Kingdom)*, v. 39, n. 6, p. 725–738, 2018. Disponível em: <http://dx.doi.org/10.1080/09593330.2017.1310307>.

DAS, Anupam. Multivariate statistical monitoring strategy for an automotive manufacturing part facility. *Materials Today: Proceedings*, v. 27, p. 2914–2917, 2019. Disponível em: <https://doi.org/10.1016/j.matpr.2020.03.515>.

DAVIES, Laurie; GATHER, Ursula. The identification of multiple outliers. *Journal of the American Statistical Association*, v. 88, n. 423, p. 782–792, 1993.

DENG, Lijuan *et al.* Biofouling and control approaches in membrane bioreactors. *Bioresource Technology*, v. 221, p. 656–665, 2016. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2016.09.105>.

DING, Yi *et al.* Long-term investigation into the membrane fouling behavior in anaerobic membrane bioreactors for municipal wastewater treatment operated at two different temperatures. *Membranes*, v. 10, n. 9, p. 1–11, 2020.

DU, Xianjun *et al.* A review on the mechanism, impacts and control methods of membrane fouling in MBR system. *Membranes*, v. 10, n. 2, p. 1–33, 2020.

ENVIRONMENTAL PROTECTION AGENCY. Manual: Nitrogen Control. n. September, p. 311, 1993.

ERIKSSON, L. *et al. Multi- and Megavariate Data Analysis: Basic Principles and Applications*. [S.l.]: Umetrics Academy, 2013.

EVENBLIJ, H. *et al.* Filtration characterisation for assessing MBR performance: Three cases compared. *Desalination*, v. 178, n. 1- 3 SPEC. ISS., p. 115–124, 2005.

EVERITT, Brian; HOTHORN, Torsten. *An Introduction to applied multivariate analysis*. [S.l: s.n.], 2008.

FENG, Suping *et al.* The effect of COD/N ratio on process performance and membrane fouling in a submerged bioreactor. *Desalination*, v. 285, p. 232–238, 2012. Disponível em: <http://dx.doi.org/10.1016/j.desal.2011.10.008>.

FERRER, Alberto. Latent structures-based multivariate statistical process control: A paradigm shift. *Quality Engineering*, v. 26, n. 1, p. 72–91, 2014.

FERRER, Alberto. Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process. *Quality Engineering*, v. 19, n. 4, p. 311–325, 2007.

FORKMAN, Johannes; JOSSE, Julie; PIEPHO, Hans Peter. Hypothesis Tests for Principal Component Analysis When Variables are Standardized. *Journal of Agricultural, Biological, and Environmental Statistics*, Sílvia me passou (17/04/20)., v. 24, n. 2, p. 289–308, 2019.

FRAGA, Florencia Arón *et al.* Evaluation of a membrane bioreactor on dairy wastewater treatment and reuse in Uruguay. *International Biodeterioration and Biodegradation*, v. 119, p. 552–564, 2017.

FRITSCH, Stefan; FRAUKE, Guenther; WRIGHT, Marvin. *neuralnet: Training of Neural Networks*. . [S.l.]: CRAN.R-project.org. Disponível em: <https://cran.r-project.org/package=neuralnet>. , 2019

GAO, Da Wen *et al.* Membrane fouling related to microbial community and extracellular polymeric substances at different temperatures. *Bioresource Technology*, v. 143, p. 172–177, 2013. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2013.05.127>.

GAO, W. J. *et al.* Influence of temperature and temperature shock on sludge properties, cake layer structure, and membrane fouling in a submerged anaerobic membrane bioreactor. *Journal of Membrane Science*, v. 421–422, p. 131–144, 2012.

GIL, José Antonio *et al.* Influence of temperature variations on the cake resistance and EPS of MBR mixed liquor fractions. *Desalination and Water Treatment*, v. 18, n. 1–3, p. 1–11, 2010.

GIWA, A. *et al.* Experimental investigation and artificial neural networks ANNs modeling of electrically-enhanced membrane bioreactor for wastewater treatment. *Journal of Water Process Engineering*, v. 11, p. 88–97, 2016. Disponível em: <http://dx.doi.org/10.1016/j.jwpe.2016.03.011>.

GKOTSIS, Petros K. *et al.* Fouling issues in Membrane Bioreactors (MBRs) for wastewater treatment: Major mechanisms, prevention and control strategies. *Processes*, v. 2, n. 4, p. 795–866, 2014.

GONZÁLEZ-CAMEJO, J. *et al.* Continuous 3-year outdoor operation of a flat-panel membrane photobioreactor to treat effluent from an anaerobic membrane bioreactor. *Water Research*, v. 169, 2020.

GRASSI, Silvia *et al.* Control and monitoring of milk renneting using FT-NIR spectroscopy as a process analytical technology tool. *Foods*, v. 8, n. 9, 2019.

GUAN, Dao *et al.* Comparison of different chemical cleaning reagents on fouling recovery in a Self-Forming dynamic membrane bioreactor (SFDMBR). *Separation and Purification Technology*, v. 206, n. January, p. 158–165, 2018. Disponível em: <https://doi.org/10.1016/j.seppur.2018.05.059>.

GUTIÉRREZ, Marina *et al.* Activated carbon coupled with advanced biological wastewater treatment: A review of the enhancement in micropollutant removal. *Science of the Total Environment*, v. 790, 2021.

GUTIÉRREZ, Marina *et al.* Removal of micropollutants using a membrane bioreactor coupled with powdered activated carbon — A statistical analysis approach. *Science of The Total Environment*, v. 840, n. April, p. 156557, 2022.

HADIAN, Hengameh; RAHIMIFARD, Ali. Multivariate statistical control chart and process capability indices for simultaneous monitoring of project duration and cost. *Computers and Industrial Engineering*, v. 130, n. March, p. 788–797, 2019. Disponível em: <https://doi.org/10.1016/j.cie.2019.03.021>.

HAGLIN, Jack M.; JIMENEZ, Genesis; ELTORAI, Adam E.M. Artificial neural networks in medicine. *Health and Technology*, v. 9, n. 1, p. 1–6, 2019. Disponível em: <http://dx.doi.org/10.1007/s12553-018-0244-4>.

HE, Fei; ZHANG, Lingying. Prediction model of end-point phosphorus content in BOF steelmaking process based on PCA and BP neural network. *Journal of Process Control*, v. 66, p. 51–58, 2018. Disponível em: <https://doi.org/10.1016/j.jprocont.2018.03.005>.

HE, Sheng bing; XUE, Gang; WANG, Bao zhen. Factors affecting simultaneous nitrification and de-nitrification (SND) and its kinetics model in membrane bioreactor. *Journal of Hazardous Materials*, v. 168, n. 2–3, p. 704–710, 2009.

HONG, Phuc Nguon *et al.* Mechanism of biofouling enhancement in a membrane bioreactor under constant trans-membrane pressure operation. *Journal of Membrane Science*, v. 592, n.

February, p. 117391, 2019a. Disponível em: <https://doi.org/10.1016/j.memsci.2019.117391>.

HONG, Phuc Nguon *et al.* Mechanism of biofouling enhancement in a membrane bioreactor under constant trans-membrane pressure operation. *Journal of Membrane Science*, v. 592, n. August, p. 117391, 15 dez. 2019b. Disponível em: <https://doi.org/10.1016/j.memsci.2019.117391>. Acesso em: 28 mar. 2021.

HOSSEINZADEH, Ahmad *et al.* Modeling water flux in osmotic membrane bioreactor by adaptive network-based fuzzy inference system and artificial neural network. *Bioresource Technology*, v. 310, n. April, p. 123391, 2020. Disponível em: <https://doi.org/10.1016/j.biortech.2020.123391>.

HOTELLING, Harold. Analysis of a complex of statistical variables into Principal Components. Jour. Educ. Psych., 24, 417-441, 498-520. *The Journal of Educational Psychology*, v. 24, p. 417–441, 1933.

HOTELLING, Harold. Multivariate Quality Control. *Techniques of Statistical Analysis*, 1947. Disponível em: <http://ci.nii.ac.jp/naid/10021322508/en/>. Acesso em: 23 jul. 2021.

HU, Yisong *et al.* A review of anaerobic membrane bioreactors for municipal wastewater treatment with a focus on multicomponent biogas and membrane fouling control. *Environmental Science: Water Research and Technology*, v. 6, n. 10, p. 2641–2663, 2020.

HUANG, Shujuan *et al.* Performance and process simulation of membrane bioreactor (MBR) treating petrochemical wastewater. *Science of the Total Environment*, v. 747, p. 141311, 2020. Disponível em: <https://doi.org/10.1016/j.scitotenv.2020.141311>.

INGLEZAKIS, V. J. *et al.* Investigating the inhibitory effect of cyanide, phenol and 4-nitrophenol on the activated sludge process employed for the treatment of petroleum wastewater. *Journal of Environmental Management*, v. 203, p. 825–830, 2017. Disponível em: <http://dx.doi.org/10.1016/j.jenvman.2016.08.066>.

JACKSON, D. A. Stopping Rules in Principal Components Analysis : A Comparison of Heuristical and Statistical Approaches Author (s): Donald A . Jackson Published by : Wiley Stable URL : http://www.jstor.org/stable/1939574 REFERENCES Linked references are available on J. *Ecology*, v. 74, n. 8, p. 2204–2214, 1993.

JACKSON, J.E. *A user's guide to principal components*. 1st. ed. New York: John Wiley & Sons, 1991.

JACQUIN, Céline *et al.* New insight into fate and fouling behavior of bulk Dissolved Organic Matter (DOM) in a full-scale membrane bioreactor for domestic wastewater treatment. *Journal of Water Process Engineering*, v. 22, n. January, p. 94–102, 2018.

JAWAD, Jasir; HAWARI, Alaa H.; JAVAID ZAIDI, Syed. Artificial neural network modeling of wastewater treatment and desalination using membrane processes: A review. *Chemical Engineering Journal*, v. 419, n. March, p. 129540, 2021. Disponível em: <https://doi.org/10.1016/j.cej.2021.129540>.

JIANG, Jiankai *et al.* Rheological characteristics of highly concentrated anaerobic digested sludge. *Biochemical Engineering Journal*, v. 86, p. 57–61, 2014. Disponível em: <http://dx.doi.org/10.1016/j.bej.2014.03.007>.

JOHNSON, A R; WICHERN, D W. *Applied Multivariate Statistical Analysis*. [S.l: s.n.], 2014. v. 44.

JOLLIFE, Ian T.; CADIMA, Jorge. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 374, n. 2065, 2016.

JOLLIFFE, I. T. *Principal Component Analysis*. New York: Springer, 2002.

JUDD, Simon. *MBR global capacity*. Disponível em: <https://www.thembrsite.com/membrane-bioreactor-global-capacity/#:~:text=The global capacity provided by,GLD (gigalitres per day).>.

JUDD, Simon. *The MBR Book Principles and Applications of Membrane Bioreactors for Water and Wastewater Treatment*. 2nd. ed. [S.l.]: Elsevier Ltd, 2011.

JUDD, Simon. The status of industrial and municipal effluent treatment with membrane bioreactor technology. *Chemical Engineering Journal*, v. 305, p. 37–45, 2016. Disponível em: <http://dx.doi.org/10.1016/j.cej.2015.08.141>.

JUDD, Simon. The status of membrane bioreactor technology. *Trends in Biotechnology*, v. 26, n. 2, p. 109–116, 2008.

JUDD, Simon; JUDD, Claire. *The 2015 MBR Survey – The Results*.

JUDD, Simon; JUDD, Claire. *The 2016 MBR Survey – The Results*.

KAISER, H F. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, v. XX, n. 1, p. 141–151, 1960.

KAMALI, Mohammadreza *et al.* Artificial intelligence as a sustainable tool in wastewater treatment using membrane bioreactors. *Chemical Engineering Journal*, n. November, p. 128070, 2020. Disponível em: <https://doi.org/10.1016/j.cej.2020.128070>.

KAMALI, Mohammadreza *et al.* Sustainability considerations in membrane-based technologies for industrial effluents treatment. *Chemical Engineering Journal*, v. 368, n. February, p. 474–494, 2019. Disponível em: <https://doi.org/10.1016/j.cej.2019.02.075>.

KARRAY, Fatma *et al.* Pilot-scale petroleum refinery wastewaters treatment systems: Performance and microbial communities' analysis. *Process Safety and Environmental Protection*, v. 141, p. 73–82, 2020. Disponível em: <https://doi.org/10.1016/j.psep.2020.05.022>.

KASSAMBARA, A.; MUNDT, F. *Extract and Visualize the Results of Multivariate Data Analyses [R package factoextra version 1.0.7]*. . [S.l.]: Comprehensive R Archive Network (CRAN). Disponível em: <https://cran.r-project.org/package=factoextra>. , 1 abr. 2020

KASSAMBARA, Alboukadel. *"ggplot2" Based Publication Ready Plots [R package ggpubr version 0.4.0]*. . [S.l.]: Comprehensive R Archive Network (CRAN). Disponível em: <https://cran.r-project.org/package=ggpubr>. , 27 jun. 2020

KASSAMBARA, Alboukadel. *PCA - Principal Component Analysis Essentials*. Disponível em: <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/>. Acesso em: 21 jul. 2021.

KASSAMBARA, Alboukadel. *Pipe-Friendly Framework for Basic Statistical Tests [R package rstatix version 0.7.0]*. . [S.l.]: Comprehensive R Archive Network (CRAN). Disponível em: <https://cran.r-project.org/package=rstatix>. , 13 fev. 2021

KIM, Jung Hoon; GUO, Xuejun; PARK, Hung Suck. Comparison study of the effects of

temperature and free ammonia concentration on nitrification and nitrite accumulation. *Process Biochemistry*, v. 43, n. 2, p. 154–160, 2008.

KOMESLI, Okan Tarik; GÖKÇAY, Celal Ferdi. Investigation of sludge viscosity and its effects on the performance of a vacuum rotation membrane bioreactor. *Environmental Technology (United Kingdom)*, v. 35, n. 5, p. 645–652, 2014.

KOURTI, Theodora. Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, v. 19, n. 4, p. 213–246, 2005.

KOURTI, Theodora. Process analysis and abnormal situation detection: from theory to practice. *IEEE Control Systems Magazine*, n. October, 2002.

KOURTI, Theodora; MACGREGOR, John F. Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, v. 28, n. 4, p. 409–428, 1996.

KOVACS, David J *et al.* Membrane fouling prediction and uncertainty analysis using machine learning : A wastewater treatment plant case study. *Journal of Membrane Science*, v. 660, n. May, p. 120817, 2022. Disponível em: <https://doi.org/10.1016/j.memsci.2022.120817>.

KRUSKAL, William H.; WALLIS, W. Allen. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, v. 47, n. 260, p. 583–621, 1952.

KRZEMINSKI, Pawel *et al.* Membrane bioreactors – A review on recent developments in energy reduction, fouling control, novel configurations, LCA and market prospects. *Journal of Membrane Science*, v. 527, n. September 2016, p. 207–227, 2017. Disponível em: <http://dx.doi.org/10.1016/j.memsci.2016.12.010>.

KULLMANN, Christoph; LAWRENCE, Darren; COSTA, Emyr. *Focus on Aquapolo Ambiental S.A., São Paulo, Brazil.* Disponível em: <http://www.thembrsite.com/features/focus-aquapolo-ambiental-s-sao-paulo-brazil.>. Acesso em: 31 maio 2021.

LE-CLECH, Pierre; CHEN, Vicki; FANE, Tony A.G. Fouling in membrane bioreactors used in wastewater treatment. *Journal of Membrane Science*, v. 284, n. 1–2, p. 17–53, 2006.

LÊ, Sébastien *et al.* FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, v. 25, n. 1, 2008. Disponível em: <http://www.jstatsoft.org/>.

LE, Thi Minh Khanh *et al.* A multivariate approach for evaluation and monitoring of water quality in mining and minerals processing industry. *Minerals Engineering*, v. 157, n. July, p. 106582, 2020. Disponível em: <https://doi.org/10.1016/j.mineng.2020.106582>.

LEBRON, Yuri Abner Rocha *et al.* A survey on experiences in leachate treatment: Common practices, differences worldwide and future perspectives. *Journal of Environmental Management*, v. 288, n. March, 2021.

LEITE, Camila Moura Diniz Ferreira. *Avaliação do desempenho da remoção de amônia no processo de tratamento de efluentes de uma refinaria com o uso de ferramentas estatísticas*. 2021. 160 f. UNIVERSIDADE FEDERAL DE MINAS GERAIS, 2021.

LEITE, Wanderli Rogério Moreira *et al.* Monitoring and Control Improvement of Single and Two Stage Thermophilic Sludge Digestion Through Multivariate Analysis. *Waste and Biomass Valorization*, v. 9, n. 6, p. 985–994, 2018.

LEMARÉCHAL, Claude. Cauchy and the Gradient Method. *Documenta Mathematica*, v.

ISMP, p. 251–254, 2012. Disponível em: <https://www.math.uni-bielefeld.de/documenta/vol-ismp/40_lemarechal-claude.pdf>.

LIU, Ya Juan *et al.* Multivariate statistical process control (MSPC) using Raman spectroscopy for in-line culture cell monitoring considering time-varying batches synchronized with correlation optimized warping (COW). *Analytica Chimica Acta*, v. 952, n. 2017, p. 9–17, 2017. Disponível em: <http://dx.doi.org/10.1016/j.aca.2016.11.064>.

LIU, Yanchen *et al.* Study of operational conditions of simultaneous nitrification and denitrification in a Carrousel oxidation ditch for domestic wastewater treatment. *Bioresource Technology*, v. 101, n. 3, p. 901–906, 2010. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2009.09.015>.

LU, Haifeng *et al.* Brewery wastewater treatment and resource recovery through long term continuous-mode operation in pilot photosynthetic bacteria-membrane bioreactor. *Science of the Total Environment*, v. 646, p. 196–205, 2019. Disponível em: <https://doi.org/10.1016/j.scitotenv.2018.07.268>.

MACHADO, Paulo Afonso da Mata. *EcoDebate*.

MAERE, Thomas *et al.* Membrane bioreactor fouling behaviour assessment through principal component analysis and fuzzy clustering. *Water Research*, v. 46, n. 18, p. 6132–6142, 2012. Disponível em: <http://dx.doi.org/10.1016/j.watres.2012.08.027>.

MANN, H.B.; WHITNEY, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, v. 18, n. 1, p. 50–60, 1947.

MAO, Xinwei *et al.* Membrane Bioreactors for Nitrogen Removal from Wastewater: A Review. *Journal of Environmental Engineering*, v. 146, n. 5, 2020.

MENG, Fangang *et al.* Fouling in membrane bioreactors: An updated review. *Water Research*, v. 114, p. 151–180, 2017. Disponível em: <http://dx.doi.org/10.1016/j.watres.2017.02.006>.

MENG, Fangang *et al.* Morphological visualization, componential characterization and microbiological identification of membrane fouling in membrane bioreactors (MBRs). *Journal of Membrane Science*, v. 361, n. 1–2, p. 1–14, 2010. Disponível em: <http://dx.doi.org/10.1016/j.memsci.2010.06.006>.

MIQUELETO, A. P. *et al.* Influence of carbon sources and C/N ratio on EPS production in anaerobic sequencing batch biofilm reactors for wastewater treatment. *Bioresource Technology*, v. 101, n. 4, p. 1324–1330, 2010. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2009.09.026>.

MIRRA, Renata *et al.* Towards energy efficient onsite wastewater treatment. *Civil Engineering Journal (Iran)*, v. 6, n. 7, p. 1218–1226, 2020.

MIWA, Toru *et al.* Role of live cell colonization in the biofilm formation process in membrane bioreactors treating actual sewage under low organic loading rate conditions. *Applied Microbiology and Biotechnology*, v. 105, n. 4, p. 1721–1729, 2021.

MOHAN, S. Mariraj; NAGALAKSHMI, S. A review on aerobic self-forming dynamic membrane bioreactor: Formation, performance, fouling and cleaning. *Journal of Water Process Engineering*, v. 37, n. May, p. 101541, 2020. Disponível em: <https://doi.org/10.1016/j.jwpe.2020.101541>.

MONTGOMERY, Douglas C. *Introdução ao Controle Estatístico da Qualidade*. 7th. ed. [S.l.]: LTC, 2016.

MOREIRA, Bruno Rafael de Almeida *et al.* Full-scale production of high-quality wood pellets assisted by multivariate statistical process control. *Biomass and Bioenergy*, v. 151, p. 106159, 2021. Disponível em: <https://doi.org/10.1016/j.biombioe.2021.106159>.

MOSER, Priscila B. *et al.* Comparison of hybrid ultrafiltration-osmotic membrane bioreactor and conventional membrane bioreactor for oil refinery effluent treatment. *Chemical Engineering Journal*, v. 378, n. April, 2019.

NAESSENS, W. *et al.* PCA as tool for intelligent ultrafiltration for reverse osmosis seawater desalination pretreatment. *Desalination*, v. 419, n. June 2016, p. 188–196, out. 2017. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0011916416306737>.

NI, Jialing *et al.* Microbial characteristics in anaerobic membrane bioreactor treating domestic sewage: Effects of HRT and process performance. *Journal of Environmental Sciences (China)*, v. 111, p. 392–399, 2022. Disponível em: <https://doi.org/10.1016/j.jes.2021.04.022>.

NORIEGA-HEVIA, G. *et al.* Experimental sulphide inhibition calibration method in nitrification processes: A case-study. *Journal of Environmental Management*, v. 274, n. April, p. 1–7, 2020.

NOSKIEVIČOVÁ, Darja. Complex control chart interpretation. *International Journal of Engineering Business Management*, v. 5, n. 1, p. 1–7, 2013.

ODRIOZOLA, Magela *et al.* Effect of sludge characteristics on optimal required dosage of flux enhancer in anaerobic membrane bioreactors. *Journal of Membrane Science*, v. 619, p. 118776, 2021. Disponível em: <https://doi.org/10.1016/j.memsci.2020.118776>.

OLIVEIRA, Caique Prado Machado De; VIANA, Marcelo Machado; AMARAL, Míriam Cristina Santos. Coupling photocatalytic degradation using a green TiO2 catalyst to membrane bioreactor for petroleum refinery wastewater reclamation. *Journal of Water Process Engineering*, v. 34, n. October, p. 101093, 2020. Disponível em: <https://doi.org/10.1016/j.jwpe.2019.101093>.

ORDÓÑEZ, Magda Liliana Ruiz. *MULTIVARIATE STATISTICAL PROCESS CONTROL AND CASE-BASED REASONING FOR SITUATION ASSESSMENT OF Magda Liliana RUIZ ORDÓÑEZ Multivariate Statistical Process Control and Case-Based Reasoning for situation assessment of Sequencing Batch Reactors Magda Liliana*. 2008. 168 f. Universitat de Girona, 2008.

PAJOUM SHARIATI, Farshid *et al.* Biomass characterization by dielectric monitoring of viability and oxygen uptake rate measurements in a novel membrane bioreactor. *Bioresource Technology*, v. 140, p. 357–362, 2013. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2013.04.099>.

PANI, Ajaya Kumar. Non-linear process monitoring using kernel principal component analysis: A review of the basic and modified techniques with industrial applications. *Brazilian Journal of Chemical Engineering*, v. 39, n. 2, p. 327–344, 2022. Disponível em: <https://doi.org/10.1007/s43153-021-00125-2>.

PEARSON, Karl. LIII. On lines and planes of closest fit to systems of points in space . *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, v. 2, n. 11, p. 559–572, 1901.

PUGLIESE, Raffaele; REGONDI, Stefano; MARINI, Riccardo. Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Science and Management*, v. 4, n. November, p. 19–29, 2021. Disponível em: <https://doi.org/10.1016/j.dsm.2021.12.002>.

QIN, Shiyi *et al.* Fungal dynamics during anaerobic digestion of sewage sludge combined with food waste at high organic loading rates in immersed membrane bioreactors. *Bioresource Technology*, v. 335, n. May, p. 125296, 2021. Disponível em: <https://doi.org/10.1016/j.biortech.2021.125296>.

R CORE TEAM. *R: The R Project for Statistical Computing*. . [S.l: s.n.]. Disponível em: <https://www.r-project.org/>. , 2020

RABAN, Daphne R.; GORDON, Avishag. The evolution of data science and big data research: A bibliometric analysis. *Scientometrics*, v. 122, n. 3, p. 1563–1581, 2020. Disponível em: <https://doi.org/10.1007/s11192-020-03371-2>.

RODRIGUEZ-SANCHEZ, Alejandro *et al.* Influent salinity conditions affect the bacterial communities of biofouling in hybrid MBBR-MBR systems. *Journal of Water Process Engineering*, v. 30, n. June 2018, p. 100650, 2019. Disponível em: <https://doi.org/10.1016/j.jwpe.2018.07.001>.

RUSSEL, Stuart; NORVIG, Peter. *Artificial Intelligence: a modern approach*. [S.l: s.n.], 2022. v. 48. Disponível em: <www.pearsonglobaleditions.com>.

SALEEM, Mubbshir; LAVAGNOLO, Maria Cristina; SPAGNI, Alessandro. Biological hydrogen production via dark fermentation by using a side-stream dynamic membrane bioreactor: Effect of substrate concentration. *Chemical Engineering Journal*, v. 349, p. 719–727, 2018. Disponível em: <https://doi.org/10.1016/j.cej.2018.05.129>.

SALES, Rafaella Figueiredo *et al.* Multivariate statistical process control charts for batch monitoring of transesterification reactions for biodiesel production based on near-infrared spectroscopy. *Computers and Chemical Engineering*, v. 94, p. 343–353, 2016. Disponível em: <http://dx.doi.org/10.1016/j.compchemeng.2016.08.013>.

SAMBUSITI, Cecilia *et al.* Influence of HRT reduction on pilot scale flat sheet submerged membrane bioreactor (sMBR) performances for Oil&Gas wastewater treatment. *Journal of Membrane Science*, v. 594, n. September 2019, p. 117459, 2020. Disponível em: <https://doi.org/10.1016/j.memsci.2019.117459>.

SANTOS, Amanda Vitória *et al.* Improving control of membrane fouling on membrane bioreactors : A data-driven approach. *Chemical Engineering Journal*, v. 426, n. June, p. 131291, 2021.

SCHMITT, Félix *et al.* Development of artificial neural networks to predict membrane fouling in an anoxic-aerobic membrane bioreactor treating domestic wastewater. *Biochemical Engineering Journal*, v. 133, p. 47–58, 2018. Disponível em: <https://doi.org/10.1016/j.bej.2018.02.001>.

SCHWERTMAN, Neil C.; OWENS, Margaret Ann; ADNAN, Robiah. A simple more general boxplot method for identifying outliers. *Computational Statistics and Data Analysis*, v. 47, n. 1, p. 165–174, 2004.

SEPEHRI, Arsalan; SARRAFZADEH, Mohammad Hossein. Effect of nitrifiers community on fouling mitigation and nitrification efficiency in a membrane bioreactor. *Chemical Engineering*

and Processing - Process Intensification, v. 128, n. December 2017, p. 10–18, 2018.

SHAO, Zhou *et al.* Tracing the evolution of AI in the past decade and forecasting the emerging. *Expert Systems With Applications*, v. 209, n. July, p. 118221, 2022. Disponível em: <https://doi.org/10.1016/j.eswa.2022.118221>.

SHARMA, Bhavender; AHLERT, R. C. Nitrification and nitrogen removal. *Water Research*, v. 11, n. 10, p. 897–925, 1977.

SONG, Hongwei; LIU, Jinrong. Forward osmosis membrane bioreactor using Bacillus and membrane distillation hybrid system for treating dairy wastewater. *Environmental Technology (United Kingdom)*, v. 0, n. 0, p. 1–22, 2019. Disponível em: <https://doi.org/10.1080/09593330.2019.1684568>.

SONG, Kyung Guen *et al.* Characteristics of simultaneous nitrogen and phosphorus removal in a pilot-scale sequencing anoxic/anaerobic membrane bioreactor at various conditions. *Desalination*, v. 250, n. 2, p. 801–804, 2010. Disponível em: <http://dx.doi.org/10.1016/j.desal.2008.11.045>.

STEWART, Rachel D. *et al.* Pilot-scale comparison of biological nutrient removal (BNR) using intermittent and continuous ammonia-based low dissolved oxygen aeration control systems. *Water Science and Technology*, v. 85, n. 2, p. 579–590, 2022.

SUN, Sheng Peng *et al.* Effective biological nitrogen removal treatment processes for domestic wastewaters with low C/N ratios: A review. *Environmental Engineering Science*, v. 27, n. 2, p. 111–126, 2010.

TAGHEZOUIT, Bilal *et al.* Multivariate statistical monitoring of photovoltaic plant operation. *Energy Conversion and Management*, v. 205, n. November 2019, p. 112317, 2020. Disponível em: <https://doi.org/10.1016/j.enconman.2019.112317>.

THE MBR SITE. *Largest MBR plants (over 100 MLD) – Worldwide*. Disponível em: <https://www.thembrsite.com/largest-membrane-bioreactor-plants-worldwide/>.

THIEMIG, C. The importance of measuring the sludge filterability at an MBR - Introduction of a new method. *Water Science and Technology*, v. 66, n. 1, p. 9–14, 2012.

TREVISAN, Vinícius. *Comparing Robustness of MAE, MSE and RMSE | by Vinícius Trevisan | Towards Data Science*. Disponível em: <https://towardsdatascience.com/comparing-robustness-of-mae-mse-and-rmse-6d69da870828>.

TURING, A.M. COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, p. 433–460, 1950.

VASSAKIS, Konstantinos; PETRAKIS, Emmanuel; KOPANAKIS, Ioannis. Big data analytics: Applications, prospects and challenges. *Lecture Notes on Data Engineering and Communications Technologies*, v. 10, p. 3–20, 2018.

VIET, Nguyen Duc; JANG, Am. Fertilizer draw solution index in osmotic membrane bioreactor for simultaneous wastewater treatment and sustainable agriculture. *Chemosphere*, v. 296, n. January, p. 134002, 2022. Disponível em: <https://doi.org/10.1016/j.chemosphere.2022.134002>.

VON SPERLING, Marcos; CHERNICHARO, Carlos. *Biological Wastewater Treatment in Warm Climate Regions - vol. 2*. [S.l.]: IWA Publishing, 2006.

WASZCZYSZYN, Zenon. *Neural Networks in the Analysis and Design of Structures*. [S.l: s.n.], 1999. v. 404. Disponível em: <10.1007/978-3-7091-2484-0%0Ahttp://link.springer.com/10.1007/978-3-7091-2484-0>.

WESTERHUIS, Johan A.; GURDEN, Stephen P.; SMILDE, Age K. Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, v. 51, n. 1, p. 95–114, 2000.

WICKHAM, H.; BRYAN, J. *Read Excel Files [R package readxl version 1.3.1]*. . [S.l.]: Comprehensive R Archive Network (CRAN). Disponível em: <https://cran.r-project.org/package=readxl>. , 13 mar. 2019

WICKHAM, H.; SEIDEL, D. *Scale Functions for Visualization [R package scales version 1.1.1]*. . [S.l.]: Comprehensive R Archive Network (CRAN). Disponível em: <https://cran.r-project.org/package=scales>. , 11 maio 2020

WICKHAM, Hadley. *ggplot2: Elegant Graphics for Data Analysis*. New york: Springer-Verlag, 2016. Disponível em: <https://ggplot2-book.org/>.

XIAO, Kang *et al.* Current state and challenges of full-scale membrane bioreactor applications: A critical review. *Bioresource Technology*, v. 271, p. 473–481, jan. 2019. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0960852418313142>.

XU, Moke; LIANG, Yu; WU, Wenjun. Predicting honors student performance using RBFNN and PCA method. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 10179 LNCS, p. 364–375, 2017.

YAMAMOTO, Kazuo *et al. Direct Solid-Liquid Separation Using Hollow Fiber Membrane in an Activated Sludge Aeration Tank*. [S.l.]: International Association on Water Pollution Research and Control, 1989. v. 21. Disponível em: <http://dx.doi.org/10.1016/B978-1-4832-8439-2.50009-2>.

YANG, Hongqun; FAN, Maohong. Considerations for the design and operation of a membrane bioreactor. *International Journal of Biotechnology*, v. 9, n. 2, p. 188–208, 2007.

YANG, Yuan *et al.* Upflow anaerobic dynamic membrane bioreactor (AnDMBR) for wastewater treatment at room temperature and short HRTs: Process characteristics and practical applicability. *Chemical Engineering Journal*, v. 383, p. 123186, 2020. Disponível em: <https://doi.org/10.1016/j.cej.2019.123186>.

YAO, Zhibin *et al.* Using hampel identifier to eliminate profile-isolated outliers in laser vision measurement. *Journal of Sensors*, v. 2019, 2019.

YU, Dawei *et al.* Fouling analysis of membrane bioreactor treating antibiotic production wastewater at different hydraulic retention times. *Environmental Science and Pollution Research*, v. 24, n. 10, p. 9026–9035, 2017.

YURTSEVER, Adem; SAHINKAYA, Erkan; ÇINAR, Özer. Performance and foulant characteristics of an anaerobic membrane bioreactor treating real textile wastewater. *Journal of Water Process Engineering*, v. 33, n. August 2019, 2020.

YUSUF, Zakariah *et al.* Permeate flux measurement and prediction of submerged membrane bioreactor filtration process using intelligent techniques. *Jurnal Teknologi*, v. 73, n. 3, p. 85–90, 2015.

ŻABCZYŃSK, S *et al.* Removal of the high ammonia nitrogen concentration at the different

sludge ages in the membrane. n. July 2019, p. 53–57, 2006.

ZANDI, Sara *et al.* Industrial biowastes treatment using membrane bioreactors (MBRs) -a scientometric study. *Journal of Environmental Management*, v. 247, n. June, p. 462–473, 2019. Disponível em: <https://doi.org/10.1016/j.jenvman.2019.06.066>.

ZAR, J. H. *Biostatistical Analysis*. 4th. ed. New Jersey: Prentice-Hall, 1999.

ZHANG, Shu-zhe; CHEN, Shuo; JIANG, Hong. A back propagation neural network model for accurately predicting the removal efficiency of ammonia nitrogen in wastewater treatment plants using different biological processes. *Water Research*, v. 222, n. March, p. 118908, 2022. Disponível em: <https://doi.org/10.1016/j.watres.2022.118908>.

ZHAO, Jie *et al.* Real-time monitoring and fault detection of pulsed-spray fluid-bed granulation using near-infrared spectroscopy and multivariate process trajectories. *Particuology*, v. 53, p. 112–123, 2020. Disponível em: <https://doi.org/10.1016/j.partic.2020.02.003>.

ZHONG, Shifa *et al.* Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environmental Science and Technology*, v. 55, n. 19, p. 12741–12754, 2021.

APPENDIX A – R code for PCA models development

```
library(readxl)
library(factoextra)
library(FactoMineR)
library(ggplot2)
library(ggpubr)
library(rstatix)
library(scales)

### DATA IMPORT AND PCA ----------------------------------------------------------------
dados_kubota =
read_excel("M:/AMANDA/Doutorado/Dados/MBRpiloto/dados_kubota.xlsx",
col_types = c("text", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric"))

res.pca = PCA(dados_kubota[,-1], scale.unit = TRUE, ncp =
ncol(dados_kubota)-1, graph = FALSE)

### EIGENVALUES ------------------------------------------------------------------------------------------------------------------------
eig_val = get_eigenvalue(res.pca)
graph_eigen = fviz_eig(res.pca, addlabels = TRUE) #cutoff 14%
ggpar(graph_eigen,
      title = "Principal Component Analysis (PCA)",
      subtitle = "Eingenvalues",
      xlab = "Principal Components", ylab = "Percentage of explained
variation",
      ggtheme = theme_bw() + theme(text = element_text(size = 25)))

### VARIABLES --------------------------------------------------------------------------------------------------------------------------
res.var = get_pca_var(res.pca)
res.var$coord
res.var$cos2
res.var$contrib

graph_var = fviz_pca_var(res.pca, axes = c(1,2),
geom.var = c("arrow","text"), arrowsize = 1, labelsize = 7,
col.var = "cos2", gradient.cols = c("blue", "yellow", "red"),
col.circle = "black",
repel = TRUE,
legend.title = "Explained variation")
ggpar(graph_var,
      title = "Loading of variables",
      xlab = "PC1 (27.50%)", ylab = "PC2 (24.76%)",
      ggtheme = theme_bw() + theme(text = element_text(size = 20)) +
theme(legend.position = "top", legend.justification = "right"))

### INDIVIDUALS ------------------------------------------------------------------------------------------------------------------------
res.ind = get_pca_ind(res.pca)
res.ind$coord
res.ind$cos2
res.ind$contrib

graph_ind = fviz_pca_ind(res.pca, axes = c(1,2),
                         geom.ind = c("point"), pointsize = 3,
                         col.ind = dados_kubota$Year, palette ="Dark2",
                         addEllipses = TRUE, ellipse.level=0.95,
```

```
                        legend.title = "Year")

ggpar(graph_ind,
      title = "Principal Component Analysis",
      subtitle = "Scores",
      xlab = "PC1 (27.50%)", ylab = "PC2 (24.76%)",
      ggtheme = theme_bw() + theme(text = element_text(size = 20)) +
theme(legend.position = "top", legend.justification = "right"))

### BIPLOT ---------------------------------------------------------------------------
graph_biplot = fviz_pca_biplot(res.pca, axes = c(1,2),
geom.ind = "point", pointsize = 2,
col.ind = dados_kubota$Year, palette = "Dark2",
addEllipses = TRUE, ellipse.level = 0.95,
geom.var = c("arrow","text"), arrowsize = 1, labelsize = 7,
col.var = "black",
repel = TRUE,
legend.title = "Year")

ggpar(graph_biplot,
      title = "Principal Component Analysis",
      subtitle = "Biplot",
      xlab = "PC1 (27.50%)", ylab = "PC2 (24.76%)",
      ggtheme = theme_bw() + theme(text = element_text(size = 20)) +
theme(legend.position = "top", legend.justification = "right"))

### BOXPLOTS -------------------------------------------------------------------------
DWC_res.kruskal = kruskal_test(SDWC ~ Year, data = dados_kubota)
DWC_dunn = dunn_test(SDWC ~ Year, data = dados_kubota, p.adjust.method
= "bonferroni")

DWC_dunn = DWC_dunn %>% add_xy_position(x = "Year")
DWC = ggboxplot(dados_kubota, x = "Year", y = "SDWC") +
      stat_pvalue_manual(DWC_dunn, hide.ns = TRUE) +
      stat_compare_means()
graph_DWC = ggpar(DWC,
title = "Sequential days without cleaning",
xlab = "Year", ylab = "SDWC",
ggtheme = theme_bw()) + theme(text = element_text(size = 18))

Perm_res.kruskal = kruskal_test(Perm ~ Year, data = dados_kubota)
Perm_dunn = dunn_test(Perm ~ Year, data = dados_kubota, p.adjust.method
= "bonferroni")

Perm_dunn = Perm_dunn %>% add_xy_position(x = "Year")
Perm = ggboxplot(dados_kubota, x = "Year", y = "Perm") +
       stat_pvalue_manual(Perm_dunn, hide.ns = TRUE) +
       stat_compare_means()
graph_Perm = ggpar(Perm,
title =  "Membrane  Permeability",
xlab = "Year", ylab = "Permeability",
ggtheme = theme_bw() + theme(text = element_text(size = 18)))

COD_res.kruskal = kruskal_test(COD ~ Year, data = dados_kubota)
COD_dunn = dunn_test(COD ~ Year, data = dados_kubota, p.adjust.method =
"bonferroni")

COD_dunn = COD_dunn %>% add_xy_position(x = "Year")
```

```
COD = ggboxplot(dados_kubota, x = "Year", y = "COD") +
      stat_pvalue_manual(COD_dunn, hide.ns = TRUE) +
      stat_compare_means()
graph_COD = ggpar(COD,
title = "COD",
xlab = "Year", ylab = "COD",
ggtheme = theme_bw()) + theme(text = element_text(size = 18))

ggarrange(graph_DWC, graph_Perm, graph_COD,
          labels = c("A", "B", "C"),
          ncol = 3, nrow = 1)

### PREDICTIVE MODEL ----------------------------------------------------------------------------------------------------
dados_kubota_pred = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/dados_kubota_PCA_II.xlsx",
col_types = c("text", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric"))

set.seed(37645)
dummy_sep = rbinom(nrow(dados_kubota_pred), 1, 0.5)
real = dados_kubota_pred[dummy_sep == 0, ]
pred = dados_kubota_pred[dummy_sep == 1, ]

res.pca_II = PCA(real[,-1], scale.unit = TRUE, ncp = ncol(real)-1,
graph = FALSE)

eig_val = get_eigenvalue(res.pca_II)
graph_eigen = fviz_eig(res.pca_II, addlabels = TRUE)#cutoff 14%
eig =ggpar(graph_eigen,
xlab = "Principal Components", ylab = "Percentage of explained
variation",
ggtheme = theme_bw() + theme(text = element_text(size = 15)))

graph_rep3 = fviz_cos2(res.pca_II, choice = "var", axes = c(1,2,3)) +
scale_y_continuous(labels = percent)
tres = ggpar(graph_rep3,
title = "Explained variation of variable",
xlab = "Variables", ylab = "Percentage of explained variation",
ggtheme = theme_bw() + theme(text = element_text(size = 15)) +
theme(axis.text.x = element_text(size = 15)))

ggarrange(eig, tres,
          labels = c("A", "B"),
          ncol = 2, nrow = 2)

#Q² e R² PARAMETERS ----------------------------------------------------------------------
X = scale(real[,-1])
X_quad = X*X
inertia = sum(X_quad)
media = sapply(real[,-1], mean)
desvio = sapply(real[,-1], sd)

cor = cor(real[,-1])
U = eigen(cor)$vector

scores = X%*%U

Xest = scores[,1:3]%*%t(U[,1:3])
```

```
erro = X - Xest
erro_quad = erro*erro

ress = sum(erro_quad)
Rquad = 1 - ress/inertia #R²
X2 = scale(pred[,-1], center = media, scale = desvio)
scores2 = X2%*%U

Xest2 = scores2[,1:3]%*%t(U[,1:3])
erro2 = X2 - Xest2
erro_quad2 = erro2*erro2

press  =  sum(erro_quad2)
Qquad = 1 - press/inertia #Q²

RMSECV = sqrt(press/(nrow(erro_quad2)*ncol(erro_quad2)))

#BIPLOTS WITH PROJECTIONS ----------------------------------------------------------
indI = fviz_pca_ind(res.pca_II, axes = c(1,2),
                    geom.ind = "point", pointsize = 2,
                    col.ind = real$Year, palette = "Dark2",
                    addEllipses = TRUE, ellipse.level = 0.95,
                    repel = TRUE,
                    legend.title = "Year")

predI = fviz_add(indI, scores2, axes = c(1,2), shape=25,
color="gray21")
CPI = ggpar(predI,
            title = "Projections on PC1 and PC2",
            xlab = "PC1 (29.4%)", ylab = "PC2 (24.2%)",
            ggtheme = theme_bw() + theme(text = element_text(size =
18)) + theme(legend.position = "top", legend.justification = "right"))

indII = fviz_pca_ind(res.pca_II, axes = c(1,3),
                     geom.ind = "point", pointsize = 2,
                     col.ind = real$Year, palette = "Dark2",
                     addEllipses = TRUE, ellipse.level = 0.95,
                     repel = TRUE,
                     legend.title = "Year")

predII = fviz_add(indII, scores2, axes = c(1,3), shape=25,
color="gray21")
CPII = ggpar(predII,
             title = "Projections on PC1 and PC3",
             xlab = "PC1 (29.4%)", ylab = "PC3 (17.0%)",
             ggtheme = theme_bw() + theme(text = element_text(size =
18)) + theme(legend.position = "top", legend.justification = "right"))

indIII = fviz_pca_ind(res.pca_II, axes = c(2,3),
                      geom.ind = "point", pointsize = 2,
                      col.ind = real$Year, palette = "Dark2",
                      addEllipses = TRUE, ellipse.level = 0.95,
                      repel = TRUE,
                      legend.title = "Year")

predIII = fviz_add(indIII, scores2, axes = c(2,3), shape=25,
color="gray21")
CPIII = ggpar(predIII,
```

```
              title = "Projections on PC2 and PC3",
              xlab = "PC2 (24.2%)", ylab = "PC3 (17.0%)",
              ggtheme = theme_bw() + theme(text = element_text(size =
18)) + theme(legend.position = "top", legend.justification = "right"))

ggarrange(CPI,CPII,CPIII,
          labels = c("A", "B", "C"),
          nrow = 1, ncol = 3)

#CROSS VALIDATION ----------------------------------------------------------------------------------------------------
cross_val =
read_excel("M:/AMANDA/Doutorado/Dados/MBRpiloto/cross_val.xlsx",
col_types = c("text", "numeric", "numeric", "text"))

graph_cross_val = ggplot(cross_val, aes(x=CP, y=Valor, fill=Parameter))
+ geom_bar(stat="identity", position=position_dodge()) +
geom_line(aes(x=CP, y=Perc, group=Parameter, linetype=Parameter)) +
geom_point(aes(x=CP, y=Perc, shape=Parameter), size=3)
ggpar(graph_cross_val,
title = "Principal Component Analysis",
subtitle = "Q² and R² values and improvement of the model",
xlab = "Number of components kept", ylab = "Q², R² and improvement(%)",
ggtheme = theme_bw() + theme(text = element_text(size = 20)) +
theme(legend.position = "top", legend.justification = "right"))

#VARIABLES PREDICTION ------------------------------------------------------------------------
previsao = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/previsao.xlsx")

filt = ggplot(previsao, aes(x=Filt, y=V1)) + geom_point() +
geom_line(aes(x=Filt, y=Filt, group=1))
a = ggpar(filt,
          title = "Sludge Filterability",
          xlab = "Real value", ylab = "Predicted value",
          ggtheme = theme_bw() + theme(text = element_text(size = 15))
+ theme(axis.title = element_text(size = 12)) +
theme(legend.position = "top", legend.justification = "right"))

VS = ggplot(previsao, aes(x=MLVSS, y=V2)) + geom_point() +
geom_line(aes(x=MLVSS, y=MLVSS, group=1))
b = ggpar(VS,
          subtitle = "MLVSS",
          xlab = "Real value", ylab = "Predicted value",
          ggtheme = theme_bw() + theme(text = element_text(size = 15))
+ theme(axis.title = element_text(size = 12)) +
theme(legend.position = "top", legend.justification = "right"))

pH = ggplot(previsao, aes(x=pH, y=V3)) + geom_point() +
geom_line(aes(x=pH, y=pH, group=1))
c = ggpar(pH,
          subtitle = "pH",
          xlab = "Real value", ylab = "Predicted value",
          ggtheme = theme_bw() + theme(text = element_text(size = 15))
+ theme(axis.title = element_text(size = 12)) +
theme(legend.position = "top", legend.justification = "right"))

COD = ggplot(previsao, aes(x=COD, y=V4)) + geom_point() +
geom_line(aes(x=COD, y=COD, group=1))
```

```
d = ggpar(COD,
          subtitle = "COD",
          xlab = "Real value", ylab = "Predicted value",
          ggtheme = theme_bw() + theme(text = element_text(size = 15))
+ theme(axis.title = element_text(size = 12)) +
theme(legend.position = "top", legend.justification = "right"))

temp = ggplot(previsao, aes(x=Temp, y=V5)) + geom_point() +
geom_line(aes(x=Temp, y=Temp, group=1))
e = ggpar(temp,
          subtitle = "Temperature",
          xlab = "Real value", ylab = "Predicted value",
          ggtheme = theme_bw() + theme(text = element_text(size = 15))
+ theme(axis.title = element_text(size = 12)) +
theme(legend.position = "top", legend.justification = "right"))

DWC = ggplot(previsao, aes(x=SDWC, y=V6)) + geom_point() +
geom_line(aes(x=SDWC, y=SDWC, group=1))
f = ggpar(DWC,
          subtitle = "Sequential days without cleaning",
          xlab = "Real value", ylab = "Predicted value",
          ggtheme = theme_bw() + theme(text = element_text(size = 15))
+ theme(axis.title = element_text(size = 12)) +
theme(legend.position = "top", legend.justification = "right"))

Perm = ggplot(previsao, aes(x=Perm, y=V7)) + geom_point() +
geom_line(aes(x=Perm, y=Perm, group=1))
g = ggpar(Perm,
          subtitle = "Membrane Permeability",
          xlab = "Real value", ylab = "Predicted value",
          ggtheme = theme_bw() + theme(text = element_text(size = 15))
+ theme(axis.title = element_text(size = 12)) +
theme(legend.position = "top", legend.justification = "right"))

ggarrange(a, b, c, d, e, f, g,
          labels = c("A", "B", "C", "D", "E", "F", "G"),
          ncol = 2, nrow = 4)
```

APPENDIX B – R code for MSPC models development

```
library(readxl)
library(ggpubr)
library(ggplot2)
library(scales)
library(rstatix)
library(FactoMineR)
library(factoextra)

### PHASE I ----------------------------------------------------------------------------------------------------------------
dados_kubota_faseI =
read_excel("M:/AMANDA/Doutorado/Dados/MBRpiloto/dados_kubota_faseI.xlsx",
col_types = c("numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric"))

#NUMBER OF COMPONENTS ----------------------------------------------------------------------
res.pca.mspc = PCA(dados_kubota_faseI, scale.unit = TRUE, ncp =
ncol(dados_kubota_faseI), graph = FALSE)

eig_val.mspc = get_eigenvalue(res.pca.mspc)
graph_eigen_mspc = fviz_eig(res.pca.mspc, addlabels = TRUE) #cutoff 14%
ggpar(graph_eigen_mspc,
      title = "Principal Component Analysis (PCA)",
      subtitle = "Eingenvalues",
      xlab = "Principal Components", ylab = "Percentage of explained
variation",
      ggtheme = theme_bw() + theme(text = element_text(size = 25)))

graph_rep3.mspc = fviz_cos2(res.pca.mspc, choice = "var", axes =
c(1,2,3))
ggpar(graph_rep3.mspc,
                  title = "Explained variation of variable",
                  subtitle = "Three components",
                  xlab = "Variables", ylab = "Percentage of explained
variation",
                  ggtheme = theme_bw() + theme(text = element_text(size
= 21)) + theme(axis.text.x = element_text(size = 21)))

#PCA MODEL ----------------------------------------------------------------------------------
media = sapply(dados_kubota_faseI, mean)
desvio = sapply(dados_kubota_faseI, sd)

cor = cor(dados_kubota_faseI)
U = eigen(cor)$vectors
lamb = eigen(cor)$values
lambida = as.data.frame(lamb)

X = scale(dados_kubota_faseI)
scores = X%*%U

X_quad = X*X
inertia = sum(X_quad)

#T² STATISTIC ------------------------------------------------------------------------------------------------------------
scores_quad = scores[,1:3]*scores[,1:3]
t2 = mapply(`/`, data.frame(scores_quad), lambida[1:3,])
T2 = rowSums(t2)
```

```
#Q STATISTIC ------------------------------------------------------------------------------
Xest = scores[,1:3]%*%t(U[,1:3])
erro = X - Xest
erro_quad = erro*erro
Q = rowSums(erro_quad)

RESS = sum(erro_quad)
R_quad = 1 - RESS/inertia #R²

### PHASE II -------------------------------------------------------------------------------
dados_kubota_faseII = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/dados_kubota_faseII.xlsx",
col_types = c("numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric"))

X2 = scale(dados_kubota_faseII, center = media, scale = desvio)
scores2 = X2%*%U

#T² STATISTIC ------------------------------------------------------------------------------
scores_quad2 = scores2[,1:3]*scores2[,1:3]
t2_2 = mapply(`/`, data.frame(scores_quad2), lambida[1:3,])
T2_2 = rowSums(t2_2)

#Q STATISTIC ------------------------------------------------------------------------------
Xest2 = scores2[,1:3]%*%t(U[,1:3])
erro2 = X2 - Xest2
erro_quad2 = erro2*erro2
Q2 = rowSums(erro_quad2)

PRESS = sum(erro_quad2)
Q_quad = 1 - PRESS/inertia #Q²

### CONTROL CHARTS -------------------------------------------------------------------------
T2_final = c(T2, T2_2)
Q_final = c(Q, Q2)
quantile(T2, 0.95)
quantile(Q, 0.95)

a = as.data.frame(seq(from = 1, to = 624, by = 1))
b = as.data.frame(T2_final)
c = as.data.frame(seq(quantile(T2, 0.95), quantile(T2, 0.95),
length.out=624))

r = cbind(a,b,c)
names(r)[1] = "Sample"
names(r)[3] = "Percentile"

graphT2 = ggplot(r[339:624,], aes(x=Sample)) +
geom_point(aes(y=T2_final, color = cut(T2_final,c(-Inf,quantile(T2,
0.95),Inf))), size = 2) +
geom_line(aes(y=Percentile, color = "95% Percentile"), size = 1) +
scale_color_manual(values = c("black", "red", "red"))
A = ggpar(graphT2,
      title = "Multivariate Statistical Process Control (MSPC)",
      subtitle = "Hotelling Control Chart",
      xlab = "Sample", ylab = "T²",
      ggtheme = theme_bw() + theme(legend.position = "none")+
```

```
theme(text = element_text(size = 15)))

d = as.data.frame(Q_final)
e = as.data.frame(seq(quantile(Q, 0.95), quantile(Q, 0.95), length.out=624))

s = cbind(a,d,e)
names(s)[1] = "Sample"
names(s)[3] = "Percentile"

graphQ = ggplot(s[339:624,], aes(x=Sample)) + geom_point(aes(y=Q_final,
color = cut(Q_final,c(-Inf,quantile(Q, 0.95),Inf))), size = 2) +
geom_line(aes(y=Percentile, color="95% Percentile"), size = 1) +
scale_color_manual(values = c("black", "red", "red"))
B = ggpar(graphQ,
      subtitle = "Q Control Chart",
      xlab = "Sample", ylab = "Q",
      ggtheme = theme_bw()+ theme(legend.position = "none")
      + theme(text = element_text(size = 15)))

#PERMEABILITY AND FILTERABILITY PLOT ----------------------------------
dados_kubota_perm = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/dados_kubota_perm_fasesIeII.xlsx",
col_types = c("numeric", "numeric","numeric"))

graph_perm = ggplot(dados_kubota_perm[339:624,], aes(x=Sample)) +
geom_point(aes(y=Perm, color = cut(Perm,c(-Inf,100,150, Inf))), size =
2) +
geom_line(aes(y=Lim, color="100 L/m².h.bar"), size = 1) + ylim(0,1000)
+scale_color_manual(values = c("red", "orange", "black", "red"))
C = ggpar(graph_perm,
      subtitle = "Permeability [L/(h.m².bar)]",
      xlab = "Sample", ylab = "Permeability",
      ggtheme = theme_bw() + theme(legend.position = "none") +
      theme(text = element_text(size = 15)) + theme(axis.title.y =
element_text(size = 12)))

dados_kubota_filt = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/dados_kubota_filt_fasesIeII.xlsx",
col_types = c("numeric", "numeric"))

graph_filt = ggplot(dados_kubota_filt[339:624,], aes(x=Sample)) +
geom_point(aes(y=Filt, color = cut(Filt,c(-Inf,10,Inf))), size = 2) +
scale_color_manual(values = c("red", "black"))
D = ggpar(graph_filt,
          subtitle = "Filterability [mL/5min]",
          xlab = "Sample", ylab = "Filterability",
          ggtheme = theme_bw() + theme(legend.position = "none") +
          theme(text = element_text(size = 15)) + theme(axis.title.y =
element_text(size = 12)))

ggarrange(A, B, C, D,
          labels = c("A", "B", "C","D"),
          ncol = 1, nrow = 4)

### CONTRIBUTION PLOTS ----------------------------------------------------------------------------------------------------
Qobservacao = as.data.frame(Q2)
cont = mapply(`/`, data.frame(erro_quad2), Qobservacao)
```

```
#OUT-OF-CONTROL OPERATION ------------------------------------------------------------
dados_kubota_cont = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/dados_kubota_cont.xlsx",
col_types = c("text","numeric","text"))

graph_cont = ggplot(dados_kubota_cont, aes(x=Sample, y=Contribution,
fill=Variable)) +
scale_fill_brewer(palette="Set2") + scale_y_continuous(labels =
percent) + geom_bar(stat="identity")
Z = ggpar(graph_cont,
          subtitle = "Out-of-control Operation",
          xlab = "Sample", ylab = "Contribution",
          ggtheme = theme_bw() + theme(legend.position = "top",
legend.justification = "right") +
          theme(text = element_text(size = 25))+ theme(axis.text.x =
element_text(size = 12, angle = 90)))

#ALARMING OPERATION ------------------------------------------------------------------
dados_kubota_cont_false = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/dados_kubota_cont_false.xlsx",
col_types = c("text","numeric","text"))

graph_cont_false = ggplot(dados_kubota_cont_false, aes(x=Sample,
y=Contribution, fill=Variable)) +
scale_fill_brewer(palette="Set2") + scale_y_continuous(labels =
percent) + geom_bar(stat="identity")
W = ggpar(graph_cont_false,
          subtitle = "Alarming Operation",
          xlab = "Sample", ylab = "Contribution",
          ggtheme = theme_bw() + theme(legend.position = "top",
legend.justification = "right") +
          theme(text = element_text(size = 25)) + theme(axis.text.x =
element_text(size = 12, angle = 90)))

ggarrange(Z, W,
          labels = c("A", "B"),
          ncol = 2, nrow = 1)

### HYPOTHESES TEST ------------------------------------------------------------------
#OUT-OF-CONTROL OPERATION ------------------------------------------------------------
dados_kubota_cont_MW = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/dados_kubota_cont_MW.xlsx")

temp = wilcox_test(Temp ~ Group, data = dados_kubota_cont_MW, exact =
FALSE)
temp = temp %>% add_xy_position(x = "Group")
temp = ggboxplot(dados_kubota_cont_MW, x = "Group", y = "Temp") +
                 stat_pvalue_manual(temp, hide.ns = TRUE)
graph_temp = ggpar(temp,
             title = "Temperature",
             ylab = "Temperature (°C)",
             ggtheme = theme_bw() + theme(text = element_text(size =
18)) + theme(axis.title.x = element_blank()))

perm = wilcox_test(Perm ~ Group, data = dados_kubota_cont_MW, exact = FALSE)
perm = perm %>% add_xy_position(x = "Group")
perm = ggboxplot(dados_kubota_cont_MW, x = "Group", y = "Perm") +
       stat_pvalue_manual(perm, hide.ns = TRUE)
```

```
graph_perm = ggpar(perm,
             title = "Membrane permeability",
             ylab = "Permability (L/h.m².bar)",
             ggtheme = theme_bw() + theme(text = element_text(size =
18)) + theme(axis.title.x = element_blank()))

dwc = wilcox_test(SDWC ~ Group, data = dados_kubota_cont_MW, exact =
FALSE)
dwc = dwc %>% add_xy_position(x = "Group")
dwc = ggboxplot(dados_kubota_cont_MW, x = "Group", y = "SDWC") +
      stat_pvalue_manual(dwc, hide.ns = TRUE)
graph_dwc = ggpar(dwc,
            title = "Sequential days without cleaning",
            ylab = "Sequential days without cleaning",
            ggtheme = theme_bw() + theme(text = element_text(size =
18)) + theme(axis.title.x = element_blank()))

#ALARMING OPERATION ------------------------------------------------------------------
dados_kubota_cont_false_MW = read_excel("M:/AMANDA/Doutorado/Dados/MBR
piloto/dados_kubota_cont_false_MW.xlsx")

temp2 = wilcox_test(Temp ~ Group, data = dados_kubota_cont_false_MW,
exact = FALSE)
temp2 = temp2 %>% add_xy_position(x = "Group")
temp2 = ggboxplot(dados_kubota_cont_false_MW, x = "Group", y = "Temp")
+ stat_pvalue_manual(temp2, hide.ns = TRUE)
graph_temp2 = ggpar(temp2,
                    title = "Temperature",
                    ylab = "Temperature (°C)",
                    ggtheme = theme_bw() + theme(text =
element_text(size = 18)) + theme(axis.title.x = element_blank()))

perm2 = wilcox_test(Perm ~ Group, data = dados_kubota_cont_false_MW,
exact = FALSE)
perm2 = perm2 %>% add_xy_position(x = "Group")
perm2 = ggboxplot(dados_kubota_cont_false_MW, x = "Group", y = "Perm")
+ stat_pvalue_manual(perm2, hide.ns = TRUE)
graph_perm2 = ggpar(perm2,
                    title = "Membrane permeability",
                    ylab = "Permability (L/h.m².bar)",
                    ggtheme = theme_bw() + theme(text =
element_text(size = 18)) + theme(axis.title.x = element_blank()))

dwc2 = wilcox_test(SDWC ~ Group, data = dados_kubota_cont_false_MW,
exact = FALSE)
dwc2 = dwc2 %>% add_xy_position(x = "Group")
dwc2 = ggboxplot(dados_kubota_cont_false_MW, x = "Group", y = "SDWC") +
       stat_pvalue_manual(dwc2, hide.ns = TRUE)
graph_dwc2 = ggpar(dwc2,
                   title = "Sequential days without cleaning",
                   ylab = "Sequential days without cleaning",
                   ggtheme = theme_bw() + theme(text =
element_text(size = 18)) + theme(axis.title.x = element_blank()))

ggarrange(graph_perm, graph_dwc, graph_temp,graph_perm2, graph_dwc2,
graph_temp2,
          labels = c("A", "B", "C", "D", "E", "F"),
          ncol = 3, nrow = 2)
```

APPENDIX C – R code for ANN models development

```
library(h2o)
library(readxl)
library(factoextra)
library(FactoMineR)
library(ggplot2)
library(ggpubr)

citation("h2o")

h2o.init()

file_path    <-    "C:/Users/Moises/Documents/AMANDA/Doutorado/Dados/MBR
Piloto/02-Dados-Amonia/dados_kubota_ann.csv"
dados <- h2o.importFile(file_path, header = TRUE, sep = ";", dec = ".")

dados_split = h2o.splitFrame(dados, ratios = c(0.7))
dados_treino = dados_split[[1]]
dados_teste = dados_split[[2]]


### Configuracao do modelo ------------------------------------------------------------
-----------

x_cols                                                                    <-
c("filt","MLVSS","pH","ammonia","sulphide","OG","phosp","fCOD","pCOD","
remCOD","temp","SRT","perm")
y_col <- "removal"

hidden_layers <- c(12,7)

actv_func <- "Rectifier"

epochs_used <- 5000

### ------------------------------------------------------------------------------------------------------------------------------------------------------
-----------

model <- h2o.deeplearning(
  x = x_cols,
  y = y_col,
  training_frame = dados_treino,
  validation_frame = dados_teste,
  activation = actv_func,
  hidden = hidden_layers,
  epochs = epochs_used
)

plot(model)

Resultado <- h2o.performance(model)
h2o.mse(Resultado)
h2o.mae(Resultado)

### Análise de sensibilidade  ----------------------------------------------------------------------------------------------------------------------
---------------------------------
```

```
h2o.varimp_plot(model)
h2o.varimp(model)


dados_kubota                                                                    =
read_excel("C:/Users/Moises/Documents/AMANDA/DOUTORADO/Dados/MBR
piloto/02-Dados-Amonia/dados_kubota_pca.xlsx")

res.pca = PCA(dados_kubota, scale.unit = TRUE, ncp = ncol(dados_kubota)-
1, graph = FALSE)

graph_eigen = fviz_eig(res.pca, addlabels = TRUE) #cutoff 14%
ggpar(graph_eigen,
      title = "Principal Component Analysis (PCA)",
      subtitle = "Eingenvalues",
      xlab = "Principal Components", ylab = "Percentage of explained
variation",
      ggtheme = theme_bw() + theme(text = element_text(size = 25)))
get_eigenvalue(res.pca)

res.var = get_pca_var(res.pca)
res.var$coord
res.var$cos2
res.var$contrib

graph_var = fviz_pca_var(res.pca, axes = c(2,3),
                         geom.var = c("arrow","text"), arrowsize = 1,
labelsize = 5,
                         col.var = "cos2", gradient.cols = c("blue",
"yellow", "red"),
                         col.circle = "black",
                         repel = TRUE,
                         legend.title = "Explained variation")
ggpar(graph_var,
      xlab = "PC2 (17.8%)", ylab = "PC3 (9.8%)",
      ggtheme = theme_bw() + theme(text = element_text(size = 18)) +
theme(legend.position = "top", legend.justification = "right"))


### Previsao ----------------------------------------------------------------------------------------------------------------
-----------

previsao <- h2o.predict(model, newdata = dados_previsao)
previsao

devtools::install_github("rstudio/tensorflow")

library(tensorflow)
install_tensorflow()

install.packages("keras")
install.packages("mlbench")
install.packages("neuralnet")

install_keras()

library(keras)
```

```
library(mlbench)
library(dplyr)
library(magrittr)
library(neuralnet)
library(readxl)
library(scales)
library(ggplot2)
library(ggpubr)

citation("keras")
citation("neuralnet")

data = read_excel("C:/Users/Moises/Documents/AMANDA/DOUTORADO/Dados/MBR
piloto/02-Dados-Amonia/dados_kubota_ann.xlsx")

str(data)
data %<>% mutate_if(is.factor, as.numeric)

n                                                                    <-
neuralnet(removal~filt+MLVSS+pH+ammonia+sulphide+OG+phosp+fCOD+pCOD+rem
COD+temp+SRT+perm,
               data = data,
               hidden = c(8,5,3),
               linear.output = F,
               lifesign = 'full',
               rep=1)

plot(n,col.hidden = 'black',
     col.hidden.synapse = 'black',
     show.weights  =  T,
     information = T,
     fill = 'grey')

data <- as.matrix(data)
dimnames(data) <- NULL

set.seed(123)
ind <- sample(2, nrow(data), replace = T, prob = c(.7, .3))
training <- data[ind==1,1:13]
test <- data[ind==2, 1:13]
trainingtarget <- data[ind==1, 14]
testtarget <- data[ind==2, 14]

m <- colMeans(training)
s <- apply(training, 2, sd)
training <- scale(training, center = m, scale = s)
test <- scale(test, center = m, scale = s)

model <- keras_model_sequential()
model %>%
  layer_dense(units = 100, activation = 'relu', input_shape = c(13)) %>%
  layer_dropout(rate=0.4)  %>%
  layer_dense(units = 50, activation = 'relu') %>%
  layer_dropout(rate=0.2)  %>%
  layer_dense(units = 1)

model %>% compile(loss = 'mse',
                  optimizer = 'rmsprop',
```

```
                  metrics = 'mae')

mymodel <- model %>%
  fit(training,trainingtarget,
      epochs = 100,
      batch_size = 32,
      validation_split = 0.2)

model %>% evaluate(test, testtarget)
pred <- model %>% predict(test)

corr = ggplot(testtarget, pred, pch=19, xlim=c(0,50), ylim=c(0,50))
ggpar(corr,
      xlab = "Real value", ylab = "Predicted value",
      ggtheme = theme_bw() + theme(text = element_text(size = 25)))
```

APPENDIX D – Test of Dunn Results

Table D.1 - Multiple comparison test of Dunn, with Bonferroni correction, for the variable sequential days without cleaning during the monitoring years

Group 1	Group 2	n_1	n_2	Statistic	p-value	p-value adjusted
Year 01	Year 02	103	185	-2.114	3.45E-02	3.45E-01
Year 01	Year 03	103	154	-2.339	1.93E-02	1.93E-01
Year 01	Year 04	103	197	8.079	6.52E-16	6.52E-15
Year 01	Year 05	103	89	13.04	7.50E-39	7.50E-38
Year 02	Year 03	185	154	-0.347	7.29E-01	1.00E+00
Year 02	Year 04	185	197	12.13	6.97E-34	6.97E-33
Year 02	Year 05	185	89	16.64	3.52E-62	3.52E-61
Year 03	Year 04	154	197	11.90	1.17E-32	1.17E-31
Year 03	Year 05	154	89	16.41	1.73E-60	1.73E-59
Year 04	Year 05	197	89	7.081	1.43E-12	1.43E-11

Table D.2 - Multiple comparison test of Dunn, with Bonferroni correction, for the variable membrane permeability during the monitoring years

Group 1	Group 2	n1	n2	Statistic	p-value	p-value adjusted
Year 01	Year 02	103	185	-6.096	1.09E-09	1.09E-08
Year 01	Year 03	103	154	-0.150	8.81E-01	1.00E+00
Year 01	Year 04	103	197	-4.468	7.91E-06	7.91E-05
Year 01	Year 05	103	89	-12.02	2.80E-33	2.80E-32
Year 02	Year 03	185	154	6.696	2.15E-11	2.15E-10
Year 02	Year 04	185	197	2.014	4.40E-02	4.40E-01
Year 02	Year 05	185	89	-7.675	1.66E-14	1.66E-13
Year 03	Year 04	154	197	-4.873	1.10E-06	1.10E-05
Year 03	Year 05	154	89	-12.92	3.44E-38	3.44E-37
Year 04	Year 05	197	89	-9.367	7.50E-21	7.50E-20

Table D.3 - Multiple comparison test of Dunn, with Bonferroni correction, for the variable COD of MBR feed during the monitoring years

Group 1	Group 2	n_1	n_2	Statistic	p-value	p-value adjusted
Year 01	Year 02	103	185	-5.345	9.02E-08	9.02E-07
Year 01	Year 03	103	154	-12.97	1.81E-38	1.81E-37
Year 01	Year 04	103	197	-11.47	1.93E-30	1.93E-29
Year 01	Year 05	103	89	-13.04	7.05E-39	7.05E-38
Year 02	Year 03	185	154	-9.110	8.24E-20	8.24E-19
Year 02	Year 04	185	197	-7.200	6.01E-13	6.01E-12
Year 02	Year 05	185	89	-9.537	1.47E-21	1.47E-20
Year 03	Year 04	154	197	2.385	1.71E-02	1.71E-01
Year 03	Year 05	154	89	-1.777	7.56E-02	7.56E-01
Year 04	Year 05	197	89	-3.861	1.13E-04	1.13E-03

APPENDIX E – Wilcoxon-Mann-Whitney Test Results for Membrane Fouling

Table E.1 - Wilcoxon-Mann-Whitney statistical test to compare samples of in-control and out-of-control operations

Group 1	Group 2	n_1	n_2	Variable	Statistic	p-value
In-control	Out-of-control	338	43	Temp	9511	0.000824
In-control	Out-of-control	338	43	Perm	14534	1.22e-26
In-control	Out-of-control	338	43	SDWC	0	1.21e-26

Table E.2 - Wilcoxon-Mann-Whitney statistical test to compare samples of in-control and alarming operations

Group 1	Group 2	n_1	n_2	Variable	Statistic	p-value
In-control	Alarming	338	86	Temp	21302	1.73e-11
In-control	Alarming	338	86	Perm	2168	3.64e-34
In-control	Alarming	338	86	SDWC	29068	1.53e-46

APPENDIX F – Wilcoxon-Mann-Whitney Test Results for Ammonia Removal

Table F1. Wilcoxon-Mann-Whitney statistical test to compare samples of in-control and out-of-control operations

Group 1	Group 2	n1	n2	Variable	Statistic	p-value
In-control	Out-of-control	272	46	Removal	12512	2.07e-27
In-control	Out-of-control	272	46	Temp	5976	1.22e-4
In-control	Out-of-control	272	46	SRT	9913	2.68e-13
In-control	Out-of-control	272	46	fCOD	9722	1.86e-9