



Development of Digital Twins for Water Treatment Systems

Topalian, Sebastian Olivier Nymann

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Topalian, S. O. N. (2024). *Development of Digital Twins for Water Treatment Systems*. Technical University of Denmark.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Development of Digital Twins for Water Treatment Systems

Sebastian Olivier Nymann Topalian
PhD Thesis

Australian Centre for Water and Environmental Biotechnology
University of Queensland

Process and Systems Engineering Centre
Department of Chemical and Biochemical Engineering
Technical University of Denmark



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

Preface

The work for this Ph.D. thesis was conducted at the Process and Systems Engineering Centre (PROSYS) at the Department of Chemical and Biochemical Engineering at the Technical University of Denmark from the 1st of October 2020 to the 30th of September 2023. The thesis relied heavily on an external collaboration between PROSYS and the Environmental Operations department at Novozymes A/S, as well as the Australian Centre for Water and Environmental Biotechnology (ACWEB).

The project was supervised by Senior Researcher Xavier Flores-Alsina, Professor Damien J. Batstone and Professor Krist V. Gernaey. Furthermore, Senior Global Technology Manager Kasper Kjellberg and Researcher Pedram Ramin have all contributed to the main chapters to the same extent as the supervisors.

The thesis is written as a collection of three manuscripts that are either published or submitted for publication:

- 1) Topalian, S. O. N., Ramin, P., Kjellberg, K., Kazadi Mbamba, C., Batstone, D. J., Gernaey, K. V. & Flores-Alsina, X., 2023. A data analytics pipeline to optimize polymer dose strategy in a semi-continuous multi-feed dewatering system, *Journal of Water Process Engineering*. 55, 15, 104048.
- 2) Topalian, S. O. N., Nazemzadeh, N., Malacara-Becerra, A., Mansouri, S. S., Kjellberg, K., Batstone, D. J., Gernaey, K. V., Flores-Alsina, X. & Ramin, P., Transfer Learning for Quantitative Image Analysis of Biosolids. Submitted to *Water Research*.
- 3) Topalian, S. O. N., Pokhrel, A., Malacara-Becerra, A., Rehn, K., Kjellberg, K., Ramin, P., Mansouri, S. S., Batstone, D. J., Gernaey, K. V. & Flores-Alsina, X., Fifty Shades of Foam: Validation of a camera-based total suspended solids soft-sensor for industrial dewatering of biosolids. Submitted to *Journal of Water Process Engineering*.

The references and figure numbering throughout the thesis resets for each chapter, and the references used outside of the paper are listed at the end of the thesis.

Additional Contributions

The following contributions were submitted to academic conferences:

Watermatex 2023, Quebec, Canada, 24-27th of September, 2023. Oral presentation: Transforming Biosolids: Linear Multimodal Modelling for Improved FTIR Based Soft Sensors.

ENBIS 2023, Valencia, Spain, 10-14th of September, 2023. Oral presentation: dPCA: A Python Library for Dynamic Principal Component Analysis. <https://waterboy96.github.io/dPCA/>

ecoSTP 2023, Girona, Spain, 26-29th of June, 2023. Poster: Transfer Learning for Quantitative Image Analysis of Biosolids

PSE 2021+, Kyoto, Japan, 19-23rd of June, 2022. Poster: Forecasting Operational Conditions: A case-study from dewatering of biomass at an industrial wastewater treatment plant

The following Master theses was carried out under the supervision of the Ph.D. student:

Würtz, A., 2023. Optimization of an Industrial Granulation Process using Dynamic Image Analysis. DTU Department of Chemical and Biochemical Engineering & Novozymes A/S

- Pokhrel, A., 2023. Data Driven Control Strategy for Dewatering of Inactivated Bio-solids in Centrifugal Decanters. DTU Department of Chemical and Biochemical Engineering & Novozymes A/S
- Malacara-Becerra, A., 2022. Instituto Tecnológico y de Estudios Superiores de Monterrey & Novozymes A/S

Acknowledgements

Writing acknowledgements has always been a challenging task for me, as there are so many wonderful people to thank, and I am sure that I will forget someone... As such figure 1 depicts a novel methodology for disseminating acknowledgements.

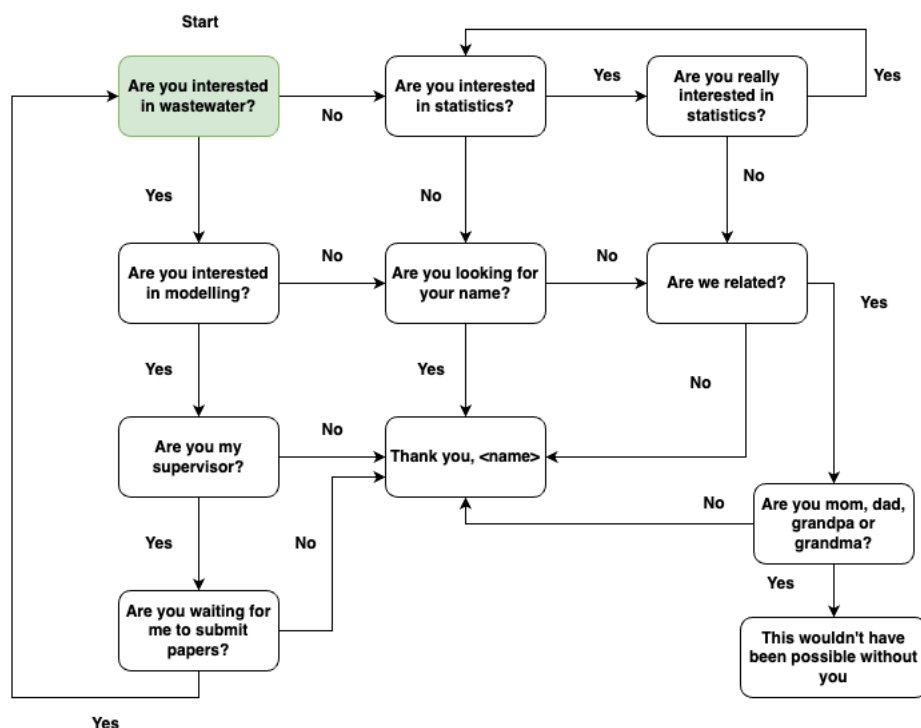


Figure 1 - Novel Acknowledgement Procedure. Start at the top left and answer the queries to find out whether you are acknowledged by the author.

The above flowchart represents a state-of-the-art system to ensure a robust acknowledgement section and leaves the disseminator with peace of mind to allocate substantial resources to a few key people of interest, which will be demonstrated in the following paragraphs.

Before I started my PhD, I already had four great supervisors for my master thesis: Krist, Xavier, Pedram and Murat. Pedram and Murat were replaced with Damien as official supervisors, but they still stuck around, and I was left with an abundance of people to spar with. What a treat. Thank you, Krist for always being encouraging and letting me explore my academic interests outside of wastewater modelling, which I can only say is rare when I look at the academic landscape of paper pushers. Thank you, Xavier, for teaching me the nuances of the academic landscape, how to be a grade A pessimist, your great comradery and impeccable taste in TV shows. Thank you, Pedram for everything you have done for me the past 4 years from our meeting at DHI, to our special course, master thesis, and everything we have done during my PhD. Of all the people mentioned on these few pages you moved the needle the most, and for that I am forever grateful. Thank you, Murat, for teaching me how to think like a statistician and all our talks on analysis of dynamical systems and modelling. Thank you, Kasper from Novozymes for introducing me to the world of industrial wastewater and how to separate wishful academic thinking from what is applicable in a real plant. Thank you, Damien, for your guidance during my stay at ACWEB, I have seldomly met

a person who is as candid as you and who showed me how to balance mathematical tools with experimental data and fundamental knowledge.

Finally, a big thank you to my past students, peers I met at conferences, colleagues past and present at ACWEB and PROSYS, friends, and last but not least, family.

Abstract

The development of digital twins for water treatment systems is a complicated task with multiple solution strategies. In this thesis three different strategies are adopted to try and connect empirical information to the task of industrial dewatering of stabilized biosolids. First, a big data approach is developed to try and utilize existing infrastructure and information from sensors and production schedules from nearby units as well as upstream production information from a biotechnological production company to forecast polymer consumption at the dewatering facility. The approach enables the analysis of how upstream categorical information such as production information propagates through the system and impacts the dewatering operation. Second, an experimentalist approach is developed where samples are collected and analyzed by combining computer vision algorithms and quantitative image analysis. The computer vision approach aptly distinguishes between particles with different morphologies, which is an improvement in terms of interpretability compared to the conventional approach where only particle size descriptors are utilized. Finally, a third approach is developed, that of the control engineer, where the validation of a partial-degassing vessel and a color-camera are used to monitor total suspended solids in a situation where traditional measurement equipment has failed before. The three different approaches are discussed, and a suggestion is made to focus on ensuring the robustness of the online reject monitoring setup and developing a monitoring setup for the cake to effectively close the mass balance around the dewatering unit.

Dansk Resumé

Udviklingen af digitale tvillinger til vandrensnings systemer er en kompliceret opgave med adskillige løsningsmuligheder. I denne afhandling bliver tre forskellige løsningstrategier undersøgt for at prøve at på kvantitativ vis forbinde empiri med driften af en industriel afvandingsproces. For det første, afprøves en moderne "big data" tilgang der udnytter eksisterende sensor infrastruktur og tilgængelig produktionsinformation for at prøve at forudsige doseringen af polymer i afvandingsanlægget. Tilgangen viser på lovende vis hvordan man kan koble produktionsinformation efter dens effekt har propageret gennem systemet og påvirket afvandingsanlægget. For det andet, afprøves en eksperimentel tilgang hvor prøver fra afvandingsanlægget opsamles og benyttes til kvantitativ billedanalyse koblet med kunstig intelligens til at skelne mellem partikler med forskellige morfologier. For det tredje, afprøves en kontrolingeniørs tilgang hvor en online måler bestående af en afgangstank og et farvekamera måler suspenderede partikler i en situation hvor traditionelt måleudstyr ikke før har set succes. De tre forskellige tilgange diskuteres, og et forslag gives til at fokusere på sidst nævnte løsningsforslag med at udvikle et målesystem, samt at udvikle endnu et målesystem til at overvåge den tunge fase fra afvandingen for at lukke massebalancen omkring afvandingsprocessen.

Table of Contents

I.	Preface	III
II.	Acknowledgements	V
III.	Abstract	VII
IV.	Dansk Resumé	VIII
1.	Introduction	1
2.	Paper 1	6
3.	Paper 2	42
4.	Paper 3	67
5.	Discussion	82
6.	Conclusion	84
7.	References	85

1. Introduction

1.1 Decision making through modelling

Modern-day decision-making relies heavily on the use of heuristics that have been derived from an imperfect mixture of empirical observations and theoretical foundations over the course of many years. It goes without saying that decision-making is a central part of everyday life, however the road towards improved decision-making is often non-trivial. One means towards improved decision-making is mathematical modelling which comes in a variety of flavours; however, all eventually refer to one of two philosophical approaches, the deterministic or stochastic approach. While these two approaches are often used to debate the existence of free will, we often use them in our everyday lives to synthesize knowledge from the systems we interact with. The deterministic approach looks for strict relationships between cause and effect, and is always interpretable, however when many deterministic relationships aggregate into a complex system, we often lose track of the interpretable properties as the sheer number of possible combinations of cause and effects explode, and systematic deterministic analysis becomes infeasible. This is where the second paradigm of stochastic thinking comes into play, i.e., for any given cause there may be several effects or no effect at all, which leads to the field of statistical modelling. Regardless of the approach, both are built on top of empirical observations, and the selected approach is then used to describe a relationship between sets of empirical observations. The knowledge derived from the empirical observations through either deterministic or statistical models form the basis of subsequent decision making. This pipeline forms the first two pillars of instrumentation, control, and automation [1]. Over the years the idea of combining a modelling approach with empirical collections evolved into the idea of the digital twin: an exact digital replica of a physical system that runs online. While the novel term was a lot catchier there was little to distinguish it from the ideals we had before the notion of the digital twin, i.e., some knowledge representation that enables decision making.

1.2 Industrial Wastewater Modelling

Within wastewater modelling several tools have been developed mainly with an offset in the context of municipal wastewater treatment. While municipal facilities are subject to variation due to dynamics across many different timescales such as the diurnal pattern of morning and evening increases in household consumption, industrial wastewater treatment plant influent variations are caused by dynamics belonging to the corresponding production facility producing the wastewater. In this thesis the emphasis will be on answering questions pertaining to dewatering of biosolids stemming from a biotechnological production company which is well-known for operating fermentation processes in batch mode, and the subsequent product purification is the final step before entering the domain of the industrial wastewater treatment plant. This presents several unique challenges for each industrial wastewater treatment plant on how to integrate knowledge

from the production with knowledge from the wastewater treatment plant to achieve better operation whether it be in terms of profitability or environmental impact. One piece of information that distinguishes industrial wastewater modelling from municipal wastewater modelling is the presence of production batch information, such as a product name or microorganism name and these are typically represented as categorical variables while conventional mechanistic wastewater treatment models deal only with continuous variables. In practice most industrial wastewater treatment plants were built after the production, but as the production dictates the subsequent waste treatment it becomes prudent to ask questions related to the upstream production information to move from sequential to integrated process design.

1.3 Dewatering

Dewatering is a common unit operation at wastewater treatment plants where a suspension of particles in water is separated into a light fraction, the reject, and a heavy fraction, the cake. Many technologies exist for dewatering from dewatering lagoons, primary clarifiers to belt filters and centrifugal decanters. The different technologies differ mainly in space requirements, dewatering efficiency and throughput, and the decision on which technology is suitable for a specific application depends on the circumstances. The centrifugal decanter is the subject of focus in this thesis where upon questions regarding the development of a digital twin are posed. The centrifugal decanter sacrifices points in efficiency for higher throughputs and lower space requirements [2]. A centrifugal decanter separates the solids from their suspension by continuously rotating the suspension at which the gravitational forces applied to the particles will accelerate them compared to the liquid if particle density is larger than the density of the liquid.

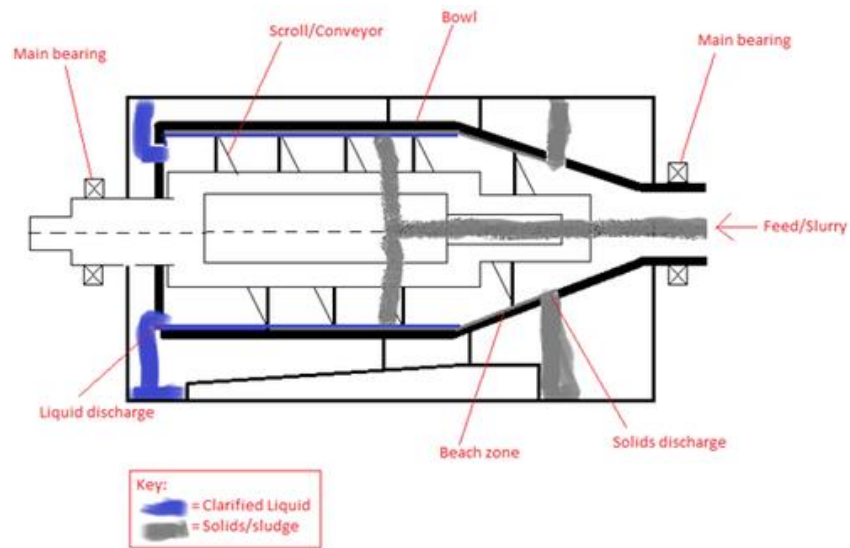


Figure 2 - Conceptual illustration of a decanter centrifuge. The feed is fed inside the scroll which rotates rapidly applying a large centrifugal force which separates the particles from the liquid based on the discrepancy in their densities producing a liquid discharge (the reject) and a heavy discharge (the cake). Illustration from Wikipedia [3].

The success of centrifugal dewatering of biological materials largely depends on the ability to form flocs that can withstand the centrifugal force applied within the machine, and this is a function of surface chemistry, particle morphology and the interplay with added chemicals such as flocculant and coagulant. While flocculant and coagulant dosages are easy to record, acquisition of data related to surface chemistry and particle morphology is less widely adopted and automated in wastewater treatment, and implementation of models that relate said properties with theoretical foundations are still missing within the wastewater literature to the knowledge of the author. As such, statistical modelling approaches seemed more suitable to address current wastewater dewatering problems, but the question remains how to obtain an adequate foundation of information upon which to build the statistical models.

1.4 Data Acquisition Strategies

For every problem a wide variety of approaches exist, and the chosen angle of attack typically depends on the education and experience of the person owning the problem. On one end of the spectrum the data scientist sits representing Industry 4.0, and on the other end of the spectrum the control engineer sits with an experimentalist situated in between.

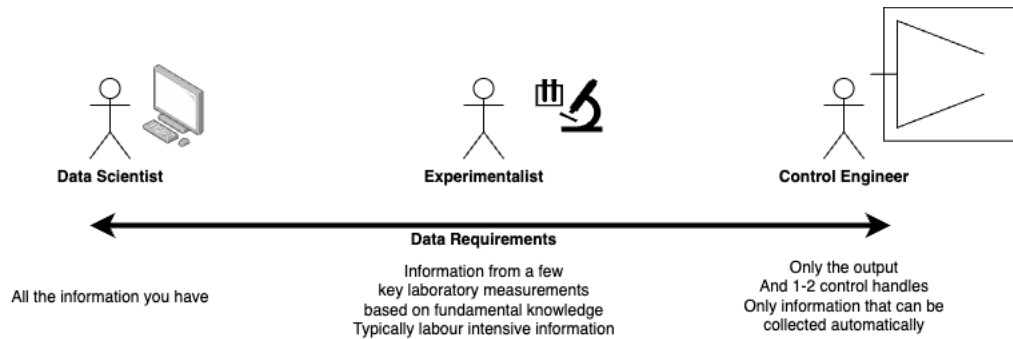


Figure 3 – A conceptual illustration of the three approaches to solving a problem with a model. On the left hand side the data scientist requiring as much prior information as possible to automatically mine through. In the middle the scientist or experimentalist, who will carry out simple and laborious experiments that have a tie-in with the fundamental principles. On the right hand side the control engineer who only requires a small set of information, namely the output and the most impactful control handles.

1.4.1 Approach of the data scientist

The “modern” approach with the rise of artificial intelligence and machine learning leans on the requirement of having as much data available as possible, and coincidentally with the rise of Industry 4.0 many plants are now storing all the information that they collect. However, that does not guarantee that they store any relevant information related to the specific problem at hand. One appealing trait of the approach is that you may uncover hidden relationships that were unknown to plant operators before, and the approach already caters to the inclusion of upstream information which can shine light on the root cause when difficult operational periods arise.

1.4.2 Approach of the experimentalist

This is the “old” school approach of due diligently collecting samples and performing relevant laboratory tests based on fundamental knowledge. This approach has been used by most academics, engineers and scientists and offers insight into obtaining a more detailed mechanistic understanding of a system, which the other two do not. One downside of the approach is that it originated for closed systems which the real world is often not, and as such it may be hard to extrapolate laboratory results to full scale, and the experimental procedures are often not within reach of automation and can only be performed when the relevant personnel is present.

1.4.3 Approach of the control engineer

The control engineer sits somewhere between the scientists of old and the data scientist of new but are still highly relevant in the engineering problem solving landscape. The control engineers, unlike the scientists, care less about the mechanisms of the system and more about solving the problem. There is one paramount prerequisite for the control engineer and that is having access and preferably easy access to measuring the controlled variable, which is often but not always equivalent to the key performance indicator of the process. Aside from frequent measurements of the controlled variable, the control engineers also require access to the manipulated variables with a large effect on the controlled variable. The largest downside to this approach is that it may not be possible to obtain reliable measurements of the controlled variable of interest, which renders the approach useless unless that variable can be estimated based on other available measurement data (soft sensor concept). However if the key variable of interest is available, then the upside lies in the parsimonious nature of the strategy, which often leads to an automatable solution.

1.5 Objectives of the thesis

The objective of the thesis is to try and approach the development of a digital twin for an industrial dewatering system from the approaches described above and evaluate the strengths and weaknesses of each. While there is no set way to do each of the approaches and the work within this thesis is by no means exhaustive, it aims to shed light on how each could be performed, so that future engineers and scientists may save deliberation time when approaching an unsolved problem with multiple angles of attack.

2. Paper 1

A data analytics pipeline to optimise polymer dose strategy in a semi-continuous multi-feed dewatering system

The following paper was accepted for publishing in the Journal of Water Process Engineering on the 10th of July 2023, and made available online on the 4th of August 2023.

The article is available at: <https://doi.org/10.1016/j.jwpe.2023.104048>

A data analytics pipeline to optimise polymer dose strategy in a semi-continuous multi-feed dewatering system

Sebastian O. N. Topalian¹, Pedram Ramin¹, Kasper Kjellberg², Christian Kazadi Mbamba³, Damien J. Batstone³, Krist V. Gernaey¹, Xavier Flores-Alsina^{1*}.

¹ Process and Systems Engineering Centre (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark. Building 228 A, 2800, Kgs. Lyngby, Denmark.

² Novozymes A/S, Hallas Alle 1, DK- 4400, Kalundborg, Denmark.

³ Australian Center for Water and Environmental Biotechnology, The University of Queensland, 4 Gehrman Laboratories Building, Research Rd, St Lucia QLD 4067, Brisbane, Australia.

***Corresponding author:**

Xavier Flores-Alsina,

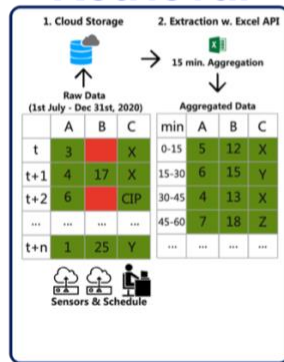
Email: xfa@kt.dtu.dk

Address: Technical University of Denmark, Department of Chemical and Biochemical Engineering, Søltofts Plads, Building 227 (Postal address: Building 228A), 2800 Kgs. Lyngby, Denmark

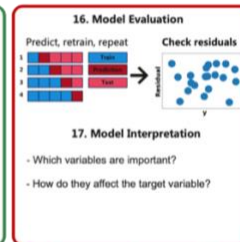
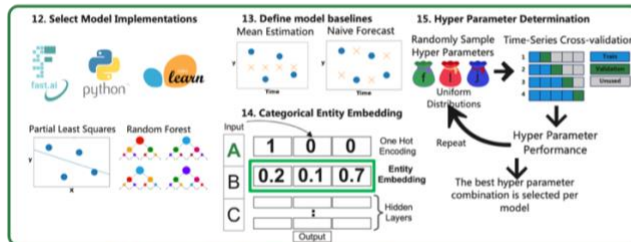
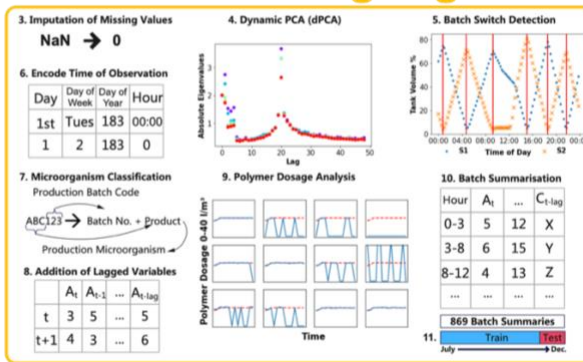
Abstract

Automated application of scientific data analytics is a critical element to enhance industrial productivity and minimize human errors. In this study a dynamic polymer dosing strategy is developed for an industrial centrifuge system receiving highly variable rates and quality of feed solids from a wastewater treatment plant. A four-section methodology is developed and tested at full-scale containing (i) data extraction for data collection and aggregation; (ii) Data wrangling including delay analysis and batch analysis; (iii) model development with predictive ability; (iv) model analysis for model evaluation and interpretation. A partial least squares (PLS) and a random forest (RF) model were validated and used to predict polymer dosages. In contrast to PLS the RF model was capable of learning structural information and describe which products related to increases or decreases in polymer dosage. An additional analysis investigating the impact of different product codes on the polymer dosage is presented, revealing that certain products generally lead to consistent changes in the polymer dosage. The proposed approach could potentially save operators 3-6 hours a day in terms of time spent on manually adjusting polymer dosages. The presented methodology for data pipelining and analysis has a generic nature and can be easily exported to other case studies.

Retrieval



Data Wrangling



Model Development

Analysis

Graphical Abstract – Overview of the specific tools applied.

Highlights

1. The methodology can be used to predict polymer dosages from upstream information
2. Categorical entity embeddings leads to model improvement and enables product analysis
3. Dynamic PCA can be used to identify process delays that match process understanding

Keywords

Data; Wastewater; Modelling; Dewatering; Flocculation

Nomenclature

U(lower bound, upper bound) – Uniform distribution with lower and upper bound

RF – Random Forest

WWTP – Wastewater Treatment Plant

industrial WWTP (iWWTP)

Prod. – Production

Machine Set. – Machine Settings

API - application-programming interface (API)

PACM – Partial Autocorrelation Matrix

Mon – Monday

Tue – Tuesday

Wed – Wednesday

Thu– Thursday

Fri – Friday

Sat – Saturday

Sun - Sunday

PLS – Partial Least Squares

dPCA – Dynamic Principal Component Analysis

RMSE – Root Mean Square Error

R^2 – Coefficient of determination

CCP – Cost Complexity Pruning

1. Introduction

A wastewater treatment plant (WWTP) plays a significant role in modern society by protecting local water environments from pollution and ensuring drinking water quality [1]. Most WWTPs use the activated sludge system, which includes a consortium of microorganisms, for removal of pollutants such as carbon, nitrogen and phosphorus [2]. Modern treatment systems are expanding to resource recovery and include biosolids management [3]. Biomass used in WWTP systems, or “activated sludge”, grows on the available nutrients in the water stream, and is disposed of as biosolids. Biosolids management accounts for approximately half the cost of wastewater treatment [4]. Biosolids management includes a sequence of unit operations including handling, conditioning, and disposal. A key step before the biosolids disposal is dewatering. These biosolids in the dewatering units are separated into a concentrated stream (cake) and a diluted stream (centrate). The cake is then disposed of and the centrate is directed/recycled back into the WWTP for further treatment [5].

The dewatering step can be performed through several alternative processes such as belt filtration, evaporation and/or centrifugation. However, for biosolids, centrifugal decanters are typically utilized due to their high throughput while maintaining a sufficient degree of separation [6]. Flocculants and coagulants are added to aid the separation process [7]. This helps the particles to form larger flocs resulting in easier separation in centrifugal decanters. Dosage determination can be a challenge in industrial systems due to several factors including (i) the dynamics of the upstream production of biosolids due to different production schedule, (ii) variations in the properties of the biosolids, (iii) variation in optimal operational settings of the decanter such as volumetric and mass loading, and centrifuge speeds [8]. To determine a sufficient chemical dosage at a specific operational window, commonly two main approaches are used in literature: experimental procedures, and modelling.

Laboratory tests aim to determine the optimal dosage by controlled mixing of biosolids and flocculant in a defined test. Peeters et al. [9] reported laboratory scale was representative of large-scale decanters. However, Gnisty et al. [10] found that while gravity belt thickeners and filter press units could be approximated in the lab, decanter centrifuges could not, since lab tests have the tendency to overestimate the dryness of biosolids compared to their full-scale counterparts. Indeed, the forces inside the decanter will cause flocs to break at a different rate than in a laboratory setup, leading to mixed results when comparing laboratory tests with full-scale performance [11]. In addition, where input solids vary, laboratory methods must be used to continually optimize polymer dose and type. This is required in different work shifts (morning, afternoon, and night). Jar testing is generally used in these cases, where the operators add polymer and polyaluminium chloride at different rates to a fixed volume of sludge and gently shake the jar by hand to judge the formation and quality of flocs. This has been identified by Gnisty et al [10] as a weak indicator of performance in full-scale operation.

Modelling has also been used to enhance the understanding of the relationship between biosolids characteristics, polymer dosage and separation efficiency in decanters. Mechanistic decanter models have been developed based on mass and energy balances in the decanter [12, 13, 14]. These models consider fundamental sedimentation mechanisms. Particle trajectory depends on a few parameters including the particle size and density. Particle separation fundamentals are used to derive separation efficiency [13]. The model by Wang et al. [13] assumes that the solids fraction and flow rate of the feed slurry is constant, which is often not the case in industrial settings. Application of this model identified that solids recovery is impacted by feed rate, solids concentration and particle size distribution [15]. The mentioned mechanistic models require the density and size of the particles as an input. Although the density and size of the particles can be measured outside of the decanter, for biosolids these properties change once the weak structure of the biosolids is exposed to the shear force in the decanter. Moreover, these properties are influenced by the addition of polymer. Leung [14] developed a mechanistic model for inferring the in-situ floc size given the recovery efficiency. However this requires plant data and hence is not a suitable approach for optimization purposes. Instead, the operators require a methodology that can utilize information obtained before the decanter to predict the optimal polymer dosage to achieve sufficient separation.

Data-driven methods, on the other hand, are not based on fundamental understanding of underlying mechanisms of the system. Rather they rely on often hidden correlations and patterns between process variables [16]. Extracting these correlations can be useful for prediction of hard-to-measure parameters such as recovery efficiencies [15]. Compared to mechanistic approaches, these models are easier to establish and faster to simulate [18]. In recent years, there has been an increasing number of studies using data-driven approaches due to their higher prediction capabilities compared to more traditional mechanistic counterparts. O. Bello et al. [19] and Jayaweera & Aziz [20] could predict surface charge and pH for a coagulation process in a water treatment plant. Hong et al. [21] predicted both dewatering performance and dosage optimization at a wastewater treatment plant. More recently, hybrid models (also called gray box models) which combine a mechanistic and data-driven approach have been also suggested [22]. In this hybrid approach, a data-driven model can compensate for uncaptured physical behavior in the mechanistic model. In a model by Menesklou et al. [23], mechanistic and artificial neural network models were used with a parallel structure with an interpolative use to predict the solids mass fraction of centrate and cake.

Although being valuable scientific contributions, previous studies did not handle some of the challenges that industrial WWTPs are facing, namely: (1) the special dynamics of dewatering systems (commonly operating in batch mode) combined with possible continuous and semi-batch operation; (2) the effect that residence time may have (and propagated) on process data, which complicates establishing causality; and, (3) the complex nature of upstream information (production schemes) that might determine the characteristics of the biosolids stream that will have to be handled.

To address these limitations this study presents a data-driven approach to assess biosolids dewatering in industrial decanters based on upstream and downstream production information. The proposed analytical pipeline includes a set of tools to identify operational modes, analyze process delays and forecast polymer dosages combining both numerical and categorical data. The plant under study is the largest industrial WWTP in Northern Europe, in Kalundborg (Denmark), and treats the wastewater produced at two biotech companies (Novozymes and Novo Nordisk). Compared to existing literature this paper goes in-depth with the process description, data-retrieval, data-analysis, model development, model selection, model analysis and solely uses data that do not have to be collected in the laboratory. In summary the objectives of the paper are: (1) demonstrate data treatment and pipelining in a biosolids dewatering unit operation, (2) develop a data-driven model procedure in a transparent fashion for forecasting and optimization of flocculants dosage, (3) show-case statistical techniques for analyzing plant-wide WWTP data, (4) deliver a data-driven decision support tool that can provide operators with suggested polymer dosages.

The paper is structured as follows: First the flow diagram of the plant under study is presented, detailing units, availability / type of sensor as well as the specifics of the inactivation and subsequent dewatering process. Next, a general overview of the methodology is given detailing the sections/steps of which it is comprised. Finally, the proposed approach is applied to a “real” case study where production data is used to come up with a model predicting polymer dosages-based plant data. This method is intended to alleviate the work of process operators.

2. Industrial site under study

2.1 Plant description

Figure 1 shows a layout of the dewatering line at the industrial WWTP (iWWTP) as well as the majority of the data signals that are investigated in this study. For a comprehensive description of the entire plant the reader is referred to the work of Monje et al. [24, 25].

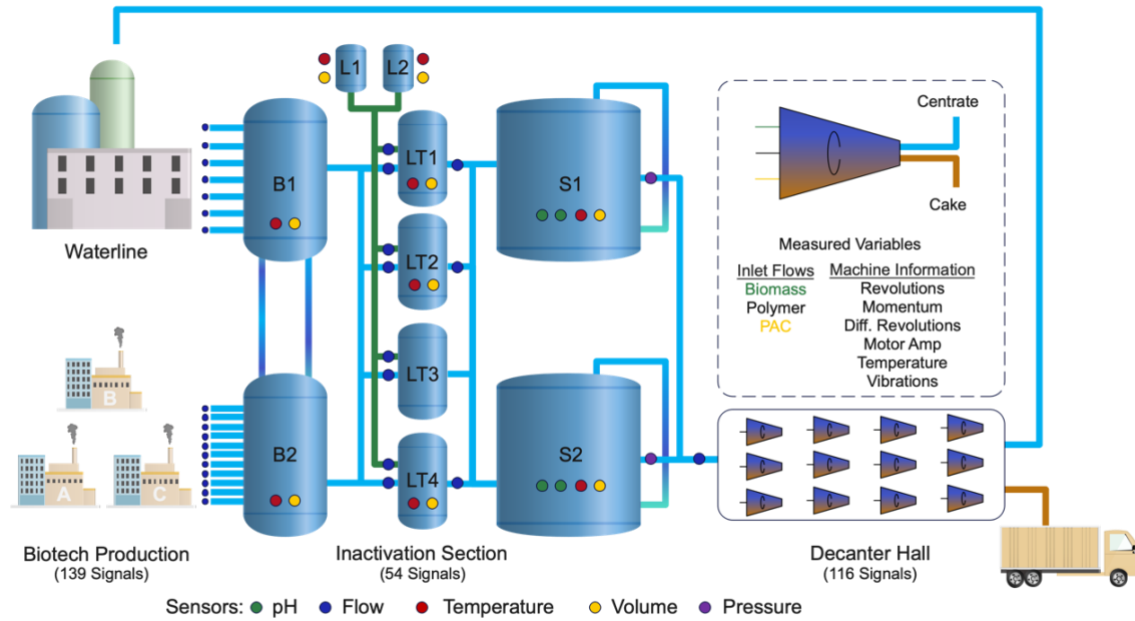


Figure 4 - Conceptual overview of the dewatering line at the iWWTP under investigation. Spent biomass and waste activated sludge are initially stored in the two buffer tanks B1 and B2, then lime solution from L1 and L2 tanks is added in tanks LT1-LT4 to inactivate the biomass before being stored in one of two storage tanks S1 and S2. The outflow is then transferred semi-continuously to decanters. The small, colorized dots denote the placement of online sensors for the inactivation section.

Briefly, the first two tanks, B1 and B2, act as buffer tanks and they receive spent biomass from three different production sites (A, B & C) as well as waste activated biosolids from the waterline. The next four tanks, LT1-LT4, act as inactivation tanks where the output of B1 and B2 is stabilized by addition of a lime solution to comply with regulations regarding the handling of genetically modified organisms [26]. The lime solution is supplied from the two tanks L1 and L2. After the addition of lime solution, the inactivated biosolids are sent to one of two storage tanks, S1 and S2, that operate semi-continuously (i.e. while one is emptying the other one is filling). The two storage tanks are recirculated and supply a series of 12 decanters in parallel operation with the feed for the dewatering operation. These centrifugal decanters are used to dewater the inactivated biomass resulting in a centrate that is fed upstream in the waterline and a cake that is disposed of. On average 250 m³ of inactivated biosolids are treated every hour with operation running around the clock.

2.2 Process description

To achieve a satisfactory dewatering of the inactivated biosolids, coagulants and flocculants are added. In this case, the plant operators are responsible for manipulating machine settings as well as the coagulant and flocculant dosage to achieve a sufficient separation in the centrifugal decanters. The operators mainly adjust the dosage of flocculant, and they can assess the chemical requirements by performing shake-flask tests or by adopting a trial-and-error based approach. This procedure takes 2-3 hours and requires their presence 24-7. As a result, the performance of the

decanters relies heavily on this manual process optimization by the operators. Process optimization is deemed necessary every time there is a switch from one storage tank to the other. This switch happens 3 to 5 times a day, and the content within each batch varies depending on the upstream production, which can vary substantially. The primary focus in this study is the optimization of flocculant dosage, (in this case an organic polymer) since it was found to be the main control handle for improving dewatering and the one most frequently adjusted by the plant operators. The proposed optimization method is based on finding patterns in operational data and building data-driven forecasting models to predict the polymer dosage and analyzing the developed models for the impact production sites A, B and C have on the predicted polymer dosage.

2.3 Data description

The data used in this study consists of three categorical variables each relating to one of the upstream production schedules for production sites A-C as well as 309 continuous variables from online sensors located in the production site and the dewatering line depicted in Figure 1. Out of the 309 continuous variables 182 (59%) relate to upstream information obtained before storage units S1 and S2, and 127 (41%) relate to downstream information obtained at and after the storage units.

Table 1 describes the number of signals collected from each section of the plant as well as the sensor type categorized by either flow, pH, temperature, batch ID, volume, pressure, machine settings or other, denoting various sensor types such as conductivity or input from operators such as chemical ratios. For the inactivation section (all sections except the production sites and the decanters), the sensor locations are shown in Figure 1. A list of the available sensors, where sensor names have been anonymized, is available in the supplementary information.

Table 1 – Overview of the sensor signals by type and section. Machine Set. Describes machine settings primarily for the decanters.

	Prod. A	Prod. B	Prod. C	Buffer	Lime	Storage	Decanters	Total (≈ %)
Flow	59	18	19	18	11	1	28	154 (50%)
pH	7	1	2	2	0	4	0	16 (5%)
Temp.	3	0	1	0	5	2	24	35 (11%)
Batch ID	1	1	1	0	0	0	0	3 (1%)
Volume	1	1	0	2	5	2	0	11(4%)
Pressure	0	0	0	0	0	2	3	5(2%)
Machine Set.	0	1	0	0	0	0	60	61(20%)
Other	13	9	1	0	0	0	1	24(7%)
Total (≈ %)	84 (27%)	31 (10%)	24(8%)	22 (7%)	21 (7%)	11 (4%)	116(8%)	

3. Methods

The proposed procedure comprises seventeen steps organized into four sections: (I) data retrieval, (II) data wrangling & (III) model development and (IV) analysis. A schematic representation of the proposed methodology is depicted in Figure 2.

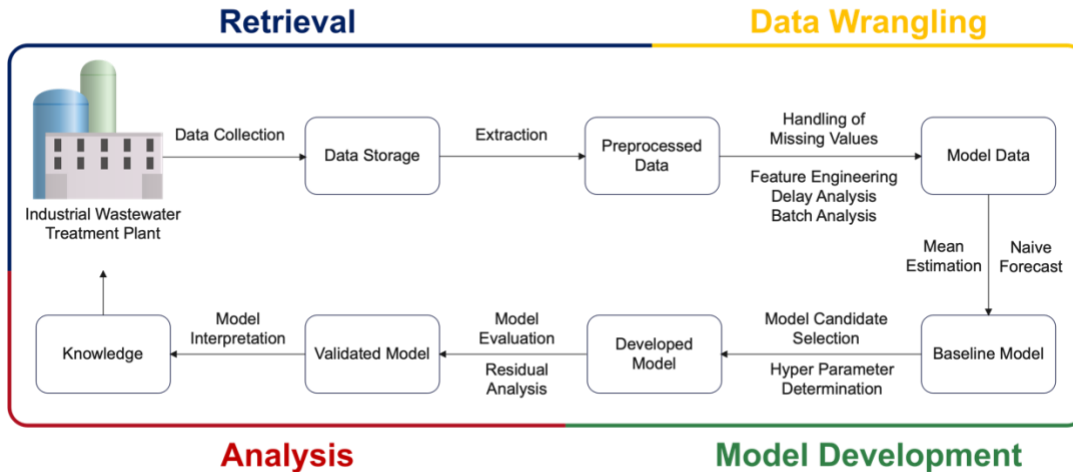


Figure 5 – Overview of the model development from data retrieval to analysis.

3.1. Section I: Data Retrieval

The first section (I) involves extraction (*Step 1*) and initial preprocessing (*Step 2*) to produce a standardized data format used in the remaining analysis. An important point to handle is an appropriate data frequency that can sufficiently represent process dynamics. In process plants, different sensors may provide signals at different frequencies. Therefore, transformations are required to produce a data set where all the signals originating from the plant sensors have the same observation frequency.

3.2. Section II: Data Wrangling

The second section (II) includes different data wrangling operations, and it is comprised of eight steps (*Steps 3- 11*). *Step 3* involves imputation of missing data. Missing logged data can be due to a variety of factors such as sensor failure or communication issues. In addition, the use of batch operations may lead to sensors not recording data during downtime resulting in no signal and missing values after the data aggregation in *Step 2*. The next step (*Step 4*) is to study process delays. Indeed, due to the semi-continuous operation of the tanks, a large delay is expected when correlating the upstream with the downstream data. The delay is equivalent to a hydraulic delay, which corresponds to the residence time in storage tanks. This step helps to identify the right determination of the number of past observations (lags) which is important for time series forecasting. *Step 5* involves batch detection and characterization to distinguish various switches from one operational model to another. At *Step 6* the start time of each batch (after an operation

switch), year / month / week / day /weekday / hour / minute are encoded. These additional time-related features represent the cyclical and seasonal nature of the input data. Next (*Step 7*) upstream information from the production site (if multiple options are possible) and the type of product being produced is to be added. Once missing data is treated, delays are studied and batches identified/characterized, the next step (*Step 8*) is adding lag variables to account for the impact of past events. In *Step 9*, an aggregate measure of the process performance is created for batch processes. This can be obtained by a simple summary statistic such as the mean or median. For normal distributions the mean is sufficient, however for machines with downtime the distribution of values may not be normal, but instead a zero-inflated normal distribution is considered with the peak at zero corresponding to machine downtime. This peak can be the observation that is observed the most rendering the median impractical for predicting performance while the machines are running. In such case, the summary statistic can instead be changed to a zero-excluded median thereby avoiding the peak corresponding to inactive machines. The last steps in this section (*Step 10 & 11*) are batch summarization and train-test splitting. In the first, a vector for each variable within a batch period is input to the transformation function, where the batch period must be specified. In the second, the data is split into a training set and a test set. The training set can be used for model development and validation whereas the test set may only be used for the evaluation of future performance on unseen data.

3.3 Section III: Model Development

The third section (III) includes four major steps. The first step (*Step 12*) involves choosing fit for purpose modelling approaches amongst multiple choices such as support vector machine and random forest. The second step (*Step 13*) is more related to determining how much the selected model candidates can learn and improve. In *Step 14* there is an extra-processing step transforming categorical variables into one-hot encodings and entity embeddings. Finally, in *Step 15* model hyperparameters are tuned to produce predictions that reproduce the data satisfactorily without overfitting.

3.4. Section IV: Model Analysis

The final section (IV) assesses how a developed model performs within a test environment, and how the model is interpreted to acquire new knowledge regarding the polymer dosing task. *Step 16* starts quantifying the prediction capabilities of the different models using appropriate metrics. The last step (*Step 17*) is to study the contribution of each variable in the overall predictive output.

Based on the information generated during sections I-IV, process engineers can exploit the power of data science procedures as presented in this study to solve practical problems such as selecting the best polymer dosage rate by taking advantage of both upstream and downstream data. To capture hidden patterns, extra information such as delays (lag variables), time-related features and entity embedding of categorical variables were analyzed and incorporated in the presented framework. The framework is built for the purpose of online application with the ability of

integration in an automation setup for online control of the dewatering process. It is important to highlight that even though the methodology seems very specific the very same procedure could be applied to solve new operational challenges within the plant i.e. dosage of lime in the inactivation tanks [24].

4. Results

4.1. Section I: Data Retrieval

4.1.1. Step 1: Data extraction

The plant is equipped with 309 online sensors logging data (see section 2.2.3). The raw signals are stored together with the 3 production schedules in a database (see section 3.2 and 3.3). The database was accessed by using an inhouse Excel application-programming interface (API). In this study, data from selected sensors were retrieved from the 1st of July to 31st of December 2020.

4.1.2. Step 2: Data aggregation

It is important to highlight that the data in the original format would lead to gaps in the data due to differing measurement frequencies. To circumvent this limitation, the data were arbitrarily aggregated in 15-minute intervals by taking the mean or median for continuous and categorical variables respectively (see Figure 3). No further consideration was given to the aggregation window. From a practical point of view, smaller intervals lead to more frequent measurements that can reveal faster process dynamics at the expense of increased computational time, but the interval length was not deemed a limiting factor worth of investigation in this study.

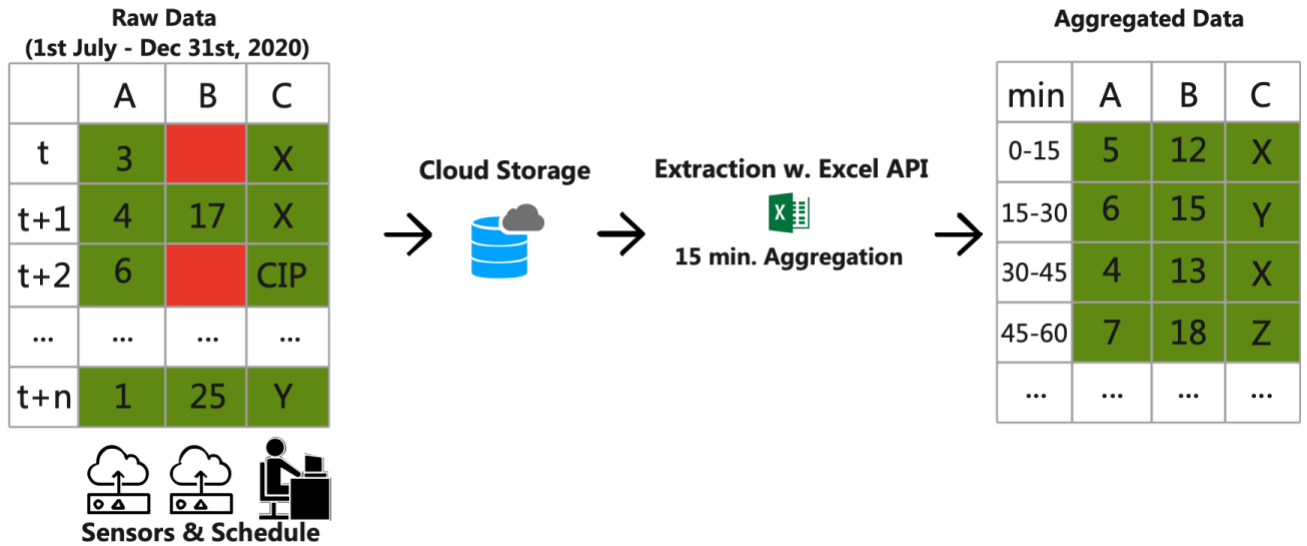


Figure 3 - Illustration of the data aggregation process. Each raw signal is stored with individual measurement frequencies and the data is then aggregated in 15-minute intervals by calculating the mean and median for continuous and categorical signals

respectively. The aggregation produces a data format where each signal has the same number of aggregate measurements simplifying the modelling phase.

4.2 Section II: Data wrangling

4.2.1. Step 3: Imputation of missing values

In this case study, it is assumed that the missing values are due to either batch downtime or equipment maintenance. The missing values are replaced by zeros automatically to enable the next steps in the methodology, thereby disregarding the potential of adding a systematic bias. The choice on replacing rather than removing the missing observations was based on preserving the size of the dataset. Figure 4a shows representations of missing data in form of a histogram. This figure indicates the fraction of missing data across online sensor data. For example, the third bar from the left indicates that 15% of the total 309 sensor data, have 20% to 30% missing values. Figure 4b is a representation of missing data across the process flowsheet i.e. from production sites to inactivation process and then to the decanter hall. On average 41% of the data is missing, however most of these variables are in the production sites and the decanter hall, which contain processes that are either batch based or at the risk of shutdown due to maintenance work. The sensors are offline during this period of downtime, and as such the data is not missing at random and does not represent a lapse in data quality. The units in the inactivation part of the plant run in a continuous fashion with far less downtime.

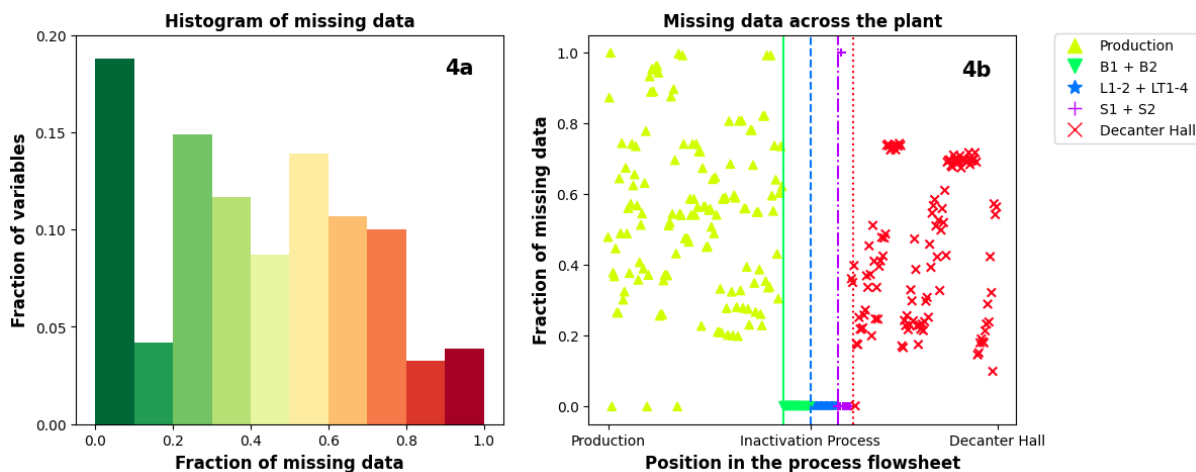


Figure 4a - Histogram of the fraction of missing data for each variable. Figure 4b - The fraction of missing data for each variable by position in the process flowsheet (Figure 1)

4.2.2. Step 4: Study of process delays

As mentioned previously analyzing delays in the process is important to establish a representative input-output relationship. In ideal mixing conditions, the delays for both soluble and particulate compounds correspond to the residence time in a tank. However, this is most likely not the case in real industrial scale with very dynamic operational conditions. Hence, process delay was investigated in more detail. A partial autocorrelation matrix (PACM) of the entire dataset after imputation is used to study process delays between unit operations [27]. The PACM of the data is

calculated for lag 0-48, in which lag 0 corresponds to no lagged states and lag 48 corresponds to states observed 12 hours earlier. Due to the multivariate nature of this analysis, absolute eigenvalues were used to investigate correlations. Figure 5 shows the five largest eigenvalues for the PACM for lags 0-48. Lags 1-9 appear to have eigenvalues that are relatively large compared to those beyond lag 9. This indicates that for lags 0-9 there appears to be a substantial correlation between the states at the current point in time and the lagged states. After lag 12, the eigenvalues begin to increase until they reach a peak around lag 20 before they taper off towards lag 48. The large spike around lag 20 (5 hours) corresponds to the residence time in the two semi-continuously operated storage tanks S1 and S2, which have a residence time of around 4-6 hours depending on the batch.

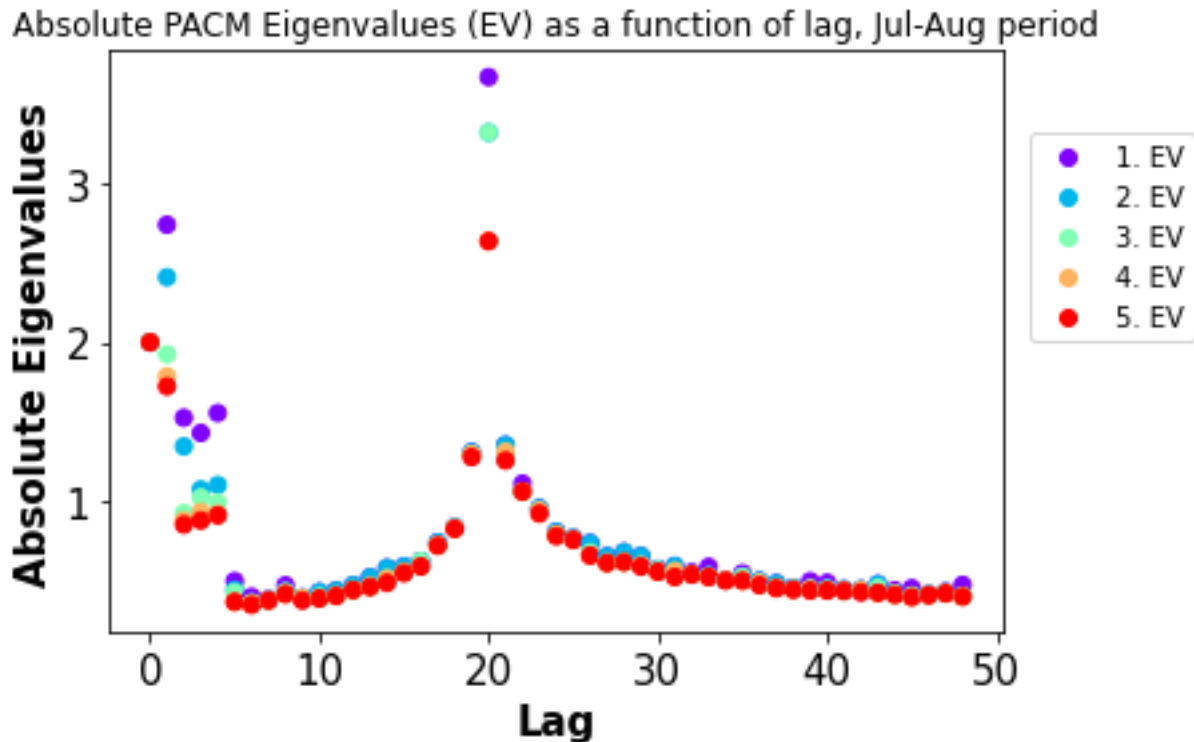


Figure 5 – Five largest absolute eigenvalues for the partial auto-correlation matrix at lags 0-48 for the period 1st of July to 30th of August. The increase in absolute eigenvalues at lag 20, corresponds to the residence time in storage tanks S1 and S2 (see Figure 1).

4.2.3. Step 5: Batch Switch Detection

To make the framework suitable for automatic online applications, a custom algorithm is developed to automatically identify when there is a change from feeding the decanter hall from S1 or S2. Six different rule-based algorithms with the following characteristics were tested using plant data:

- 1st (I1): Check the value at the previous and next time-step for one tank to assign a minimum or maximum,

- 2nd (I2): For each time-step, if one tank has a local minimum and the other has a local maximum, identify a batch switch
- 3rd (I3): For each time-step, if one tank has a local maximum, identify a batch switch
- 4th (I4): For each time-step, if one tank has a local maximum above 500m³ (42% volume), identify a batch switch
- 5th (I5): For each time-step, if one tank has a window size 4 moving average local maximum above 500m³ (42%), identify a batch switch
- 6th (I6): For each time-step, if one tank has a local maximum above 500m³ (42%) and the following holds $V(t) - V(t+2) > 5\text{m}^3$, identify a batch switch

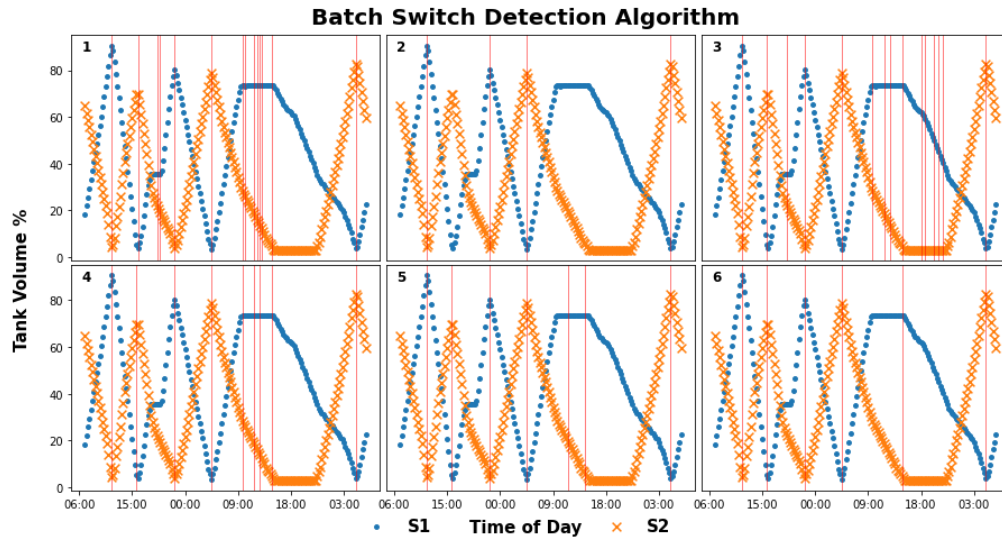


Figure 6 - The six rule-based batch switch detection algorithms applied to a challenging period in the dataset. Vertical red lines are indicative of identified batch switch times. Numbers in the figures refer to batch switch detection algorithm candidates 1 to 6.

Figure 6 depicts the semi-continuous operation of tanks S1 and S2 together with the prediction capabilities of the different algorithms. In the absence of valve position values, tank volume, as the main indicator of batch switch, was investigated during a challenging 24 hour operation period. I1, I3 and I4 tend to assign too many batch switches, whereas iteration I2 assigns too few. I5 and I6 are evenly matched, however in figure 6 both iteration 5 and 6 miss assigning a switch around 09:00, but I5 also assigns a switch that is not present in the period right after.

The 6th iteration (I6) section was applied to the entire period of data and validated manually by visual inspection. The algorithm identified 877 batch switches. However only 869 were found when validated manually. One of the errors was an unlabeled batch switch whereas the remaining 9 errors were mislabeled batch switches. Hence, the accuracy of the applied algorithm is 99%. Identified start and end time points for each batch were used to create the batch summaries

4.2.4. Step 6 & 7: Encoding time observation and microorganisms classification.

The production schedules include alpha-numerical codes that characterize what product is being produced, and the batch number. The product code is classified based on its production organism

to allow the study of patterns associated with specific stream characteristics. An illustration is given in figure 7.

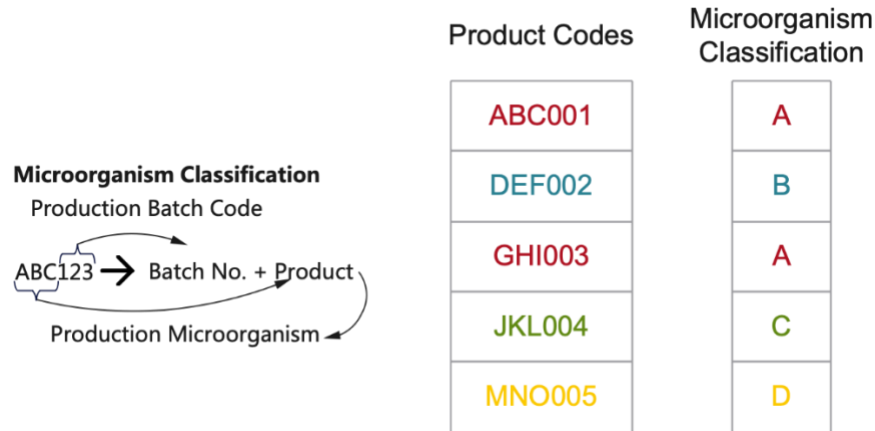


Figure 7 - The product codes from the three different production sites are classified by a senior scientist based on which microorganism is used for the given production.

4.2.5. Step 8: Addition of the lagged variables

Based on the analysis described in section 4.2.2 (Step 4) it was decided to add lags until lag 9 to capture relevant relationships and interactions between variables located at different sections of the plant. The lagged variables correspond to the originally observed variable at the prior timestep through to the variable observed 9 timesteps earlier.

4.2.6. Step 9: Analysis of the current mode of operation

In the current mode of operation, the different centrifugal decanters in the hall are typically dosed with the same polymer dosage at the same point in time, except for when a machine is shut down or when testing is performed. Hence, predicting the mean polymer dosage would not be indicative of the optimal dosage decided by the operator, since the machines operating abnormally would systematically bias the mean polymer dosage rate to a lower polymer dosage rate than used in the active machines. Therefore a zero-excluded median final polymer dosage for each batch is used as the target variable for the modeling part of the study. Apart from the zero-excluded median polymer dosage a set of statistics are calculated:

- Start dosage, l/m³: The starting zero-excluded median polymer dosage across all machines
- Final dosage, l/m³: The final zero-excluded median polymer dosage across all machines after operator optimization
- Changed optimum: Whether or not the start and final dosages are equivalent
- Condition change: Whether or not the zero-excluded median polymer dosage was changed during the batch
- Time to optimization: Number of time-steps before the final dosage is observed the first time
- Time with optimized conditions: Fraction of time that the final dosage was utilized

- Number of machines on spec: The number of machines utilizing the zero-excluded median dosage
- Number of idle machines: The number of idle machines during a batch

The purpose of these statistics is to investigate if there are any inter-day trends due to the work shift, or any weekly patterns.

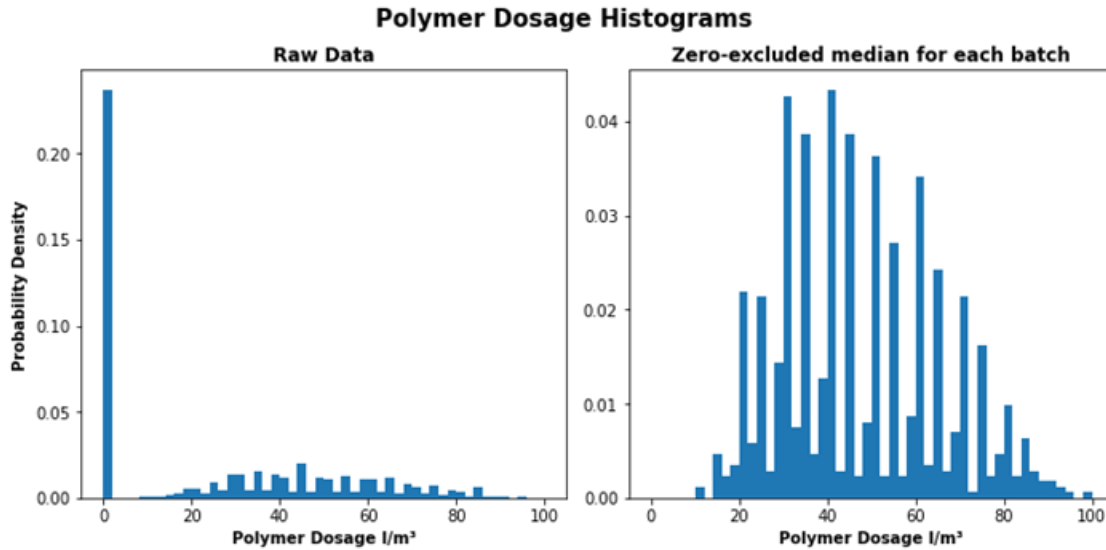


Figure 8 - Polymer dosage histograms. Left: Histogram based on the raw data for all 12 decanters at each time step. Right: Histogram based on the final zero-excluded median of all decanters for each batch.

Figure 8 shows two histograms of the polymer dosage. The left histogram is based on the raw data before any processing is done, and here it is seen that almost every fourth observation is null, corresponding to a machine out of order. The right histogram shows the zero-excluded median polymer dosage across all decanters at the final time-step, i.e. the one optimized by the operator, for each batch as identified by the batch switch detection step.

Table 2 shows the average of the polymer dosage analysis statistics across all the batches from June to December 2020. The values are averages according to which operator work-shift or which weekday it is. The polymer dosage analysis is meant to be a simple indicator of how often polymer dosages are changed to look for patterns across shifts. The most notable difference in **Table 2** is that the morning shift consumes approx. 10% more polymer than the evening and night shifts.

Table 2 - Polymer dosage analysis by shift, and by weekday.

	Morning	Evening	Night	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Start dosage, l/m³	49.6	45.7	46.2	48.6	49.0	45.1	47.4	47.6	45.6	47.2
Changed optimum %	83.1	87.0	84.2	84.1	88.7	86.3	87.2	83.1	82.9	81.3
Condition change %	94.2	98.0	93.9	96.8	96.8	92.4	95.2	97.6	95.5	93.8
Time to optimization	1.7	2.7	2.2	1.9	2.3	3.0	1.8	2.0	2.1	2.3
Time with optimized conditions	0.4	0.3	0.4	0.4	0.3	0.3	0.4	0.4	0.3	0.4
Number of machines on spec	3.9	3.6	3.6	3.8	3.5	3.8	3.7	3.7	3.6	3.8
Number of idle machines	5.1	5.3	5.5	5.1	5.5	5.4	5.3	5.2	5.5	5.1

4.2.7. Step 10: Batch summarization

The two storage tanks switch between emptying and filling approximately 3-5 times a day. In order to provide a meaningful target it was decided to predict the final zero-excluded median polymer dosage for each batch based on data summaries of past data, i.e. upstream variables collected before the storage tanks, and the historical condition of the decanters. The batch switches detected by the algorithm are used to define a batch period wherein the remaining data is summarized with a mean or median for continuous and categorical variables respectively.

4.2.8. Step 11: Train-Test Splitting

The upstream information from the polymer dosage is summarized for a given batch by calculating the mean and median for continuous and categorical variables respectively. In total 869 batch summaries are calculated. For these 869 batch summaries, the upstream information associated with the first 868 batches is used to predict the downstream polymer dosage of the last 868 batches (= one batch ahead prediction). In this way, the upstream information collected during the filling of S1 and S2 is considered the set of independent variables, and the polymer dosage for emptying a respective batch is the dependent variable. A train-test split of 4:1 is chosen arbitrarily so that for these 868 batches the first 80% are used for training and the last 20% are used for evaluation of the final model performance.

4.3. Section III: Model development

4.3.1. Step 12: Select model implementation

Two model candidates are selected: the linear model partial least squares (PLS) and the non-linear model random forests (RF). PLS is commonly used within the field of chemometrics while RF is commonly used within the field of data science for regression and classification of tabular data, such as the data investigated here [28]. Furthermore, for tabular data RF typically reaches similar accuracies compared to neural network-based methods with less training time [29]. For mathematical details on PLS and RF the reader is referred to Wold et al. [30] and Breiman [31] respectively. The model implementations used for PLS and RF are from the Sci-kit Learn library [32].

4.3.2. Step 13: Defining model baselines

In this study, two baselines are used: mean estimation and naïve forecast. The mean estimate is defined as using the mean of the training population of the dependent variable to estimate the dependent variable for the test set. This results in a root-mean square error (RMSE) of 18.41 l/m³. The naïve forecast is defined as using the prior value of the dependent variable as the estimate for the dependent variable at the subsequent time-step. In this case, the calculated RMSE is 17.09 l/m³

4.3.3. Step 14: Categorical entity embedding

To convert categorical data such as the production schedules to a numerical form with entity embedding, two forecasting datasets are created and compared as illustrated in **Figure 9**. One set uses one-hot encoding to process the categorical variables whereas the other set utilizes categorical entity embeddings instead. An embedding is a vector representation of a categorical variable. In one-hot encoding a category is replaced with a binary (dummy) variable. Entity embedding is a similar transformation but keeping informative relations between a feature's values, revealing the inherent continuity of the data [33]. The categorical entity embedding is carried out utilizing the fastai python package [34].

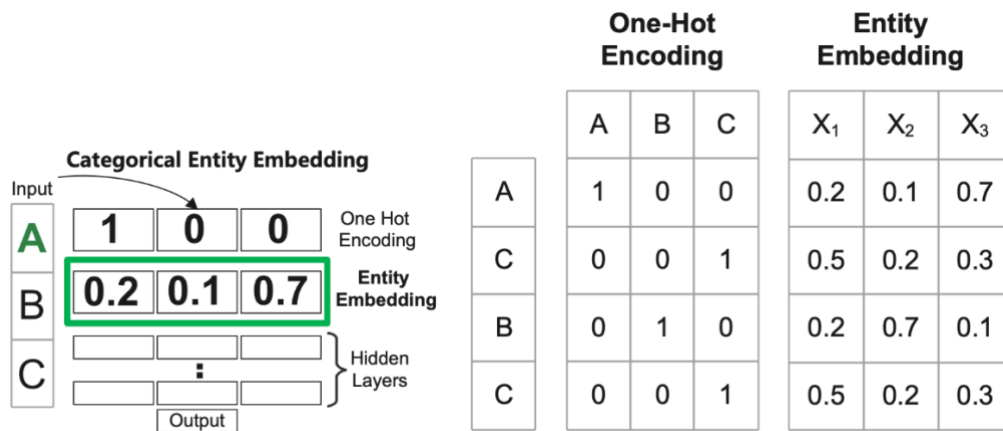


Figure 9 – A categorical entity embedding is obtained by training a neural network on the batch summarized data to predict the target polymer dosage. The entity embedding is used to encode the categorical variables in a continuous fashion rather than the

discrete one-hot encoding. In this example A , B and C represent 3 different levels of a categorical variable, and X_1 , X_2 , X_3 represent a continuous linear combination of A , B , C after entity embedding.

Table 3 shows the addition of variables in the one-hot encoded and entity embedded version of the dataset. The initial feature engineering adds 19 extra variables: 8 variables related to time, 4 related to microorganism classification, 2 related to the pH of the tank during filling, 1 related to the volume during filling, 1 related to the time it takes to fill a batch, 1 total lime flow into LT1-4, a variable expressing the ratio between biomass from the production sites and the waterline and finally the target polymer dosage discussed previously. The one-hot encoding adds 106 variables based on the 3 product codes, as well as the 4 variables from the microorganism classification. The addition of 9 lagged states multiplies the number of variables by 10, however lags are not added for the filling time or the polymer dosage resulting in the addition of 3888 and 2934 variables for the one-hot encoded and entity-embedded dataset respectively. The last step in producing the entity-embedded dataset is the entity-embedding which adds 427 variables based on the 3 product codes, as well as the 4 variables from the microorganism classification. The final total variable count is 4322 and 3689 for the one-hot encoded and entity-embedded datasets respectively.

Table 3 – Overview of the number of variables after each section.

	Start	Feature Engineering	One Hot Encoding	Addition of lags	Entity Embedding	Total
One-hot encoded	309	+19	+106	+3888	+0	4322
Entity embedded	309	+19	+0	+2934	+427	3689

4.3.4. Step 15: Hyperparameter selection

4.3.4.1 Random Forest Hyperparameter determination

For the data set with entity embedded variables, and that with one-hot encoded variables, the distributions that the maximum features are drawn from are uniform with lower and upper bounds of $U(1, 3688)$ and $U(1, 4321)$ respectively, whereas the distributions are unchanged for the other hyperparameters. The RF hyperparameters are tuned to prevent overfitting and they regularize the size of the models to ensure that they are not over specified and can generalize well. Instead of using a conventional validation split, the hyperparameter combinations are evaluated by time-series cross validation with 3 splits, which was chosen as to avoid over-fitting on the training set, and to avoid the increase in computational time from increasing the number of folds. 200 hyperparameter combinations were randomly drawn from the distributions and evaluated.

The best cross-validation coefficient of determination (R^2) for the one-hot encoded data was 0.09, whereas the best cross-validation R^2 for the entity embedded data was 0.12. The mean cross-validation score for the one-hot encoded data was 0.03 whereas for the entity embedded data the

mean cross-validation score was 0.06. Therefore, the entity embedded data is used over one-hot encoded for the subsequent analysis.

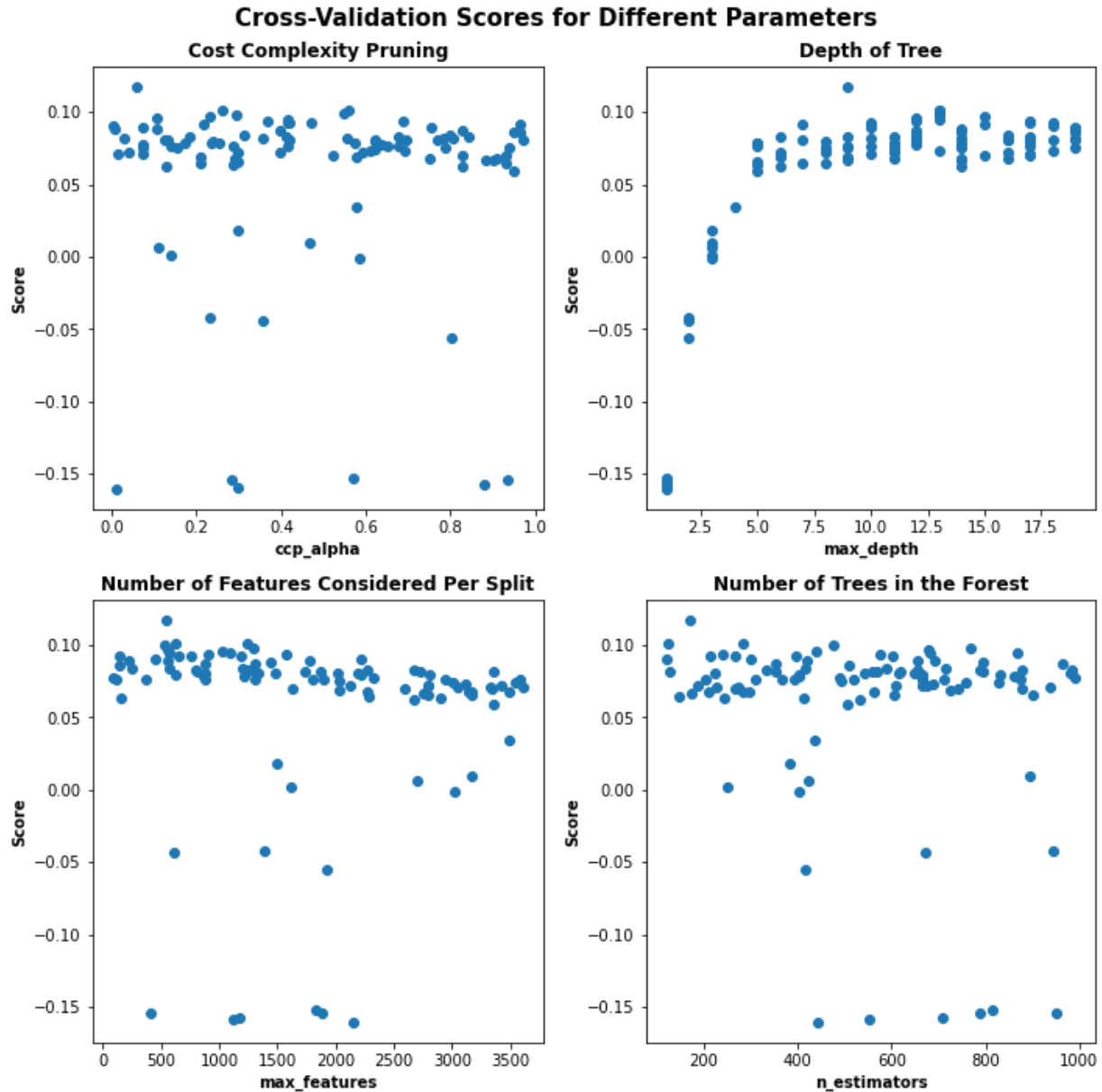


Figure 10 - Results of time-series cross-validation. Each plot shows the cross-validated score as a function of changing a given hyperparameter.

Figure 10 shows the results from the time-series cross-validation based on the four hyperparameters. It appears that the depth of the tree, maximum depth, has a clear positive correlation with the performance until a drop-off occurs around depth 5. Calculating the standardized regression coefficient between the hyperparameters and their cross-validation scores in table 3 reveals a similar trend that the depth of the tree has the highest correlation with model performance, the max features hyperparameter is insignificant, and CCP and the number of trees

have a lesser impact on the model performance. The best random forest model was built with the parameters shown in Table 4.

Table 4 - Standardized regression coefficients of the cross-validation results.

	CCP	Max depth	Max features	Number of trees
Standardized regression coefficient	-0.008	0.041	0.000	-0.005

4.3.4.2 Partial Least Squares hyperparameter determination

In a similar fashion the PLS model was cross-validated to determine the number of components to retain, but the PLS models best scores were all negative indicating that they were arbitrarily worse than the mean. The PLS models were overall better while built on the entity embedded data set with a mean cross-validation score of -1835.36, where the ones built on the one-hot encoded one had a mean cross-validation score of -5277.58. The PLS models are dropped from further analysis since no combination of data and number of components showed a positive cross-validation score.

4.4. Section IV: Model analysis

4.4.1. Step 16: Model evaluation

The RMSE of the RF model on the unseen test set is 17.75 l/m^3 , which compared to the baselines is 3.6% lower than mean estimation, but 3.9% higher than a naïve forecast. One advantage that the naïve forecast has over the RF is that it uses information of the prior batch as an independent variable, which the built RF does not. In an attempt to alleviate this drawback, a new data set is constructed where the entity embedded filling and emptying information of the prior batch is added to the set of independent variables resulting in a dataset with 7378 independent variables, and the hyperparameter optimization procedure is repeated to train a new random forest model. The new RF has an RMSE of 16.19 l/m^3 , 12.1% and 5.3% lower than the mean estimation and naïve forecast respectively and was trained with hyperparameter values of 0.69, 18, 7212 and 688 for the CCP, tree depth, max features and number of trees respectively. Figure 12 shows the predicted vs actual polymer dosages for the test set, the final 20% of the data. The model appears to have a problem with predicting low polymer dosages compared to high polymer dosages since its estimations are all approximately within the $40\text{-}70 \text{ l/m}^3$ range whereas the observations approximately lie within the $10\text{-}85 \text{ l/m}^3$ range. Most of the predictions lie within the $40\text{-}60 \text{ l/m}^3$ range which could be due to an over-representation of those values in the training data. In an online scenario, it would be possible to update the model before the next prediction has to be made. To give a better idea of how the model would perform if we update it after every new batch, the model is repeatedly retrained before predicting the polymer dosage of the next batch. The RF feature importance is used to evaluate prediction capability of the independent variables for predicting the dependent variable. The feature importance of all the variables sum to one for a given model and describe how much a given feature reduces the error of the model, and a decision is made to drop any variable with a feature importance lower than 0.0001. This reduces the total number of independent variables from 7378 to 1375. After this the RMSE on the test set is reduced to 14.66 l/m^3 , 9.5 % lower than the model prior to retraining and dropping variables. From the right hand side plot in Figure 11 it is evident that the model can predict within a wider range after dropping redundant features and retraining after each prediction as seen by the increased spread in the blue predictions.

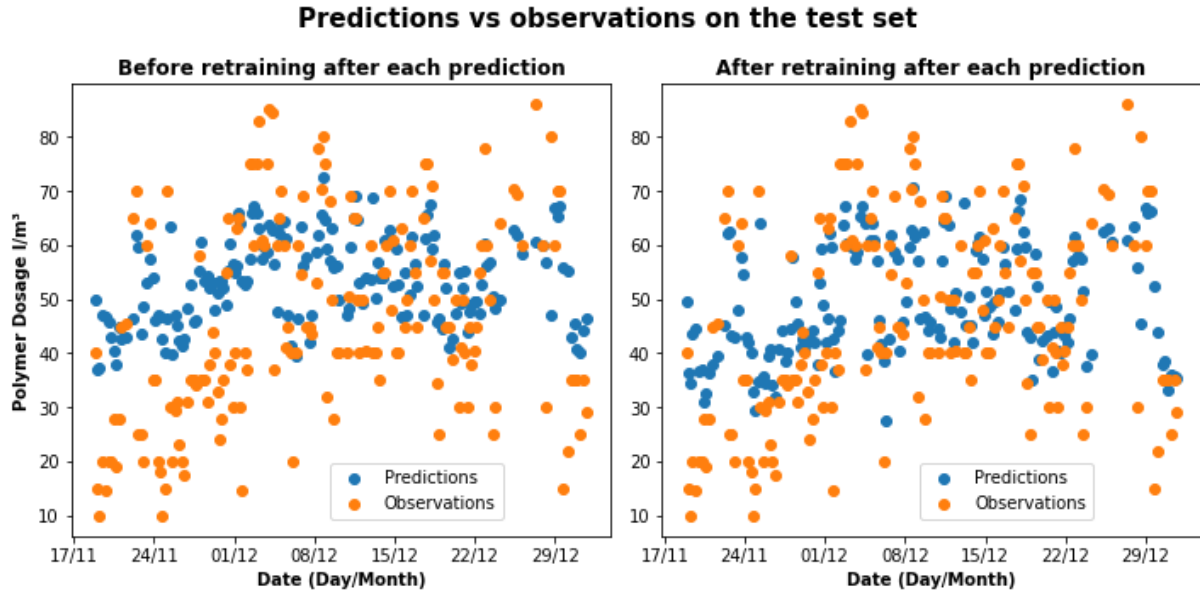


Figure 11 – Predictions (blue) vs observations (orange) for the test set before and after dropping redundant variables and retraining the model after each prediction.

4.4.2. Step 17: Model interpretation

4.4.2.1 Interpretation of Random Forest feature importances

In Table 5 and Table 6 the highest feature importance is shown in terms of weight as well as which batch wise lagged version of the variable it is. The most important feature according to the model is the past polymer dosage, i.e., the autoregressive component. The flow from one of the production sites, here denoted production A, is the second most important albeit with a much lower weight. The past cake production has the third highest feature importance and can also be considered as an autoregressive component, that signals when cake production is of a certain level then the following cake production should be similar. The batch ID for the same production site also shows up as the fourth most important variable. For most of the variables except the past polymer dosages, the variables are present for the past batch, and the batch before the past batch, and then for different lagged versions.

The autoregressive variables describe the last two polymer dosages with the prior polymer dosage making up almost the entirety of that sections feature importance. Both the time and cake production information do not appear to have an easily interpretable relationship between their value and the contribution of that variable to the polymer dosage estimate. Production A appears to have higher feature importance in the model, compared to productions B and C. The buffer, storage and lime sections are all less important according to the model than either of the production sites suggesting that the principal source of variation comes from the upstream production sites and not the current treatment procedure applied downstream. The buffer, storage and lime sections are tightly controlled and as a consequence contain less variance than the upstream production which leads to lower predictive power.

Table 5 – RF feature importance of the full model with filling and emptying information from the prior batch.

Feature	Polymer Dosage	Production A Flow	Cake Production	Production A Batch ID	Day of year
Batchwise Lag	1	2	1	2	2
Weight	0.3884	0.0073	0.0063	0.0058	0.0051

Table 6 – RF feature importance of the full model summarized for different sections of the plant

	Autoregres sive	Prod. A	Decanters	Time	Prod. C	Prod. B	Buffer Tanks	Storage Tanks	Lime Tanks	Cake Production
Weight	0.39	0.1890	0.1693	0.0677	0.0434	0.0387	0.0291	0.0262	0.2336	0.0146

5. Analysis of the production sites

In addition to analysis of the results reported in Table 5, it is also of interest to interpret the three main production sites considered in the model independently. These three production sites denoted A, B and C have their own unique characteristics in terms of layout and products being generated (see Figure 1)

In order to estimate the effect of the different products, the contributions for all embedded and lagged versions of the batch ID variables are summarized for each prediction when the product is of a certain value. All the product codes have been anonymized and denoted 1-20. For each production site, the procedure is repeated but where the products are classified as described in section 4.2.5 to look for differences across microorganisms. In Figures 13-15 boxplots of the contribution from the different product codes are shown for the different production sections. For each prediction in the test set the contributions for each variable related to the associated batch ID are summarized, and the boxplots are then generated from this result across the test set. This metric does not account for interactions between the product codes and other variables such as the flows from the corresponding section.

For production A the procedure is repeated after categorizing the product codes by one of three process treatments, j, h, k, and repeated again for which microorganism, a or Z, was used in the production. Figure 12 shows that for some products such as product 1-5 the model generally decreases its estimate of the required polymer dosage, and for some products such as 6 and 9 the model generally increases its estimate of the required polymer dosage. For treatment h the model suggests lower polymer dosages than for j and k. On the other hand, j and k seem to be indistinguishable from each other. For microorganisms Z the model generally suggests lowering the polymer dosage in comparison with microorganism a where there is a much larger spread than for microorganism Z.

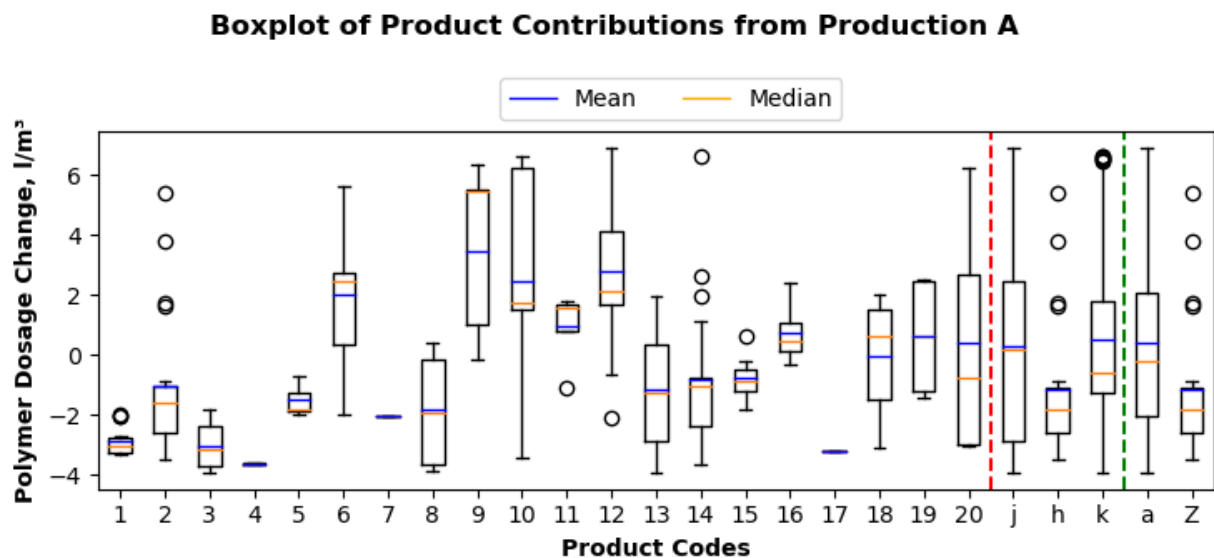


Figure 12 – Boxplots of the influence of product codes from production A on the predicted polymer dosage. The boxplots between the red and green lines are the product codes categorized by one of three processing methods, whereas the boxplots after the green line are the product codes categorized by which microorganism is used in production.

For site B the effects of the product on the estimated polymer dosage are lower (compared to A) as seen evident by the scale of the y-axis in Figure 13. More specifically, for product 1, 5, and 16 the model suggests lowering the polymer dosage whereas for the other products the picture is less clear. No immediately noticeable trends are evident in the boxplots that were categorized by production microorganism

For production C illustrated in Figure 14 the effects of the product on the estimated polymer dosage is higher than for production B but lower than for production A. For product 15 and 18 the model suggests lowering the polymer dosage, whereas for product 3 and 12 the model suggests increasing the polymer dosage. For the microorganism classified boxplots the model suggests increasing and decreasing the polymer dosage for products f and b respectively.

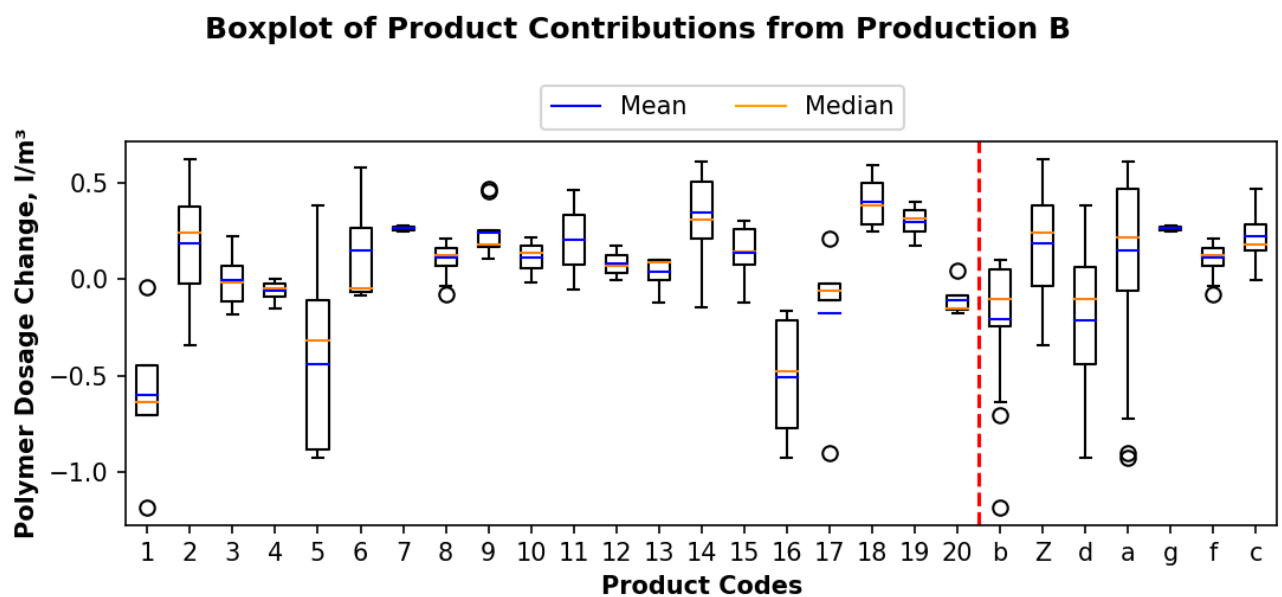


Figure 13 – Boxplots of the influence of product codes from production B on the predicted polymer dosage. The boxplots after the red line are the product codes categorized by which microorganism is used in the production.

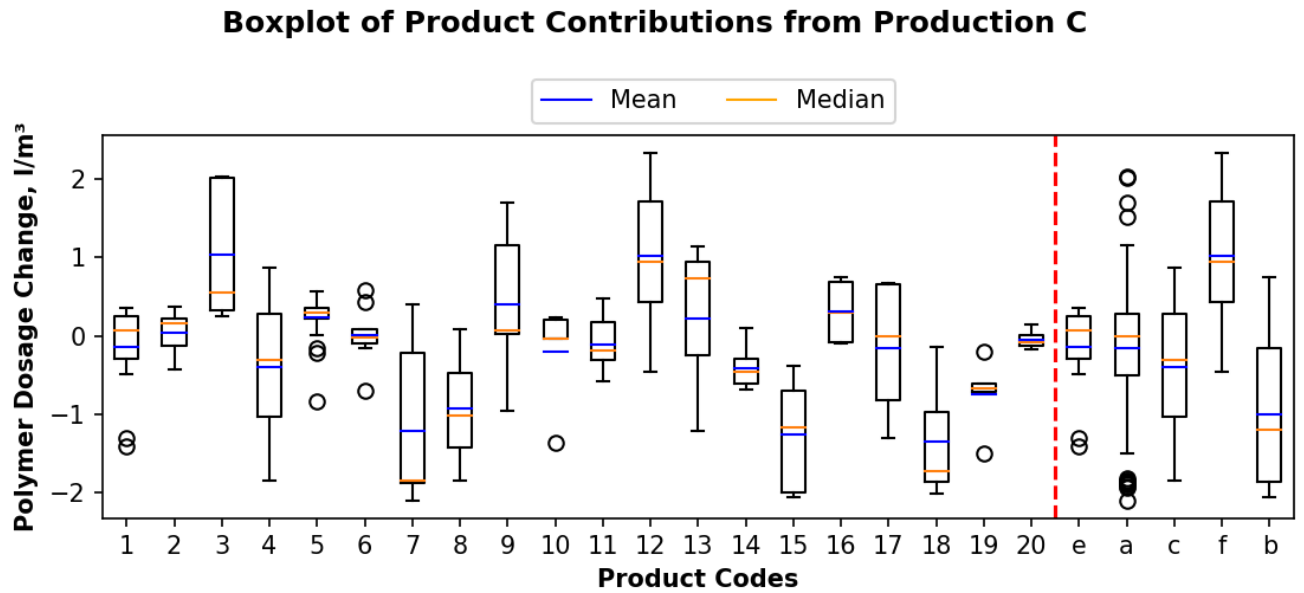


Figure 14 - Boxplots of the influence of product codes from production C on the predicted polymer dosage. The boxplots after the red line are the product codes categorized by which microorganism is used in the production.

6. Discussion

6.1. Data aggregating and wrangling

The raw data was aggregated into averages over 15 minutes where the 15-minute window was chosen arbitrarily. Decreasing the length of the aggregation window could capture small infrequent changes, however considering that, the 15-minute aggregations are later summarized by batch, this should not result in major changes to the rest of the methodology. Decreasing the length of the aggregation window would affect the dPCA by increasing the number of lags required to account for the same period which would result in an increase of the required computational time. Increasing the length of the window would make the dPCA faster and more lags could be accounted for at the cost of capturing less of the process dynamics. For most of the variables, one of the lagged versions had higher feature importance than the original variable indicating that the addition of lagged variables could be beneficial despite the increase in number of variables. While dPCA has traditionally been used for statistical monitoring and fault detection at large industrial plants and chemical processes this shows the application of dPCA as a feature engineering tool for predictive tasks within a similar system [27, 35]. The methodology serves as a nice supplement to residence time analysis in cases where process dynamics are present, and unlike residence time analysis there is no need to know the specifications of the tanks and pipes at the plant making it a more generalizable tool to study process delays at large plants.

The developed batch switch detection algorithm had an accuracy of 99% and it provides an easy and automated way to label a switch from one batch to another. If the valve positions were available, they could be used instead of the tank volume however, there would still be a need to account for stops in emptying and filling where there is no switch between batches. The developed batch switch detection algorithm utilizes the 2 future values to detect a batch switch

i.e. the change cannot be identified before there has already been a switch between the two batches. In practice however from the past batch switch one can start calculating a mean summary used as the input for the model as the model does not require the batch to be finished before a prediction can be given, and the process dynamics at the plant are considered slow enough that a few extra observations in the batch summarization would not change the batch description significantly. The batch switch detection algorithm can be used for other systems operating in semi-continuous mode to label the occurrence of a new batch and enable further data analysis. The added time variables do not appear to encapsulate seasonal dynamics and should be interpreted as being a way for the model to pick up on similar values located in time as a type of autoregressive component. The prior polymer dosage is the variable with the highest feature importance and combining the time variables also accounts for approximately 6.8% of the error reduction when building the model. The autoregressive variables in the model suggest that the polymer dosage should be akin to the prior polymer dosage and as such the autoregressive variables do not aid interpretability but can increase the forecast accuracy in systems with a sequential nature such as this one where the polymer dosages exhibit some correlation with past polymer dosages. The validity of the model interpretation relies on the accuracy of the model, and while including the autoregressive components can boost the accuracy, they do not offer direct interpretation. However, AR components help establish a more credible model with a higher accuracy which can still be used to interpret the effect of the non-autoregressive components, such as the flows and product codes. The autoregressive components can easily be dropped which will raise the importance measure of the other variables as the variable importances are interdependent however the accuracy will drop partially reducing the model credibility.

The microorganism encoding enables analysis of the effect that different types of biosolids have on the dewatering operation. From a modelling perspective the microorganism encoded variables had lower feature importance than the entity embedded product codes, so the microorganism encoded variable may be excluded from the model and used instead for simplifying the analysis of which products affect the polymer dosage. The entity embedding of microorganisms can be seen as a use case for utilizing categorical information found in large industrial processes and chemical plants such as for instance product codes, treatment types or equipment types and entity embedding the categorical information should improve the ability of predictive models to utilize this information resulting in higher accuracies and enabling a more cohesive analysis of the categorical variables as they are better utilized in the model after entity embedding [33].

Analysis of the polymer dosage patterns was necessary to generate a prediction task that aligned with solving the task of forecasting the polymer dosage determined by the operators. Without the polymer dosage analysis, the model would be heavily affected by inactive machines. This approach can be extended to other equipment with similar on/off usage patterns to only investigate the desired behavior. The polymer dosage analysis developed here may be used as a performance indicator for how the operators handle each batch. However, it should be mentioned that the operators have other tasks than dosing polymer for the decanters, and the indicators are not well developed or tested for actual implementation as key performance

indicators. Nevertheless, they may form the basis of a monitoring scheme which could be combined with an operator schedule to determine if some operators consistently perform differently from others over a longer period of time.

6.2. Model development and analysis

In this study, naïve forecasts, mean estimation, PLS and RF models were used to forecast the polymer dosage. The developed RF outperformed both established baselines and could predict trends in the polymer dosage on the test set. However, it appeared that the model did not predict any values above or below 70 l/m³ or 30 l/m³ respectively. This may be due to an overrepresentation of observations in the range of 30-50 l/m³ that skews the model towards predicting within said range. There exists a wide variety of data-driven techniques such as neural network-based methodologies, and other methods from the field of statistics that could be applied to the problem at hand. However, the purpose of this paper was not to exhaustively test available data-driven prediction models, but to demonstrate data treatment and pipelining in a biosolids dewatering unit operation.

The RF was optimized by drawing hyperparameters from predefined distributions. By plotting the cross-validation score and calculating the standardized regression coefficients it was revealed that mainly the tree depth influenced the performance of the model. Only 200 hyperparameter combinations were tested for each RF model, and this number may be increased however it is believed that the score should only improve marginally in this case. When evaluating the RF model on the test-set the model was retrained to predict the next polymer dosage, and this greatly improved the performance of the model. The number of time-series cross-validation folds could be increased during the time-series cross-validation in the training phase to better imitate the evaluation scenario at the expense of computational time.

Interpreting multivariate models for non-linear systems may be challenging and in this study the analysis of the model was done by summarizing the attribution of each feature category for a prediction and then averaging across the observations in the test set, albeit limited to the product codes since these were of primary interest. The analysis of product codes suggested that some products may on average result in lower polymer dosages than others opening for an internal discussion about which characteristics that distinguish these products from each other and how that may impact the dewatering step. For future work, the RF-based feature importance analysis could be supplemented with partial dependence plots by similarly changing a feature and the corresponding lagged versions simultaneously to see the effect on the dependent variable.

6.3. Operator schedule and turbidity measurements

One fundamental assumption in this study was that the operator determined polymer dosage corresponds to the dosage that gives the best separation on the decanters. The optimal polymer dosage is prone to errors such as operator mistakes and validity of the test method. As such, the quality of ground truth is questionable at best. This implies that the model was built on noisier data than could be anticipated, and conversely it may also mean that replacing the polymer dosage by installing a turbidity sensor and using its measurement as the target variable

in the future could lead to better predictions, where the polymer dosage could instead be analyzed as an independent variable. Including the operator as an independent categorical variable in the model may also be beneficial since different operators may dose differently. It was estimated that this procedure could save the operators 3-5 hours of time daily, as the optimization task can take anywhere from 30 to 90 minutes when performed and the task is performed 3-6 times a day.

6.4. Cost of Implementation

The cost of implementation of the present work as a decision support tool can be broken down into 1) Data storage and access, 2) Computational costs for model development, 3) Model maintenance, 4) Development of potential user interfaces for operators. The storage of data is covered by a companywide budget and extra traffic that a solution like this would create is considered negligible as data would need to be accessed 5-6 times a day which is in line with regular usage.

The computational costs for model development are considered negligible as it can be developed using open-source software on a standard laptop without further modifications (MacBook Pro 18,1). The model maintenance in terms of retraining and validation of new hyperparameters takes approximately 1-2 hours with the specified hardware and is suggested to be carried out once a month, however this procedure can be automated through continuous integration and development procedures which would require some initial effort. For operators to use the developed model as a decision support tool an interface would most likely be required which can be developed with open-source technologies such as Streamlit. The authors estimate that both the automation with continuous integration and development procedures as well as the interface development would amount to approximately a week's worth of work for the first author or an engineer with an equivalent skill-set.

7. Conclusion

The main findings of this study can be summarized in the following points:

- A novel methodology based on data-science provided with polymer dosages forecast for centrifugal decanters treating biosolids within industrial sites. The proposed approach is comprised of four main sections and eighteen steps and allowed to handle important challenges within production such as 1) the semi-continuous nature of process operation, 2) the existence of multiple time lags between upstream (production, water treatment, inactivation units) and downstream (decanter hall) information and 3) the combination of both numeric and categorical variables.
- Different data extraction and wrangling methods were developed and tested at full scale to handle the following activities: harmonize sensor signals to the same frequency, tackle missing values due to down-time / equipment maintenance, automatically detect batch switches, identify (and incorporate) process delays between upstream and downstream operations and finally analyze (polymer) dosages patterns (daily / weekly) & codify / categorize streams generated during production.

- Model development and analysis sections showed that it was possible to forecast chemical dosages (RMSE = 14.66 l/m³, 14.2% reduction from baseline) based on past polymer and cake production (autoregressive component), flow from production site A and batch ID. The model had problems when predicting low and high polymer dosages, however the authors suspect this is due to the noisy and uncertain nature of the data.
- A more in dept analysis of the product codes, process treatment and type of microorganisms reveals their impact on the polymer dosage. This exercise is done for production sites A, B and C. The study shows what products involve increasing and/or decreasing the quantity of polymer. It is also revealed that production sites A and C have a more significant impact on dosage than B.
- The general methodology can be used as a tool to come up with data-driven support tools. This method can be used as guidance to estimate polymer dosages based on upstream production information and downstream operational data. The proposed approach will substantially reduce the required operator time (3-6 h) needed in the lab to estimate polymer dosages based on wet experiments.

8. Software & Data Availability

Most of the code is available via the wwtmodels Github repository (<https://github.com/wwtmodels>), and for undisclosed sections the reader is encouraged to contact the first author by email: sebttop@kt.dtu.dk. The data is not available for release due to confidentiality requirements from the involved company.

9. Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

10. Acknowledgements

The authors would like to acknowledge the financial and academic support of the Technical University of Denmark and Novozymes for their contributions to the project. Dr Ramin and Dr Flores-Alsina thanks the support of the Danish Innovation Fund under the project GREENTAN (Contract-No: 0177-001009A). Dr Flores-Alsina and Prof. Gernaey are grateful to the Miljøministeriets Miljøteknologisk Udviklings og Demonstrationsprogram (MUDP) project NACAT (Contact-No: 2021-20015). Part of this research was conducted when Mr Topalian was an academic visitor at the Australian Center for Water and Environmental Biotechnology (ACWEB). A preliminary version of this study was presented at the Process Systems Engineering (PSE) conference in Kyoto (19-23 June 2022). Mr. Topalian would like to acknowledge Otto Mønstedts Fond and Idella Foundation for their financial support which enabled his stay at PSE and ACWEB respectively.

11. References

- [1] Progress on wastewater treatment – Global status and acceleration needs for SDG indicator 6.3.1. United Nations Human Settlements Programme (UN-Habitat) and World Health Organization (WHO), Geneva, 2021.
- [2] D. Brdjanovic, G. A. Ekama, M. C. M. van Loosdrecht and M. Henze, Biological Wastewater Treatment - Principles, Modelling and Design, IWA Publishing, 2008, pp. 1-6.
<https://doi.org/10.2166/9781780401867>
- [3] K. E. Holmgren, H. Li, W. Verstrate and P. Cornel, State of the Art Compendium Report on Resource Recovery from Water, IWA Resource Recovery Cluster, 2016.
- [4] O. Nowak, Optimizing the Use of Sludge Treatment Facilities at Municipal WWTPs, Journal of Environmental Science and Health Part A. 41 (2006) 9 1807-1817.
- [5] M. L. Christensen, K. Keiding, P. H. Nielsen, M. K. Jørgensen, Dewatering in biological wastewater treatment: A review, Water Research. (2015) 82 14-24.
- [6] R. J. Wakeman, Separation technologies for sludge dewatering, Journal of Hazardous Materials. 144 (2007) 3 614-619.
- [7] Z. Yan, B. Örmeci, J. Zhang, Effect of Sludge Conditioning Temperature on the Thickening and Dewatering Performance of Polymers, Journal of Residuals Science & Technology. 13 (2016) 3 215-224.
- [8] A. Records and, K. Sutherland, Decanter Centrifuge Handbook, Elsevier: Amsterdam, The Netherlands, 2001.
- [9] B. Peeters, R. Dewil, J. F. Van Impe, L. Vernimmen, W. Meeusen, and I. Y. Smets, Polyelectrolyte flocculation of waste activated sludge in decanter centrifuge applications: Lab evaluation by a centrifugal compaction test, Environ. Eng. Sci., 28 (2011) 11 765–773.
- [10] P. Ginisty, Laboratory tests to optimize sludge coagulation / flocculation process before thickening or dewatering, American Filtration and Separations Society 2005 – 18th Annual Conference, Afs, 2005.
- [11] P. Ginisty, R. Mailler, and V. Rocher, 2021. Sludge conditioning, thickening and dewatering optimization in a screw centrifuge decanter: Which means for which result?, Journal of Environmental Management. 280 111745.
- [12] M. Gleiss, and H. Nirschl, Modeling Separation Processes in Decanter Centrifuges by Considering the Sediment Build-Up, Chemical Engineering & Technology. 38 (2015) 10 1873-1882.
- [13] C. Bai, H. Park, and L. Wang, 2021. Modelling solid-liquid separation and particle size classification in decanter centrifuges, Sep. Purif. Technol. 263 118408.
- [14] W. W. F. Leung, Inferring in-situ floc size, predicting solids recovery, and scaling-up using the Leung number in separating flocculated suspension in decanter centrifuges, Sep. Purif. Technol. 171 (2016) 69–79.
- [15] C. Bai, H. Park, and L. Wang, 2022. A Model–Based Parametric Study of Centrifugal Dewatering of Mineral Slurries, Minerals. 12 1288.
- [16] T. Hastie, R. Tibshirani and J. Friedman, 2017, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer New York, 2001.

- [17] X. Kiang, C. Liuxi, Y. Li, X. Gao, and G. Bai, 2022. Investigation on the Separation Performance and Multiparameter Optimization of Decanter Centrifuges, *Processes*. 10, 7, 1284.
- [18] B. Ráduly, K. V. Gernaey, A. G. Capodaglio, P. S. Mikkelsen and M. Henze, Artificial neural networks for rapid WWTP performance evaluation: Methodology and case study, *Environmental Modelling & Software*. 22 (2007) 1208-1216.
- [19] O. Bello, Y. Hamam, and K. Djouani, Modelling of a coagulation chemical dosing unit for water treatment plants using fuzzy inference system, *IFAC Proceedings Volumes (IFAC-PapersOnline)*. 19 (2014) 3985–3991.
- [20] C. D. Jayaweera and N. Aziz, An efficient neural network model for aiding the coagulation process of water treatment plants, *Environ. Dev. Sustain*. 24 (2022) 1069–1085.
- [21] E. Hong, A. M. Yeneneh, T. K. Sen, H. M. Ang, and A. Kayaalp, ANFIS based Modelling of dewatering performance and polymer dose optimization in a wastewater treatment plant, *J. Environ. Chem. Eng*. 6 (2018) 2 1957–1968.
- [22] J. L. Pitarch, A. Sala and C. de Prada, 2019. A Systematic Grey-Box Modeling Methodology via Data Reconciliation and SOS Constrained Regression, *Processes*. 7 3 170.
- [23] P. Menesklou, T. Sinn, H. Nirschl and M. Gleiss, 2021. Grey Box Modelling of Decanter Centrifuges by Coupling a Numerical Process Model with a Neural Network, *Minerals*. 11 7 755.
- [24] V. Monje, H. Junicke, D. J. Batstone, K. Kjellberg, K. Gernaey and X. Flores-Alsina, 2022. Prediction of mass and volumetric flows in a full-scale industrial waste treatment plant, *Chemical Engineering Journal*. 445 136774.
- [25] V. Monje, M. Owsianiak, H. M. Junicke, K. Kjellberg, K. V. Gernaey & X. Flores-Alsina, 2022. Economic, technical, and environmental evaluation of retrofitting scenarios in a full-scale industrial wastewater treatment system, *Water Research*. 223 14 118997.
- [26] European Parliament and the Council, DIRECTIVE 2009/41/EC on the contained use of genetically modified micro-organisms. *Official Journal of the European Union*, 75–97, 2009.
- [27] E. Vanhatalo, M. Kulahci and B. Bergquist, On the structure of dynamic principal component analysis used in statistical process monitoring, *Chemometrics and Intelligent Laboratory Systems*. 167 (2017) 1-11.
- [28] A. Goldbloom, How to Win Kaggle Competitions, *Weights & Biases*, <https://www.youtube.com/watch?v=0ZJQ2Vsgwf0>, visited 21st of July 2022.
- [29] T. Hastie, R. Tibshirani and J. Friedman, 2017, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, 350-352, 2001.
- [30] S. Wold, M. Sjöström and L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*. 58 (2001) 2 109-130.
- [31] L. Breiman, Random Forests, *Machine Learning*. 45 (2001) 5–32.
- [32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Scikit-learn: Machine Learning in Python*, *JMLR*. 12 (2011) 2825-2830.
- [33] C. Guo and F. Berkhahn, Entity Embeddings of Categorical Variables, *CoRR*, abs/1604.06737, 2016.
- [34] Howard et al., 2018, fastai, GitHub, <https://github.com/fastai/fastai>, visited 21st of July 2022.

[35] L. Hongbin, Z. Hao, J. Yi and Z. Fengshan, Online fault detection of complex wastewater treatment process using dynamic kernel PCA, Journal of Jiangsu University. 42 (2021) 2 215-220.

3. Paper 2

Transfer Learning for Quantitative Image Analysis of Biosolids

The paper was submitted to Water Research.

A glossary of terms describing several modelling concepts is presented at the end of section 6.

An online tutorial of presented image clustering is available at:

https://colab.research.google.com/github/waterboy96/ImageAnalysis/blob/main/Quantitative_Image_Analysis_of_Biosolids_Evaluating_morphological_parameters_versus_visual_features.ipynb

Transfer Learning for Quantitative Image Analysis of Biosolids

Sebastian O. N. Topalian¹, Nima Nazemzadeh¹, Alonso Malacara-Becerra^{1,2}, Seyed Soheil Mansouri¹, Kasper Kjellberg³, Damien J. Batstone⁴, Krist V. Gernaey¹, Xavier Flores-Alsina¹, Pedram Ramin¹

¹ Process and Systems Engineering Centre (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark. Building 228 A, 2800, Kgs. Lyngby, Denmark.

²Tecnologico de Monterrey, School of Engineering and Sciences, Monterrey 64849, Mexico.

³ Novozymes A/S, Hallas Alle 1, DK- 4400, Kalundborg, Denmark.

⁴Australian Center for Water and Environmental Biotechnology, The University of Queensland,

4 Gehrman Laboratories Building, Research Rd, St Lucia QLD 4067, Brisbane, Australia.

Abstract

Leveraging data analytics, particularly quantitative image analysis, can boost process efficiency within an industrial context. This study examines such potential in dewatering of stabilized biosolids from a major industrial wastewater treatment plant in Northern Europe. The centrifugal decanters are used for dewatering, with polymers improving separation. The study aims to develop a transparent and systematic analysis workflow encompassing data integration from various sources to predict decanter organic solids recovery. During two dedicated measurement campaigns, data were collected from operational conditions and laboratory measurements. In addition, auxiliary data were collected from image analysis and generated by transfer learning techniques using a readily available online database. Partial Least Squares (PLS) and Random Forest (RF) models were tested using different combination of data sources. The campaign results revealed variable correlation between polymer dosage and organic solids recovery due to complex dynamics of solids characteristics originated from biotech production upstream of the decanters. Following clustering of segmented images of individual particles and predicting recovery with a RF model, the presence of specific crystalline particles was found to be significant, linking important recovery dependency to these particles. The best recovery prediction was obtained using a RF model utilizing both process and laboratory data in combination with transfer learning, improving the prediction by 14% as compared to baseline prediction (using average values). In general, the RF model outperformed the PLS model in predicting recovery, although both models lack consistency in prediction across the organic solids concentration range. Overall, the present study can assist operators to gain deeper insight in the factors influencing dewatering efficiency and can be used as a preliminary tool for system diagnosis. The workflow developed here can be exported to other case studies involving heterogeneous mixtures of particle morphologies.

Keywords

Transfer learning, Dewatering, Sludge, Modelling, Images

Abbreviations

Wastewater treatment plant (WWTP)
Sludge volume index (SVI)
Quantitative Image Analysis (QIA)
Multiple Linear Regression (MLR)
Partial Least Squares (PLS)
Deep Learning (DL)
Convolutional Neural Network (CNN)
Cross-validation (CV)
Transfer Learning (TL)
Polyaluminium Chloride (PAC)
Equivalent circular diameter (ECD)
Principal Component Analysis (PCA)
Random Forest (RF)
Partial Dependence Plot (PDP)
Negative mean square error (NMSE)
Principal components (PCs)
Laboratory measurements (L)
Process measurements (P)

1. Introduction

1.1 Dewatering and Performance Monitoring

Wastewater treatment plays a crucial role in modern society ensuring potable drinking water and safe aquatic environments [1]. A key step in this process is the dewatering of biosolids, where microbes accumulated during secondary treatment are separated from the effluent, accounting for approximately half the cost of wastewater treatment [2]. Centrifugal decanters are commonly used for dewatering due to their high throughput while maintaining a good separation efficiency [3]. To improve separation efficiency, flocculants such as polymer are commonly used which promotes the formation of larger particle aggregates [4]. However, dosing these chemicals presents operational challenges at wastewater treatment plants (WWTP) due to changing process dynamics and biosolid characteristics [5].

A common measure of dewaterability is the sludge volume index (SVI) which serves as an indirect quality measure, partially correlating with the dewatering efficacy of a process. Assuming that there is a relationship between SVI and the process performance, this indirect quality measure can then be utilized for practical process optimization under different conditions, avoiding the cost of full-scale trial and error. However, these indirect quality

measures may not always accurately reflect the behavior observed at full-scale [6,7]. This implies that additional considerations, such as frequent correlation updates, may be necessary if SVI is used as a measure to ensure optimal dewatering performance in real-world scenarios.

Direct quality measures, such as recovery efficiency, demand slightly more resources for acquisition, as both feed and reject streams need to be sampled and analyzed. However, the primary drawback of using direct quality measures is the setback associated with simulating the process under different conditions, as full-scale trial and error can be costly.

1.2 Quantitative Image Analysis within sludge characterization

Microscopy techniques have long been used within wastewater analysis, particularly in identifying morphological characteristics of biomass that are relevant to dewaterability measures such as SVI. However, often the analysis of the microscopy information is labour intensive and requires a trained expert to distinguish for instance between different morphologies that could cause sludge bulking [8, 9]. To automate the analytical procedure quantitative image analysis (QIA) and chemometrics tools have been combined [10]. QIA applies a wide set of mathematical tools to segment images and extract morphological parameters such as length, area and circumference of particles present in the image. The information obtained from QIA can then be utilized to map out relationships with operational conditions using tools such as multiple linear regression (MLR) or partial least squares (PLS) [11]. Most studies that use image analysis for sludge characterization, predict SVI based on morphological parameters such as convexity, compactness, equivalent diameter with the help of either PLS, MLR or Pearson correlation [12-15]. Other studies have investigated the relationship between morphological parameters and settling velocity, bulking conditions, sludge blanket height, and effluent quality, among others [16-19]. In addition, a few studies focused on the image collection process and qualitative characterization, investigating the impact of factors such as dilution and magnification on morphological characteristics [20-22]. A common workflow in these studies is to segment the particles observed within a photo and calculate morphological parameters for each particle. This process results in a distribution of morphological parameters for each photo which can then be summarized using, for instance, a mean to obtain a single metric per sample. This metric can be correlated with the corresponding sample properties of interest, such as SVI. A model can establish the relationship between the images obtained under different operational conditions and the selected performance measures. Such a model can then be utilized to guide and troubleshoot the operations by evaluating control measures at lab scale reducing the drawbacks associated with full scale trial and error. The morphological parameters used are typically derived from an algebraic expression involving the perimeter, width, length and area of the observed particle, and the properties of another shape such as a convex area fitted around the observed particle. The used morphological parameters often only contain information regarding shape, and not the texture (e.g. the roughness) of the observed particles.

1.3 Transfer Learning with Computer Vision Models

In recent years, deep learning (DL) models, particularly convolutional neural networks (CNN) have excelled at computer vision (CV) tasks such as segmentation and classification. Although these approaches have shown great results in many CV tasks not all challenges have been fully solved using these models [23, 24]. DL models exemplify end-to-end learning where they start with the raw input image, internally generate features, and then establish a relationship between these features and the output of interest [25]. The features learned by DL models span a wide range of features, starting from simple edges and vertices and progressing to more complex elements such as meshes and textures, and even abstract concepts such as facial features and composition [26]. One belief of the DL philosophy is that the end-to-end learning approach fosters the generation of richer features. These features are constructed in a supervised manner which allows DL models to uncover higher-level concepts. In contrast, the features generated by human visual taxonomy rely on unsupervised methods, and are in this case constrained to describing shape, and not more advanced concepts. This could potentially be used to distinguish between microbial matter and inorganic matter, as these particles typically have drastically different densities and surface properties that can affect dewatering processes. By leveraging the power of DL to extract meaningful features, researchers and practitioners can potentially enhance their understanding and classification of wastewater particles based on their origin.

In the context of biosolids characterisation, while using DL models is not well adopted, a single use-case of DL models was found for sludge characterization relating to the classification of dispersed and aggregated flocs, as well as classifying the presence or absence of filamentous bacteria [27]. However, DL models are notorious for requiring substantial amounts of labelled data which can be laborious to collect. To alleviate this challenge, transfer learning (TL) aims to apply models already trained on other tasks. This approach eliminates the need for excessive data labelling and developing models from scratch [28]. For CV tasks, DL models can leverage transfer learning using the ImageNet database, which contains a large array of images, making it a valuable resource for pre-training models. Availability of software implementations makes TL more accessible and efficient for implementation [29-31]. This way, DL models can be repurposed and adapted for specific sludge characterization tasks, enhancing their utility and effectiveness.

1.4 Monitoring of stabilized biosolids at an industrial wastewater treatment plant

Majority of the studies focusing on image characterization of sludge primarily deal with activated sludge from municipal plants. However, the industrial wastewater plants exhibit challenges as the characteristics of the produced sludge heavily depend on the production process, with patterns being dictated by the production schedule. In this study, we attempt to characterize stabilized biosolids from the largest industrial WWTP in Northern Europe located in Kalundborg (Denmark) which treats the wastewater produced at two biotech companies. The purpose of sludge characterization is to establish a meaningful correlation between visual

characteristics of the biosolids and the solids content found in the reject stream. By understanding this correlation, we can gain valuable insights into the dewatering process and its efficiency. This is similar to the process of mapping visual characteristics to settling velocities in other dewatering processes. The aim of this study is:

- (i) System characterization and analysis of the dewatering unit in the case study through the implementation of two dedicated measurement campaigns
- (ii) Establish a systematic analysis pipeline/workflow that includes image segmentation, feature generation, data integration and prediction
- (iii) Develop data-driven models to predict organic solids recovery using different data sources.

The systematic characterization and workflow establishment would enable the operators to gain more insights for detailed process monitoring and diagnosis.

2. Methods & Materials

2.1 Plant description

An overview of the plant is presented in figure 1. In this study, the biosolids originate from the production of the two companies and must be stabilized. The stabilization takes place in two storage tanks, S1 and S2, containing biosolids treated with quick lime (CaO) to $\text{pH} > 11$ to secure the stability during storage and to improve the dewatering characteristics. The two storage tanks operate in a semi-continuous fashion, with 4-6 batches processed each day. Following the stabilizing step, the biosolids are dewatered in a decanter hall, where several centrifugal decanters operate in parallel. For a thorough description of the plant layout the reader is referred to our previous work [32-34]. Plant operators manipulate machine settings and adjust polymer dosage in centrifugal decanters to achieve effective separation. Their primary focus is on adjusting the polymer dosage, which they determine through shake-flask tests or trial-and-error methods.

2.2 Measurement Campaign

To gain insight into the dewatering process, two measurement campaigns were conducted: A fall campaign from the 21st of September to 8th of October 2021, and a spring campaign from the 28th of April to the 16th of May 2022. For each batch investigated in the two campaigns, one feed sample of the stabilized biosolids was obtained, along with the three reject samples corresponding to three different polymer dosages: -25%, mean, +25%. The mean polymer dosage was determined by the operators working a given shift. During the fall and spring campaigns, 21 and 12 batches were collected, respectively. It is important to note that each batch was not collected from the same decanter, and notably, the spring campaign was carried out using a decanter from a new manufacturer.

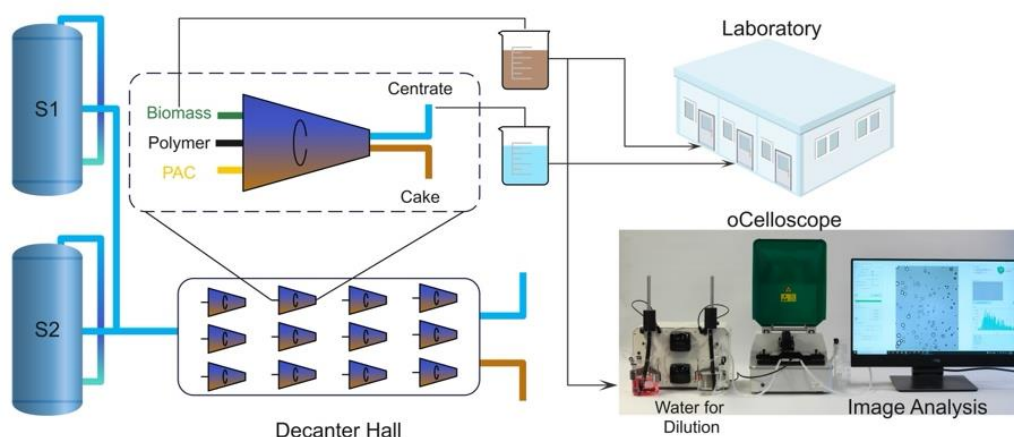


Figure 1. Overview of the plant and the experimental setup in this study. The decanter hall receives stabilized biomass from S1 and S2 tanks. Grab samples were collected from biomass and centrate of a selected decanter, and they were subsequently analysed with the ParticleTech flow cell system.

2.3 Experimental Setup

Stabilized biosolids grab samples were collected from the outflow of the storage tanks. The samples were then analysed for pH, solids content, as well as detailed imaged-based characterization. In addition, operational data, such as flow rates, polyaluminium chloride (PAC) dosage and polymer dosage, were also collected. For the image analysis, to ensure that the images obtained were not overcrowded, a volumetric dilution factor of 20:1 was used following trial and error. Since the operating pH affects the interaction energies of the particles observed in the images, a sodium hydroxide solution was added to adjust the pH to its value before dilution (approx. pH 12). The images were collected and processed off-line using a ParticleTech flow cell system, along with its corresponding software [35]. The morphological parameters for each particle were then recorded including area, equivalent circular diameter (ECD), Feret diameter (minimum, maximum, and mean), the Feret ratio, perimeter length, compactness, and circularity. An overview of the data source and data type used in this study is presented in Table 1. The Experimental protocol is presented in the supplementary information.

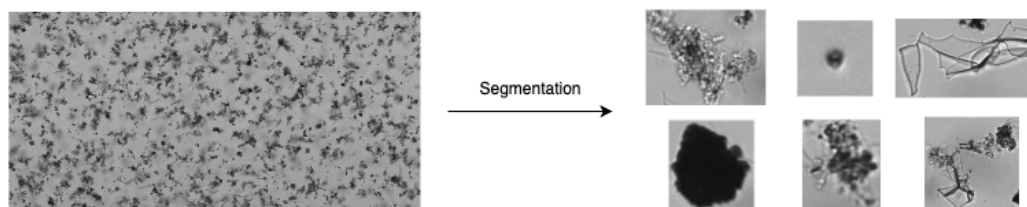
Table 2 – Overview of the operational conditions as well as the laboratory measurements collected.

Data source	Data type							
Operational Conditions	Differential Speed	Revolutions	Feed rate	Flow	Polymer dosage	PAC dosage	Seasonality	
Laboratory Measurements	pH sample	Adjusted pH post dilution	Feed Dissolved Solids (TDS)	Total	Feed Total Suspended Solids (TSS)	Feed Total Solids (TS)		
Image analysis (geometric features)	Area	equivalent circular diameter (ECD)	Feret diameter (min, mean, max)		Feret ratio (Feret min / Feret max)	Perimeter length	Compactness	Circularity

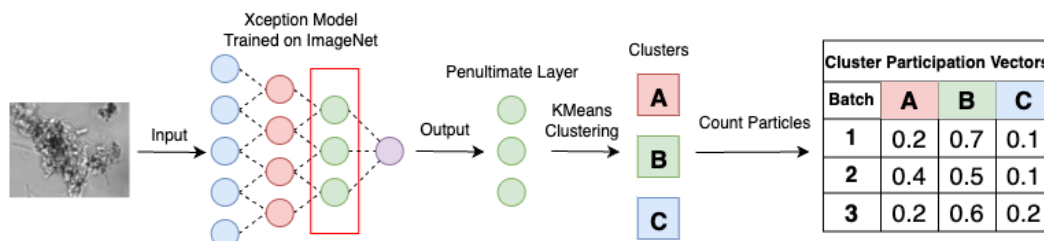
2.4 Modelling Procedure

Figure 1 provides a conceptual overview of the modelling procedure, including three steps. Step 1 involves obtaining and segmenting photos into individual particles. In Step 2, descriptive information is extracted from the images in the form of latent representations, utilizing a pretrained image model (Xception model). With this model it is possible to generate more detailed features (here we used the nodes in the second layer to the last). This information is then used to produce a standardized data format with K-means clustering. The result is membership of the images into recognized clusters, here represented by A, B, C. In step 3, the generated features based on membership can be supplemented with the other sources such as laboratory measurements. The produced dataset is then used in a supervised modelling procedure to estimate the organic solids recovery.

1. Photos are segmented with the ParticleTech software



2. Obtain latent representations from pretrained image model and cluster them



3. Use feature count with laboratory measurements to predict removal efficiencies

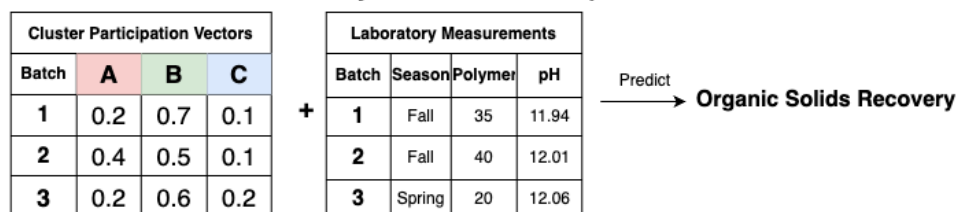


Figure 2 – Analysis pipeline for predicting organic solids recovery based on images and laboratory measurements.

2.4.1 Image Model for Transfer Learning

For generating feature vectors (in addition to the ones listed in Table 1), the convolutional neural network model Xception is selected. This model represents an excellent trade-off between a high accuracy and a low computational demand [36]. While comparing different DL architectures for feature vector generation goes beyond the scope of this study, it is worth noting

that models with higher accuracy on ImageNet typically lead to better transferability to other applications [30]. The use of Xception in this study allows for effective extraction of meaningful features, which are essential for the subsequent analysis and understanding of the data.

2.4.2 Principal Component Analysis & K-means Clustering

Each segmented photo has a set of descriptors based on either classical geometrical properties or the latent representation from the image model (Xception model). Some of these features may be redundant. For instance, the Xception model was trained on coloured photos, and the images used in this study are black and white, latent features that represent changes in colour would exhibit low variances across all the samples.

Principal component analysis (PCA) is commonly used for dimensionality reduction, and in this case, it can create a more compact representation for our feature vectors by eliminating redundant variables. This process not only enhances the efficiency of clustering in the subsequent step but also helps maintain the essential information in a reduced format. For more details on PCA the reader is referred to Wold et al. [37]. It is important to acknowledge that each sample photo contains a varying number of particles, resulting in different feature vectors, due to the varying number of segmented photos, for each sample. For developing a model to predict organic solids recovery, a standard data format is desired. For each sample the data from image segmentation consists of n rows, representing the number of segmented particles, and m columns, representing the different features describing the segmented particle. For the feature vectors obtained through the Xception model, in this study, this would correspond to 2048 columns and for the geometric properties 9 columns. The challenge here is that the number of segmented particles, n , can vary from one sample to another and a common format should be used to describe each sample, i.e., a 1 by m vector. One approach is to average data points across the number of particles, n , for each feature, m . However, in this way, the information on the distribution across particles is lost. Another approach is to cluster the particles and calculate the participation of each cluster in k predefined numbers of clusters. It is calculated as the ratio of the number of particles in a cluster to the total number of particles (n) and would produce a 1 by k vector for each sample as desired. The latter approach is adopted here and this is illustrated in figure 2 where three clusters, $k=3$ (A, B, C), are used as an example. A large variety of clustering algorithms exist and here we use K-means clustering. For details on K-means clustering the reader is referred to Lloyd [38].

2.4.3 Partial Least Squares (PLS) & Random Forest (RF)

To predict the organic solids recovery, a regression tool such as PLS or RF is required. For sludge characterization, PLS is a widely used technique, whereas RF has not been used thus far to the knowledge of the authors. In comparison to PLS, RF is a non-linear technique that can capture non-linear relationships between the cluster participation vectors, laboratory measurements and the organic solids recovery. Like PLS coefficients, RF also provides a feature importance metric, enabling investigation of the independent variable's significance in predicting the organic solids recovery. For further details on PLS and RF, readers can refer to Wold et al. for PLS and Breiman for RF [39, 40].

2.4.5 Model interpretation

PLS models offer easily interpretable linear coefficients for each variable, whereas the non-linear RF models do not. Instead, RF models can be interpreted in terms of feature importance and partial dependence plots (PDP). Feature importance in RF indicates the reduction in model error attributed to each variable. A feature importance of 0 for a variable means the model does not rely on it, while a value of 1 indicates the model solely relies on that variable. Partial dependence plots are model-agnostic and show the predicted value of the model when adjusting a single variable while keeping other variables fixed. These plots display the average response on the test set as a single variable is adjusted.

2.4.6 Batch Leave-one-out Cross-validation & Train-Test Splitting

Both PLS and RF require hyperparameters tuning to avoid overfitting. These parameters were determined through batch leave-one-out cross-validation, where one batch was used for evaluation while the remaining batches were used to train the model. This procedure was repeated for each batch, obtaining a cross-validation score in form of the negative mean square error (NMSE). For PLS, the number of components was optimized, while for RF, the maximum depth was optimized according to the NMSE score. The model with the highest NMSE score was then evaluated on the test set, comprising 8 batches (approximately 25%). The test set was generated in a stratified fashion by randomly selecting 5 out of 21 and 3 out of 12 batches from the fall and spring campaign, respectively, ensuring a similar ratio between batches from both seasons in both training and test sets.

3. Results

3.1 Measurement Campaigns

Figure 3 illustrates the organic solids recovery for different polymer dosage in each batch. For each batch, a second order polynomial has been fitted to visualize the apparent relationship between polymer dosage and the reject quality in terms of organic solids recovery. A concave curve (e.g. batch 4) demonstrates the trade-off between destabilizing the suspension initially and stabilizing it with a sufficiently high polymer dosage, making the separation infeasible. It is noteworthy that the observed variation among the curves is linked to changes in biotech production upstream from the WWTP.

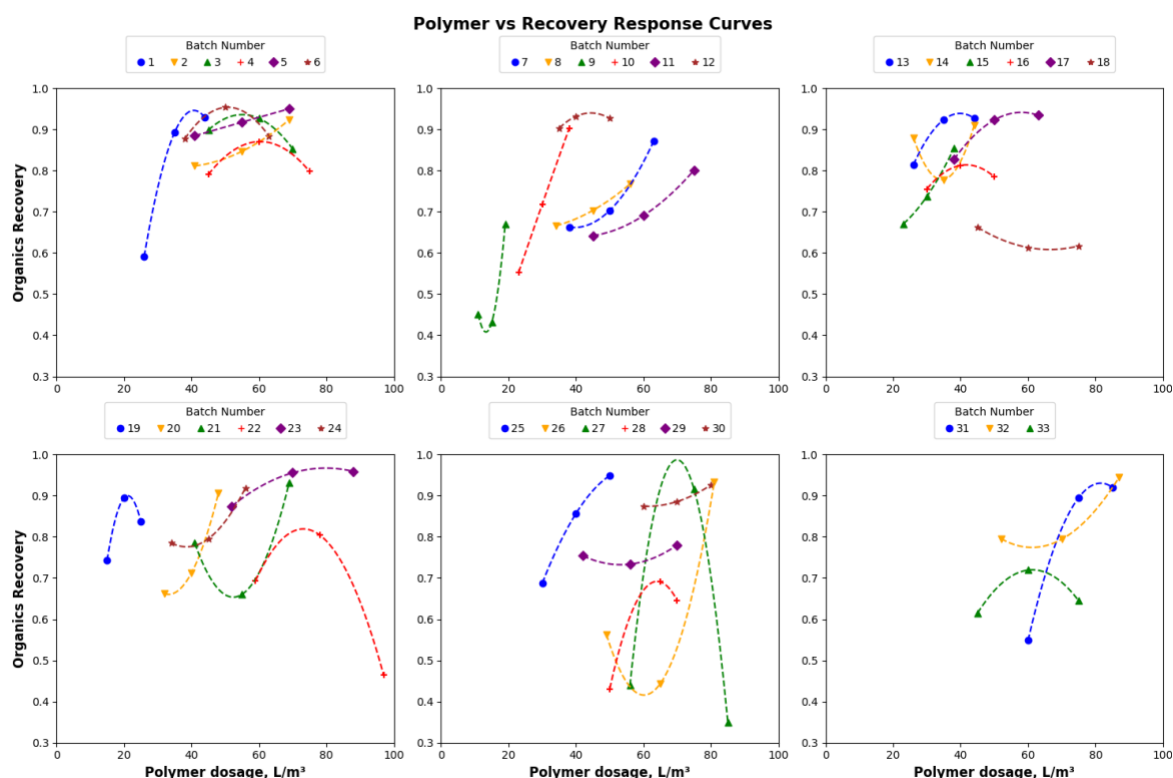


Figure 3 – Organic solids recovery at different polymer dosages. Each parabola corresponds to a polynomial fit based on the three reject samples collected from each batch. Batches 1-21 and 22-33 are from the fall and spring campaigns, respectively.

3.2 PCA of Xception and Geometric Feature Vectors

The photos of segmented particles from the feed sample of the training batches are fed to the Xception model to obtain one feature vector per photo from the penultimate model layer. These feature vectors are reduced into principal components (PCs) using PCA. Geometric features are also reduced into PCs. **Figure 4** displays two plots for the PCA analysis depicting the cumulative variance explained ratio versus the number of PCs for the Xception and geometric feature vectors. A threshold is selected to retain 99% of the variance, which corresponds to 107 PCs for the Xception feature vectors, reducing the dimensionality from 2048 to 107 (95% reduction), and 4 PCs for the geometric feature vectors, reducing the dimensionality from 9 to 4 (55% reduction). The threshold of 99% is chosen, to retain nearly all variance while significantly reducing the number of variables to speed up the clustering process in the subsequent section.

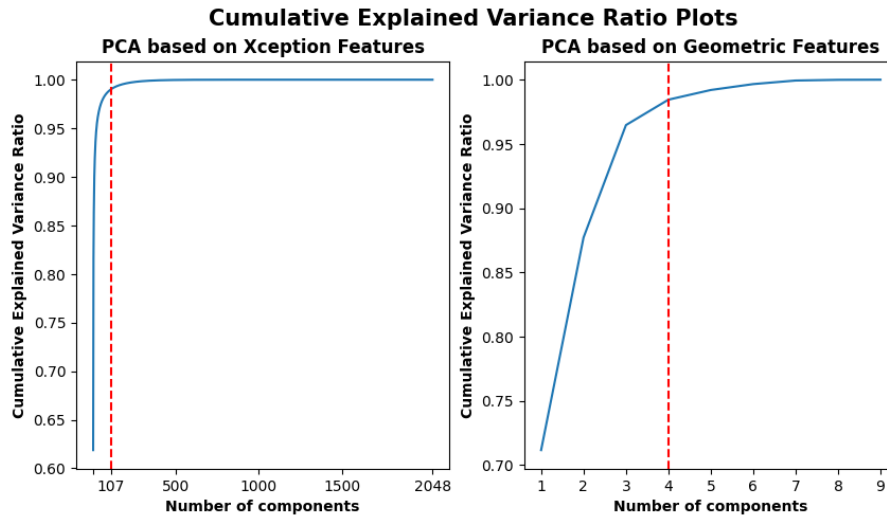


Figure 4 – Cumulative explained variance ratio plots for PCA on the Xception features (left) and Geometric features (right) for the 25 training batches. The dashed red lines indicate the cut-off point at 107 components for the Xception features and 3 components for the Geometric features, chosen to explain 99% of the cumulative variance.

3.3 K-means clustering

Once the thresholds of 107 and 4 PCs are determined, K-Means clustering was performed. The number of clusters evaluated varied from 1 to 25. For each cluster, the inertia score was calculated, which measures the distance from each observation in the cluster to its centroid. As the number of clusters increases, the inertia tends to decrease since smaller clusters are formed. However, beyond a certain point, additional clusters provide little performance gain, making further subdivision redundant.

Figure 5 displays the min-max scaled inertia plotted against the number of clusters. Beyond 10 clusters, the inertia decreases at a relatively constant rate, leading to the selection of 10 clusters as a compromise to avoid redundant clusters and unnecessary information in the subsequent model. Using the training data, each sample's segmented particles are then assigned to the clusters creating a participation vector that describes the proportion of each cluster in the overall composition of particles.

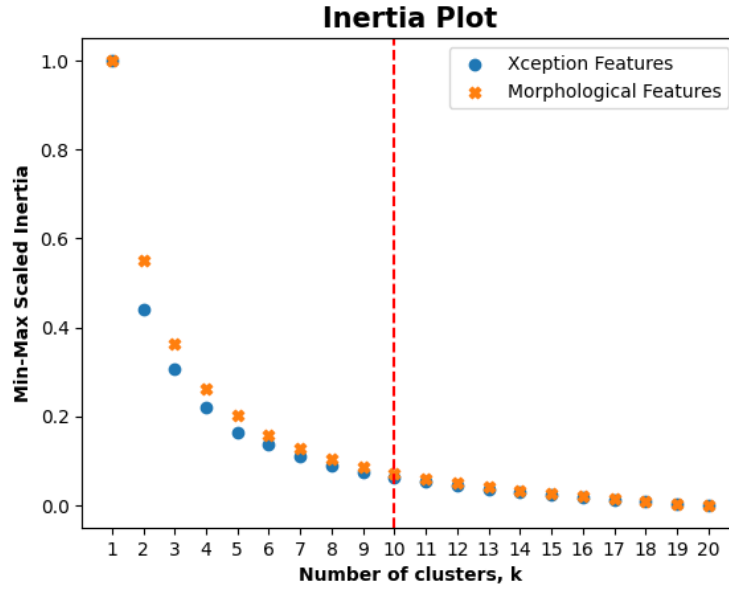


Figure 5 – Min-max scaled inertia plotted against the number of clusters in K-means clustering. The data for this analysis are from two data sources: principal components of Xception (Blue dots) and geometric feature vectors (Orange crosses).

3.4 Supervised modelling with RF

Apart from the cluster participation vectors based on the Xception feature vectors and the geometric properties, another image description was used which is the means of the geometric properties for each sample. Each of these three types of image information were used for constructing regression models with either no auxiliary information (None), laboratory measurements (L), process measurements (P) or both laboratory and process measurements (L+P). For each combination, both RF and PLS model were constructed and trained using leave-one-out cross-validation on the training set. The hyperparameters, maximum depth, and number of components for RF and PLS, respectively, were validated during this process. The performance of the models was then evaluated in terms of root mean square errors (RMSE) on the 8 batches not used for training and cross-validation. Table 2 displays the test set RMSE for the RF and PLS models, respectively.

Table 3 – RMSE of RF/PLS models on the test batches for each combination of image processing technique and auxiliary information such as laboratory measurements or process information. * Denotes the RMSE of estimating that the predicted recovery is the average recovery of the training set.

RF/PLS Models	Image Information				
		None	Xception Clusters	Geometric Means	Geometric Clusters
Auxiliary Information	None	0.160*	0.151/0.157	0.144/0.165	0.195/0.206
	Laboratory (L)	0.149/0.174	0.143/0.167	0.139/0.189	0.152/0.218
	Process (P)	0.192/0.182	0.144/0.160	0.146/0.197	0.174/0.180
	Both (L+P)	0.145/0.212	0.138/0.183	0.142/0.204	0.144/0.207

In this analysis, twelve out of fifteen and one out of fifteen RF and PLS models, respectively, display a lower expected error than the baseline. This indicates that the non-linear RF model is better suited than PLS for modelling the organic recovery rates. In terms of image information, the model that obtains the lowest RMSE is the one built with Xception feature vectors and both process and laboratory information. These findings highlight the relative performance of the different models and their respective image processing techniques, underscoring the potential advantages of using Xception-based cluster participation vectors in certain scenarios while noting the limitations of PLS models in this specific context.

Figure 6 shows the parity plots for the models utilising all auxiliary information and the three different image processing methods. Batch 9 and 26 exhibit large prediction errors which could indicate that these batches are outliers.

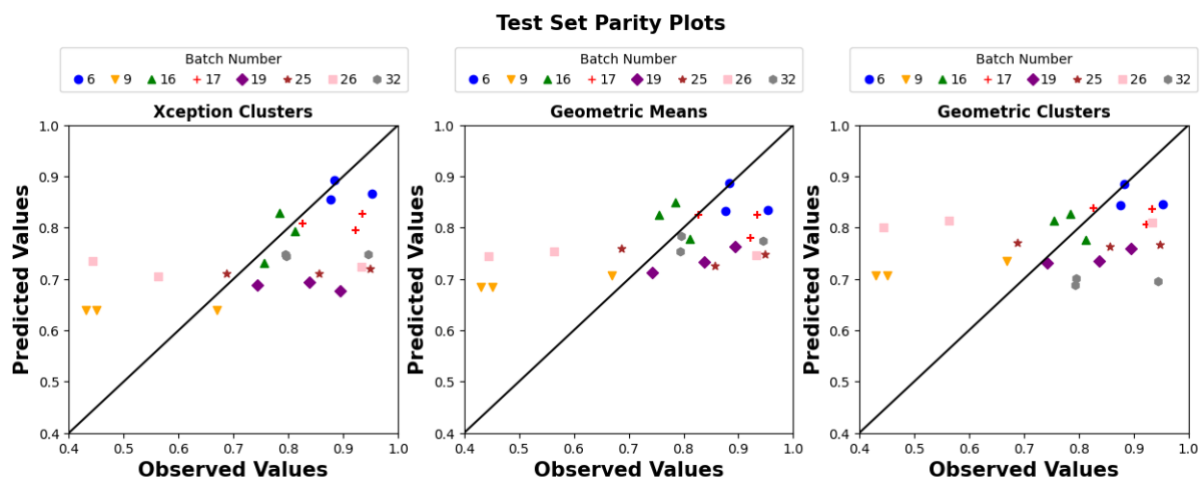


Figure 6 – Parity plots for the three different random forest models using laboratory and process measurements as well as the Xception participation vectors (left), geometric means (middle), and geometric (right) clusters respectively.

3.5 Model Interpretation

Figure 7 shows three subplots corresponding to the feature importance of the auxiliary information, cluster features, and geometric means for the three RF models using all auxiliary information and the three different image processing methods. Polymer, pH, feed TSS, feed TDS and to an extent the differential speed of the decanter play a large role, whereas the feed flow, rotational speed and PAC dosage play a lesser role. The seasonality appears to have little to no effect in each case. A reason for this could be that the impact the seasonality would have, would be a change to the dosage of lime in the upstream stabilizing process which would change the TDS content due to the solubility of lime being higher during colder parts of the year. The model using geometric clusters seems to highly utilise one type of cluster (cluster 5) whereas the feature importance in the Xception clusters is more evenly distributed. For the geometric means the features hold a lower feature importance than the auxiliary information in general with the ferret ratio being the most prominent feature.

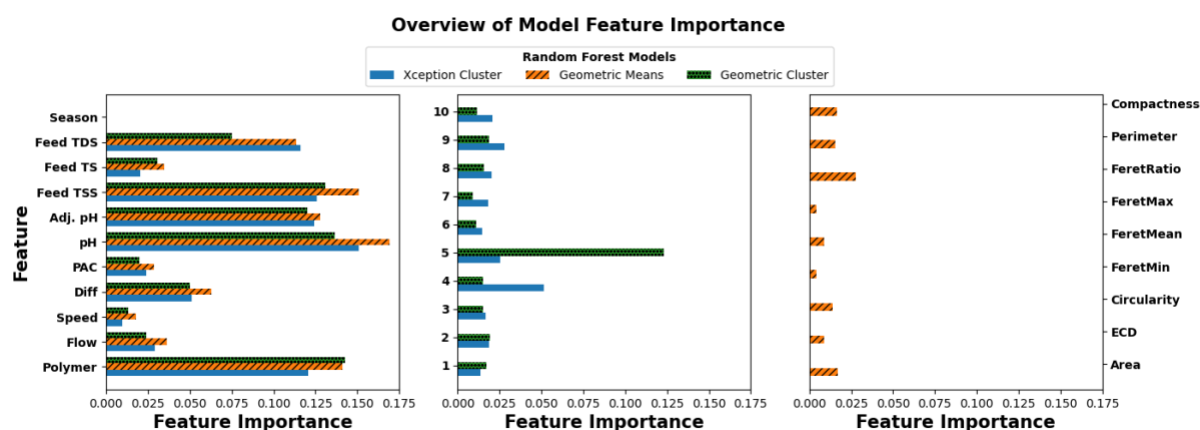


Figure 7 – Overview of the RF feature importance for the 3 different models. The feature importance of the auxiliary information, cluster features and geometric properties left, center and right respectively.

Figure 8 shows the partial dependence plots for the RF model, utilising the Xception clusters and all the auxiliary information. Of particular interest is the partial dependence plot for polymer which shows that an increase in polymer is expected to increase recovery until a threshold (around 70 l/m³), beyond which the recovery declines. The plot for the flow reveals a nearly linear inverse relationship: higher flow, lower recovery. Increasing the differential speed appears to increase the expected recovery in a relationship resembling an S curve. The pH and the adjusted pH plots suggest that higher pH leads to worse recovery. The feed TS appears to have no effect on the expected recovery compared to the feed TSS and TDS, which are linked to positive and negative relationships with recovery, respectively. Many of the clusters exhibit less pronounced effects on the recovery, potentially due to partial autocorrelation with the laboratory measurements. For instance, the feed TDS and the cluster 3, with a correlation coefficient of 0.95, differ in impact. Table 3 and 4 show example images from the Xception and geometric clusters, respectively. The example images are taken by analysing a random batch and looking at their corresponding labels. The Xception clusters appear to better group what looks like crystalline particles in cluster 3, which is the cluster highly correlated with TDS, whereas the geometric clusters show less groupings overall.

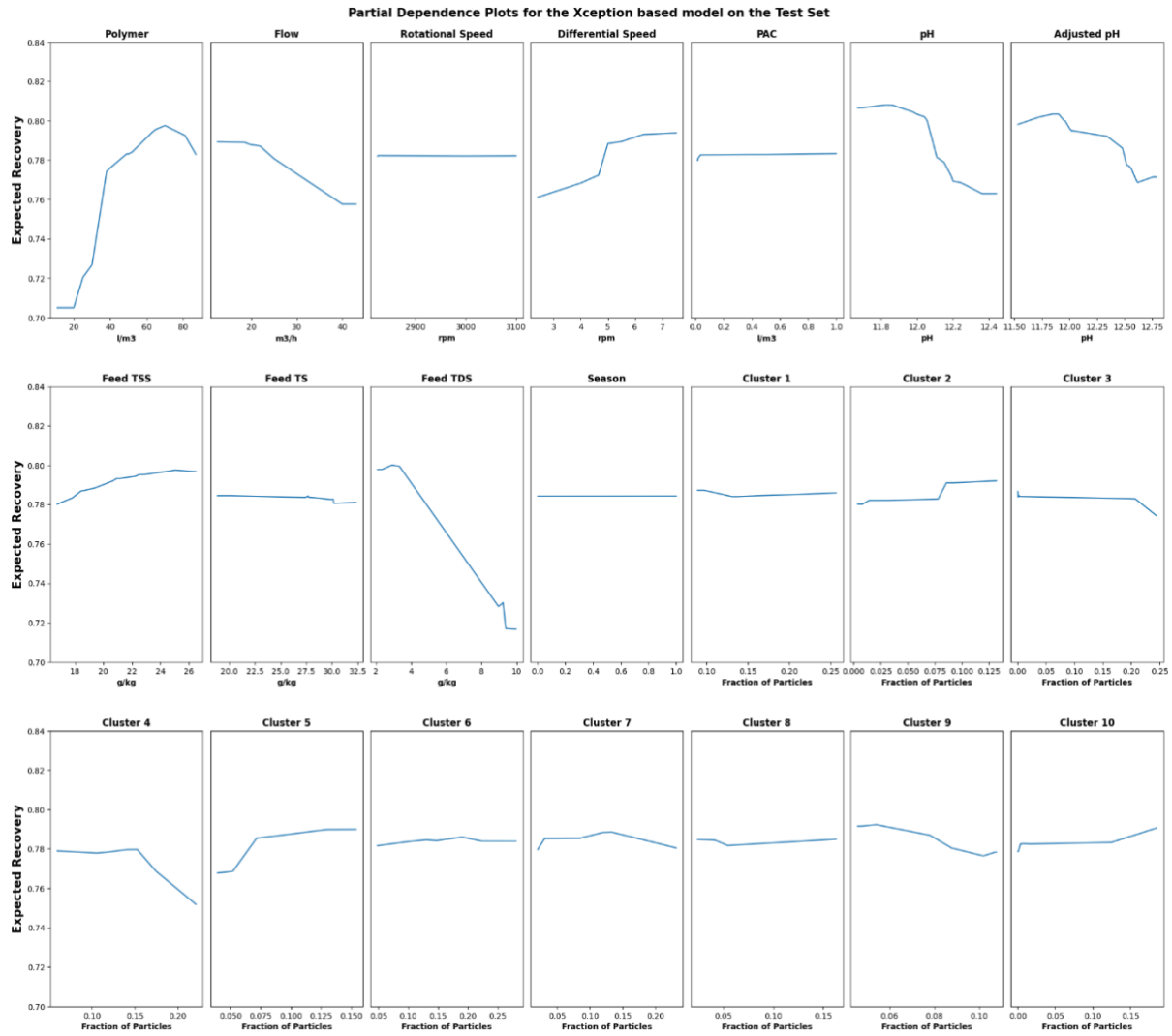


Figure 8 – Overview of the partial dependence plots for the RF model using all auxiliary information (i.e. laboratory and process) and clusters built on the Xception feature vectors. The partial dependence plots show the average expected change to the recovery as each variable is perturbed one at a time for the test data set.

Table 4 – Randomly sampled images from a batch belonging to each of the 10 different Xception clusters.

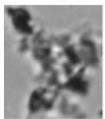
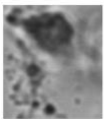


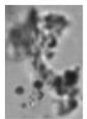
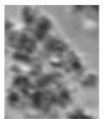
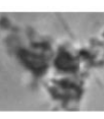
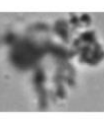
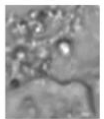
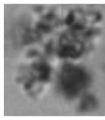
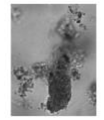
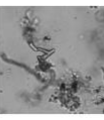
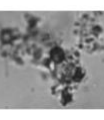

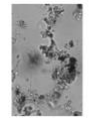
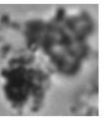
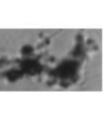
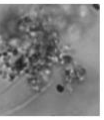
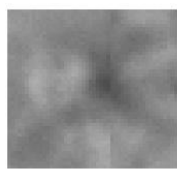
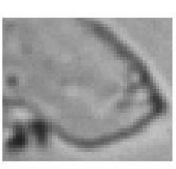
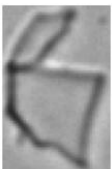
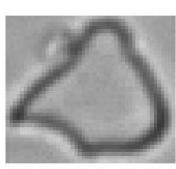
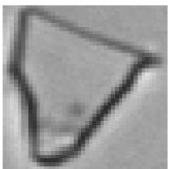


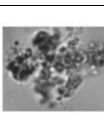
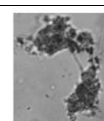
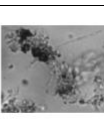

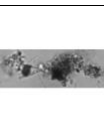
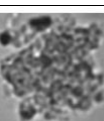

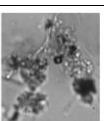

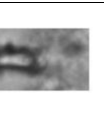
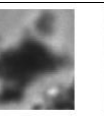
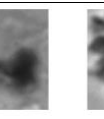
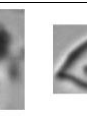
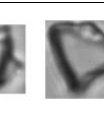
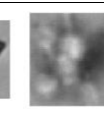
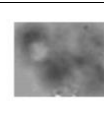
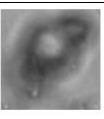
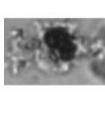
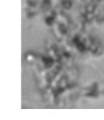

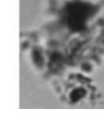
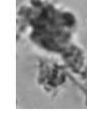
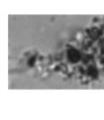

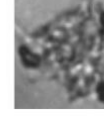
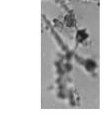


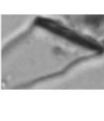
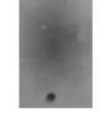

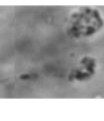
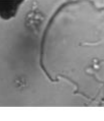
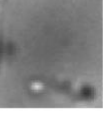



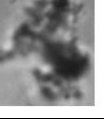

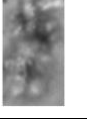
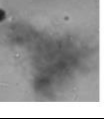
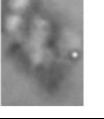
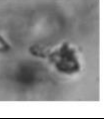

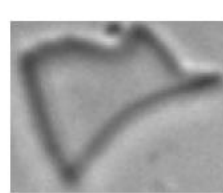

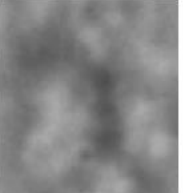



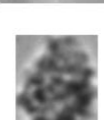






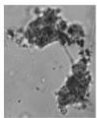

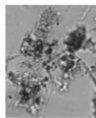
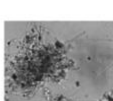
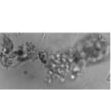
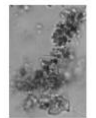
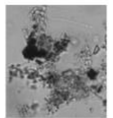

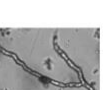

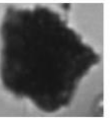
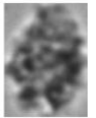
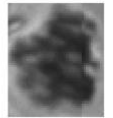
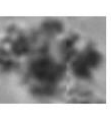
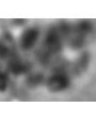

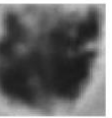
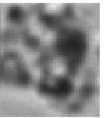
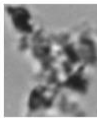
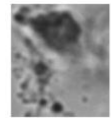
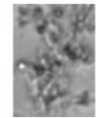
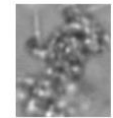
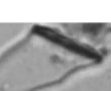
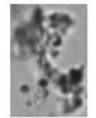
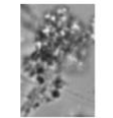
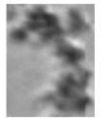
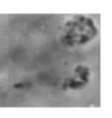
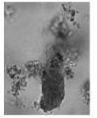
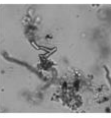
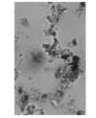
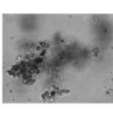
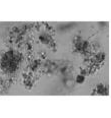
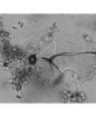
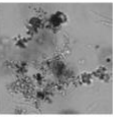
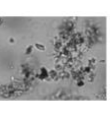
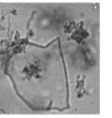
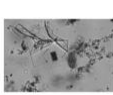
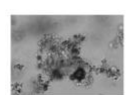

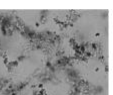
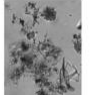
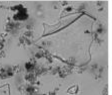
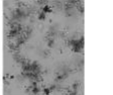
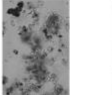
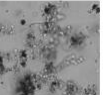
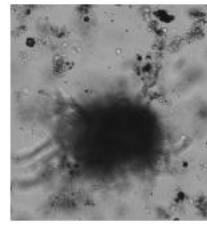

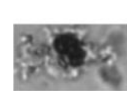
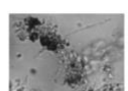

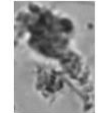

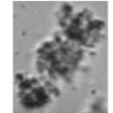
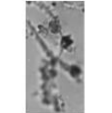
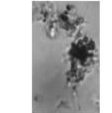
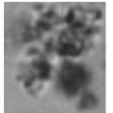
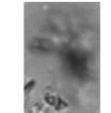
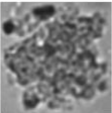

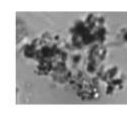
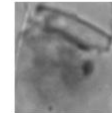

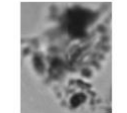
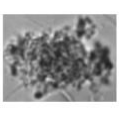

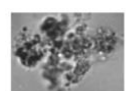

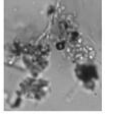
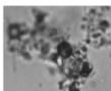
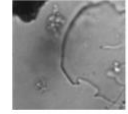
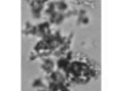


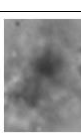

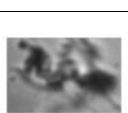
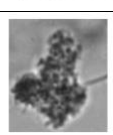



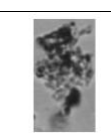

1									
2									
3									
4									
5									
6									
7									
8									
9									
10									

Table 5 - Randomly sampled images from a batch belonging to each of the 10 different Geometric clusters.

1									
2									
3									
4									
5									
6									
7									
8									
9									
10									

4. Discussion

Significant advances have been made within the field of DL-based computer vision, and newer models than the one used here are now available. However, it is important to note that larger

models with higher accuracy also come with a steep increase in the number of parameters compared to the gain in accuracy, making them more computationally demanding [21].

If certain species are of interest, such as cluster 3 for predicting TDS, this approach can be used for development of an initial identification system. The system can be fine-tuned with labels if needed. For instance, in this study, the crystalline particles displayed are of particular interest to the company. Similarly, other industrial wastewater treatment plants may have species with specific impacts on their dewatering operations, making this approach valuable for their monitoring as well. Estimating recovery in centrifugal decanters based solely on laboratory measurements is challenging, as machine operation significantly affects results. Many methods aiming to extrapolate lab to full-scale decanter dewatering overlook this operational influence [41]. The errors seen in the developed models were high (see Figure 6), which we attribute to the challenging nature of the data, particularly because one of the decanters was being commissioned during the spring campaign. Outliers possibly emerged due to extreme reject qualities, yet they were retained for analysis. No outlier analysis was pursued, as model interpretation remained satisfactory. We expect that applying the same setup to estimate settling velocities or capillary suction time, instead of reject quality in centrifugal decanters, would likely yield similar or even better results as the impact of the machinery does not play a role. The methodology presented here can be readily adapted to other industrial or municipal systems for biomass characterisation and prediction of settling properties, or potentially fermentation or crystallisation processes. The clustering and deep learning treatment is believed to be only beneficial for particle processes with several different morphologies, whereas geometric features are believed to be adequate for mixtures with only one morphology.

The developed framework presents a methodology for predicting recovery, aiding decisions via sample collection and stabilized sludge property analysis. This guides polymer dosage optimization to enhance expected recovery. The imaging system and sample collection is fully automatable in contrast to conventional gravimetric analysis. Coupled with adequate analysis, this approach demonstrates a framework that streamlines image-to-model analysis, enabling potential process automation, which can save operators considerable time. In terms of method development, future work could look into integrating the entire pipeline, such that everything from the extracted features to the clustering is affected by the downstream task of predicting recovery rates.

5. Conclusion

We present a novel framework for utilising the power of DL vision models to extract features and track species in stabilized biosolids. These features are then used to predict reject quality in an industrial setting.

- Image analysis and deep learning have been combined to cluster particles into different clusters for developing a characterization system for each batch which can be used for supervised modelling of recovery efficiencies
- Among the developed models, linear PLS models exhibited no potential, but nonlinear RF models, particularly the Xception-based cluster model, outperformed those built on traditional geometric properties and clusters.
- Random Forest partial dependence plots revealed key relationships between process parameters and the expected recovery, which can then be used to increase process understanding.
- Model errors were relatively high, and we suspect this is due to the commissioning of a new decanter while collecting the data set and the inherent complexity of predicting properties in centrifugal decanters.

6. Software & Data Availability

An example of the generation of the participation coefficient vectors is available via the wwtmodels Github repository, and for undisclosed sections the reader is encouraged to contact the first author by email: sebttop@kt.dtu.dk. A sample of the image data is made available for distribution.

7. Glossary

Convolutional Neural Networks (CNNs): Convolutional Neural Networks (CNNs) are deep learning algorithms specifically designed for analyzing visual data. CNNs employ convolutional layers to capture local patterns, pooling layers to reduce dimensionality, and fully connected layers to make predictions based on learned representations. By leveraging these specialized layers, CNNs excel at tasks like image classification, object detection, and image segmentation.

Computer Vision (CV): Computer vision is a field of artificial intelligence and computer science that focuses on enabling computers to interpret and understand visual information from digital images or videos. It involves developing algorithms and techniques to extract meaningful insights, recognize patterns, and make sense of visual data.

Cross-validation: Cross-validation is a process that involves partitioning the available data into multiple subsets or folds. The model is then trained and evaluated multiple times, using different combinations of these subsets.

Deep learning (DL): Deep learning is a subfield of artificial intelligence (AI) that focuses on training and building neural networks to learn and make intelligent decisions from vast amounts of data. Deep learning algorithms learn to recognize patterns and extract meaningful insights from data through multiple layers of interconnected nodes or neurons. This approach has proven highly effective in tasks such as image and speech recognition.

Feret diameter: The Feret diameter provides an estimate of the object's size in a particular direction, capturing its maximum extent. The Feret diameter is determined by placing two perpendicular lines (calipers) across an object's 2D image or silhouette. The Feret diameter is

the maximum distance between these calipers, representing the object's size or width in a specific direction.

ImageNet Database: is a database of images organized according to the WordNet hierarchy, where each node in the hierarchy is represented by hundreds or thousands of images. ImageNet is widely used for training and evaluating image classification models.

Inertia in K-Means: In K-Means clustering, inertia measures how tightly data points are clustered around their centroids, aiming to minimize this value for well-defined clusters. It is commonly calculated by summing the squared distances between data points and their centroid within one cluster.

K-Means clustering: it is an unsupervised machine learning algorithm used to group data points into K clusters based on their similarities. It iteratively assigns data points to the nearest cluster centroid and updates the centroids by calculating the mean of the points within each cluster. The process continues until convergence, resulting in well-defined clusters with minimized variance within each cluster.

Latent representation: In machine learning, a latent representation refers to a learned, compressed, or hidden representation of input data that captures the underlying structure and relevant features of the data. It is often obtained through an intermediate layer or set of layers in a neural network

Partial dependence plots (PDPs): In machine learning clustering tasks, they show how predicted outcomes (cluster membership) change with variations in specific features while keeping others fixed. PDPs help understand the impact of individual features on cluster formation, aiding in feature selection and identifying influential factors.

Partial Least Squares (PLS): PLS seeks to find a set of latent variables, known as components, that capture the maximum covariance between the predictor variables and the response variables. Unlike traditional regression methods, PLS performs a simultaneous decomposition of both the predictor and response variables, allowing for effective modeling of the interdependencies.

Principle Component Analysis (PCA): Principal Component Analysis (PCA) is a dimensionality reduction technique used in statistics and machine learning. Its main objective is to transform a high-dimensional dataset into a lower-dimensional representation while preserving the most important information. It is commonly used for exploratory data analysis, feature extraction, and data compression.

Random Forest (RF): In a Random Forest, a collection of decision trees is created using a technique called bootstrap aggregating, or bagging. Each decision tree is trained on a randomly sampled subset of the training data, and during the training process, each tree is provided with different subsets of features. This randomness introduces diversity among the individual trees, making them less prone to overfitting and improving their generalization capability. To make predictions with a Random Forest, the input data is passed through each decision tree, and each tree independently produces a prediction. The final prediction is determined through a voting or averaging process, where the predictions from all the trees are combined to reach a consensus.

Sludge volume index (SVI): Sludge Volume Index (SVI) is a measurement used in wastewater treatment to assess the settling characteristics of activated sludge. It provides an indication of the sludge's ability to separate from the treated wastewater during the settling process. A low

SVI indicates good settling characteristics, meaning the sludge settles quickly, and the treated wastewater can be effectively separated from the sludge

Transfer Learning (TL): Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a similar but different task. This technique reduces the amount of data and computational resources required to train the task at hand.

8. Acknowledgements

The authors would like to acknowledge the financial and academic support of the Technical University of Denmark and Novozymes for their contributions to the project. Dr. Ramin and Dr. Flores-Alsina thank the support of the Danish Innovation Fund under the project GREENTAN (Contract-No: 0177-001009A). Dr. Flores-Alsina and Prof. Gernaey are grateful to the Miljøministeriets Miljøteknologisk Udviklings og Demonstrationsprogram (MUDP) project NACAT (Contract-No: 2021-20015). Part of this research was conducted when Mr. Topalian was an academic visitor at the Australian Center for Water and Environmental Biotechnology (ACWEB). Mr. Topalian would like to acknowledge Idella Foundation for their financial support which enabled his stay at ACWEB. Mr. Malacara-Becerra is grateful to the Helix Lab Fellowship program, to the Consejo Nacional de Ciencia y Tecnología (CONACyT) and to Tecnológico de Monterrey for providing the financial and academic support during his participation in this research.

9. References

- [1] Progress on wastewater treatment – Global status and acceleration needs for SDG indicator 6.3.1. United Nations Human Settlements Programme (UN-Habitat) and World Health Organization (WHO), Geneva, 2021.
- [2] O. Nowak, Optimizing the Use of Sludge Treatment Facilities at Municipal WWTPs, *Journal of Environmental Science and Health Part A*. 41 (2006) 9 1807-1817.
- [3] R. J. Wakeman, Separation technologies for sludge dewatering, *Journal of Hazardous Materials*. 144 (2007) 3 614-619.
- [4] Z. Yan, B. Örmeci, J. Zhang, Effect of Sludge Conditioning Temperature on the Thickening and Dewatering Performance of Polymers, *Journal of Residuals Science & Technology*. 13 (2016) 3 215-224.
- [5] A. Records and, K. Sutherland, *Decanter Centrifuge Handbook*, Elsevier: Amsterdam, The Netherlands, 2001.
- [6] P. Ginisty, Laboratory tests to optimize sludge coagulation / flocculation process before thickening or dewatering, *American Filtration and Separations Society 2005 – 18th Annual Conference*, Afs, (2005).
- [7] V. H. P. To, T. V. Nguyen, S. Vigneswaran, H. H. Ngo, A review on sludge dewatering indices, *Water Science & Technology*, 74 (2016) 1 1-16.
- [8] D. P. Mesquita, A. L. Amaral and E. C. Ferreira, Activated sludge characterization through microscopy: A review on quantitative image analysis and chemometric techniques, *Analytica Chimica Acta*. 802 (2013) 14-28.
- [9] E. Koivuranta, T. Stoor, J. Hattuniemi, J. Niinimäki, On-line optical monitoring of activated sludge floc morphology, *Journal of Water Process Engineering*, 5 (2015) 28-34.
- [10] M. B. Khan, H. Nisar, C. A. Ng, P. K. Lo, V. V. Yap, Generalized classification modeling of activated sludge process based on microscopic image analysis, *Environmental Technology*, 39 (2018) 24-34.
- [11] J. C. Costa, D. P. Mesquita, A. L. Amaral, M. M. Alves and E. C. Ferreira, Quantitative image analysis for the characterization of microbial aggregates in biological wastewater treatment: a review, *Environmental Science and Pollution Research*. 20 (2013) 5887–5912.
- [12] M. da Motta, M.-N. Pons, N. Roche, Study of filamentous bacteria by image analysis and relation with settleability, *Water Science & Technology*. 46 (2002) 1-2 363-369.
- [13] D. P. Mesquita, O. Dias, A. L. Amaral, E. C. Ferreira, Correlation between sludge settling ability and image analysis information using partial least squares, *Analytica Chimica Acta*. 642 (2009) 1-2 94-101.
- [14] R. Jenné, E. N. Banadda, I. Smets, J. Deurinck, J. Van Impe, Detection of filamentous bulking problems: Developing an image analysis system for sludge composition monitoring, *Microscopy and Microanalysis*. 13 (2007) 36-41.
- [15] C. Leal, A. V. del Río, D. P. Mesquita, A. L. Amaral, E. C. Ferreira, Prediction of sludge settleability, density and suspended solids of aerobic granular sludge in the presence of

- pharmaceutically active compounds by quantitative image analysis and chemometric tools, *Journal of Environmental Chemical Engineering*. 10 (2022) 2 107136.
- [16] C. Leal, M. Lopes, A. V. del Río, C. Quintelas, P. M. L. Castro, E. C. Ferreira, A. L. Amaral, D. P. Mesquita, Assessment of an aerobic granular sludge system in the presence of pharmaceutically active compounds by quantitative image analysis and chemometric techniques, *Journal of Environmental Management*. 289 (2021) 112474.
 - [17] D. Kang, D. Xu, T. Yu, C. Feng, Y. Li, M. Zhang, P. Zheng, Texture of anammox sludge bed: Composition feature, visual characterization and formation mechanism, *Water Research*. 154 (2019) 180-188.
 - [18] E. Koivuranta, T. Stoor, J. Hattuniemi, J. Niinimäki, On-line optical monitoring of activated sludge floc morphology, *Journal of Water Process Engineering*. 5 (2015) 28-34.
 - [19] N. Derlon, C. Thürlimann, D. Dürrenmatt, K. Villez, Batch settling curve registration via image data modeling, *Water Research*. 114 (2017) 327-337.
 - [20] D. P. Mesquita, O. Dias, R. A. V. Elias, A. L. Amaral, E. C. Ferreira, Dilution and magnification effects on image analysis applications in activated sludge characterization, *Microscopy and Microanalysis*. 16 (2010) 5 561-568.
 - [21] D. Snidaro, F. Zartarian, F. Jorand, J.-Y. Bottero, J.-C. Block, J. Manem, Characterization of activated sludge flocs structure, *Water Science & Technology*. 36 (1997) 4 313-320.
 - [22] M. Schmid, A. Thill, U. Purkhold, M. Walcher, J. Y. Bottero, P. Ginestet, P. H. Nielsen, S. Wuertz, M. Wagner, Characterization of activated sludge flocs by confocal laser scanning microscopy and image analysis, *Water Research*. 37 (2003) 9 2043-2052.
 - [23] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, 2018. Deep Learning for Computer Vision: A Brief Review, *Computational Intelligence and Neuroscience*, 7068349.
 - [24] N. O' Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, Lenka Krpalkova, D. Riordan, J. Walsh, Deep Learning vs. Traditional Computer Vision, *Advances in Computer Vision Proceedings of the 2019 Computer Vision Conference (CVC)*. Springer Nature Switzerland AG (2019) 128-144.
 - [25] J. Wang, Y. Ma, L. Zhang, R. X. Gao, D. Wu, Deep learning for smart manufacturing: Methods and applications, *Journal of Manufacturing Systems*. 48 (2018) C 144-156.
 - [26] M. D. Zeiler, R. Fergus, Visualizing and Understanding Convolutional Networks, *European conference on computer vision*. (2014) 818-833.
 - [27] H. Satoh, Y. Kashimoto, N. Takahashi, T. Tsujimura, Deep learning-based morphology classification of activated sludge flocs in wastewater treatment plants, *Environmental Science: Water & Technology*. 7 (2021) 2 298-305.
 - [28] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A Comprehensive Survey on Transfer Learning, *Proceedings of the IEEE*. 109.1 (2020) 43-76.
 - [29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. (2009) 248–255.

- [30] S. Kornblith, J. Shlens, Q. Le, Do Better ImageNet Models Transfer Better?, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 2661-2671.
- [31] F. Chollet et al., Keras, 2015. <https://keras.io>, visited 6th of January 2023.
- [32] S. O. N. Topalian, P. Ramin, K. Kjellberg, C. K. Mbamba, D. J. Batstone, K. V. Gernaey, X. Flores-Alsina, 2023. A data analytics pipeline to optimise polymer dose strategy in a semi-continuous multi-feed dewatering system, Journal of Water Process Engineering. 55 104048.
- [33] V. Monje, H. Junicke, D. J. Batstone, K. Kjellberg, K. Gernaey and X. Flores-Alsina, 2022. Prediction of mass and volumetric flows in a full-scale industrial waste treatment plant, Chemical Engineering Journal. 445 136774.
- [34] V. Monje, M. Owsianiak, H. M. Junicke, K. Kjellberg, K. V. Gernaey & X. Flores-Alsina, 2022. Economic, technical, and environmental evaluation of retrofitting scenarios in a full-scale industrial wastewater treatment system, Water Research. 223 14 118997.
- [35] Image Equipment Utilised form ParticleTech, <https://particletech.dk>, visited 9th of January 9, 2023.
- [36] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1251-1258.
- [37] S. Wold, Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems. 2 (1987) 1-3 37-52.
- [38] S. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory. 28 (1982) 2 129-137.
- [39] S. Wold, M. Sjöström and L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems. 58 (2001) 2 109-130.
- [40] L. Breiman, Random Forests, Machine Learning. 45 (2001) 5–32.
- [41] P. Ginisty, Laboratory tests to optimize sludge coagulation / flocculation process before thickening or dewatering, American Filtration and Separations Society 2005 – 18th Annual Conference, Afs, (2005).

4. Paper 3

Fifty Shades of Foam: Validation of a camera-based total suspended solids soft-sensor for industrial dewatering of biosolids.

The paper was submitted to the Journal of Water Process Engineering.

Fifty Shades of Foam: Validation of a camera-based total suspended solids soft-sensor for industrial dewatering of biosolids

Sebastian Olivier Nymann Topalian¹, Alex Pokhrel¹, Alonso Malacara-Becerra^{1,2}, Kasper Rehn³, Kasper Kjellberg³, Pedram Ramin¹, Seyed Soheil Mansouri¹, Damien J. Batstone⁴, Krist V. Gernaey¹, Xavier Flores-Alsina¹

¹ Process and Systems Engineering Centre (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark. Building 228 A, 2800, Kgs. Lyngby, Denmark.

²Tecnologico de Monterrey, School of Engineering and Sciences, Monterrey 64849, Mexico.

³ Novozymes A/S, Hallas Alle 1, DK- 4400, Kalundborg, Denmark.

⁴Australian Center for Water and Environmental Biotechnology, The University of Queensland, 4 Gehrmann Laboratories Building, Research Rd, St Lucia QLD 4067, Brisbane, Australia.

***Corresponding author:**

Xavier Flores-Alsina,

Email: xfa@kt.dtu.dk

Address: Technical University of Denmark, Department of Chemical and Biochemical Engineering, Søtofts Plads, Building 227 (Postal address: Building 228A), 2800 Kgs. Lyngby, Denmark

Abstract

The use of optical back scattering (OBS) devices such as turbidity sensors for measuring solids content in water application is wide-spread, however for applications where phase transitions occur such as between the liquid phase and air bubbles, the air bubbles can lead to considerable measurement errors. Here we present the validation of a setup with a partial degassing vessel combined with a colour-camera for monitoring solids of an industrial dewatering process in terms of reject solids. The setup was validated in two different iterations in two different measurement campaigns. The sensor works in the presence of air with strong intraday linear relationships between the total suspended solids (TSS) and camera signals (R^2 : 0.53-0.98). Finally, a proposed explanation is given as to the

mechanism by which the solution works and maintenance requirements are discussed.

Highlights

1. A standard camera with a blitz can be used for monitoring total suspended solids.
2. Faulty conditions such as no-foam and blockages can easily be detected.
3. High fidelity data for development of control strategies can be collected.

Keywords

Wastewater; Dewatering; Camera; Instrumentation; Monitoring

Nomenclature

WWTP - Wastewater Treatment Plant

TSS – Total Suspended Solids

OBS – Optical Back Scattering

R^2 - Coefficient of Determination

EPS - Extra-Polymeric Substances

OBS – Optical Back Scattering

1. Introduction

Wastewater treatment plays a vital role in modern society by ensuring the availability of safe drinking water and maintaining a healthy aquatic environment [1]. A crucial step in this process is the dewatering of biosolids, where microorganisms accumulated during secondary treatment are separated from the liquid portion. Biosolids management accounts for upwards of half of the overall operational cost of wastewater treatment [2]. Centrifugal decanters are commonly employed for dewatering due to their ability to handle large volumes while maintaining a high separation efficiency [3]. To enhance the effectiveness of separation, flocculants like polymers are used, which encourage the formation of larger particle aggregates leading to improved dewatering [4]. However, the dosing of chemicals presents operational challenges at wastewater treatment plants (WWTPs) due to the dynamic nature of the processes involved and the varying characteristics of biosolids [5]. The complexity associated with dewatering leaves room for improvement; better optimised chemical dosage leads to more efficient solids recovery which may provide two benefits: a reduction in reject solids that are directed to subsequent bioreactors reduces the aeration requirements, and an increase in solids available for digestion which can be used to generate biogas for electricity production, and thereby yielding a net profit in terms of operational costs. The costs associated with improved operation

such as an increased electricity demand of the dewatering equipment or potential increased costs of chemicals are considered negligible compared to the benefits that can be obtained through improved chemical dosage. Most dewatering processes are evaluated in terms of the reject quality, cake solids, or the solids capture and a change in chemical dosage can therefore be evaluated with said metrics to determine whether the change was beneficial or not. To the authors knowledge there is no way to automatically carry out the gravimetric laboratory analysis required for determining solids content used for calculating the aforementioned quality criteria. However, a variety of techniques exists for approximating the solids content through methods such as turbidimetry or microscopy [6]. Particularly common in water and water related applications is the turbidity sensor, which is an optical back scattering (OBS) device [7]. While OBS devices such as the turbidity sensor are widely adopted, it has been shown that for heterogeneous mixtures, such as when air bubbles are present, that this can render the sensors unreliable as the air bubbles mimic the backscattering from the particles that the devices use to determine their output [8]. Instrumentation, control, and automation is the ubiquitous technology for process improvement [9], and this paper describes the road towards advanced instrumentation of an industrial dewatering process treating inactivated biosolids. Here we present the validation of an approach with a regular camera that based on colour values of reject water from a centrifugal decanter estimates the total suspended solids content, i.e., the quality measure for the dewatering operation, which can in turn be used for control and optimization purposes [10].

2. Methods and Materials

2.1 Degassing Setup & Image Processing

The reject outlet of a centrifugal decanter is fitted with a degassing setup as depicted in figure 1. The purpose of the degassing setup is to compress the liquid thereby reducing the air content from centrifugation and limiting the refraction of light to produce a less foamy reject where colour differences can be traced with a camera (SensoPart Object, CMOS colour camera).

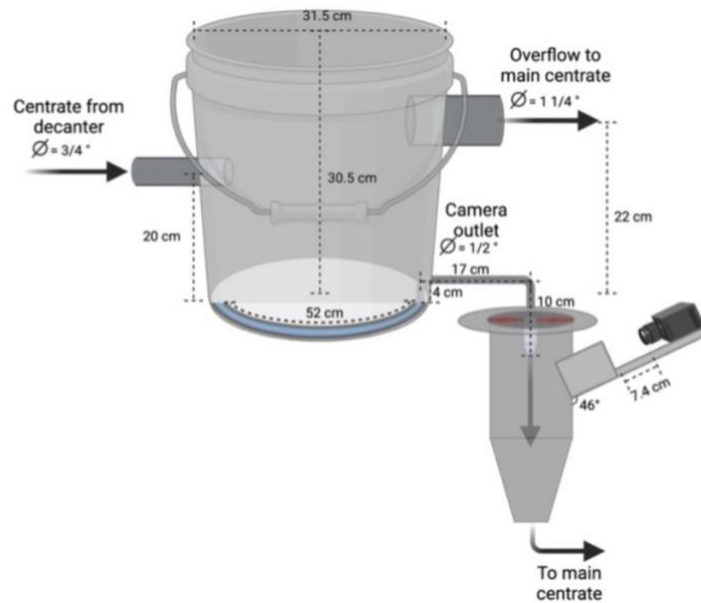


Figure 6 - Initial Degassing Setup. A plastic bucket is lined with a tube at the bottom that collects the degassed reject. The bucket also features an overflow which is useful for removing potentially accumulating foam.

A grab sample of the reject is taken from the outlet of the degassing chamber after the camera has captured an image as indicated by a blitz. Figure 2 displays an example of an image captured by the camera. The camera uses either a static window or a tracking window to localize the beam of liquid in each photo which may vary its position as the flow changes over time. Within the window simple color related computations can be calculated such as the contrast or the average colour values from several colour spaces (CIE 1931 RGB, CIELAB or HSV) and sent to the control system of the centrifugal decanter and then stored in an inhouse database or collected locally with a connected PC. In this study, the tracking window is placed by a vertical edge detection algorithm which looks for a vertical edge that exceeds a contrast threshold between the liquid beam and the background. Several tracking windows can be placed if desired.

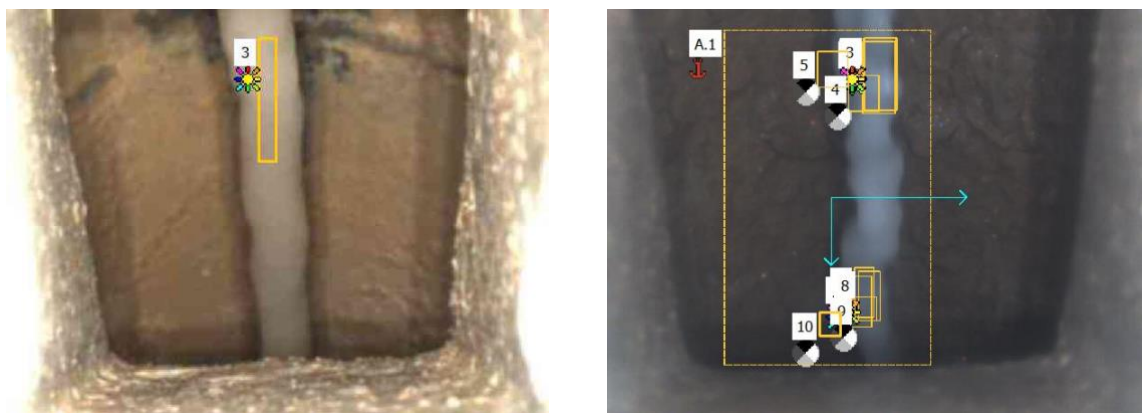


Figure 7 - Example photo from the camera setup with a static window (left), and two tracking windows for each of the three color spaces in the top and bottom (right).

2.2 Measuring campaign

After preliminary investigations (Appendix A) showed good intraday correlation between the observed camera values in the CIELAB colour scale and the collected total suspended solids content after the degassing setup, a larger measuring campaign (M1) was carried out in the Spring of 2022 (28th of April to the 15th of May). Afterwards a second measuring campaign was conducted in Spring 2023 (25th of April to the 5th of June 2023) to validate a stainless steel version of the plastic bucket prototype.

2.3 Determination of Total Suspended Solids

For each reject grab sample the total suspended solids (TSS) content is determined. A clean and dry glass microfiber filter (Whatman ® glass microfiber filters, Grade GF/A, 1.6 µm pore size) was used for each TSS measurement. The clean filter weight is recorded after which the filter is positioned on top of a vacuum filtration system and between 5-10 mL of liquid is deposited onto the filter before applying suction. After the liquid has been vacuumed off the remaining filter paper with sample is dried at 105° C in an MA35 infrared moisture analyzer (Sartorius). The TSS is then calculated as the dried sample minus the filter weight as a fraction of the sample weight.

3. Results

3.1 Spring Campaign 2022 (M1)

The results from the campaign are shown in figure 3 where an acceptable linear relationship ($R^2 = 0.814$) is present between the red-green dimension, A, in the CIELAB colour space and the collected TSS measurements. Based on this it was decided to upgrade the degassing vessel from plastic to stainless steel and repeat the validation experiment.

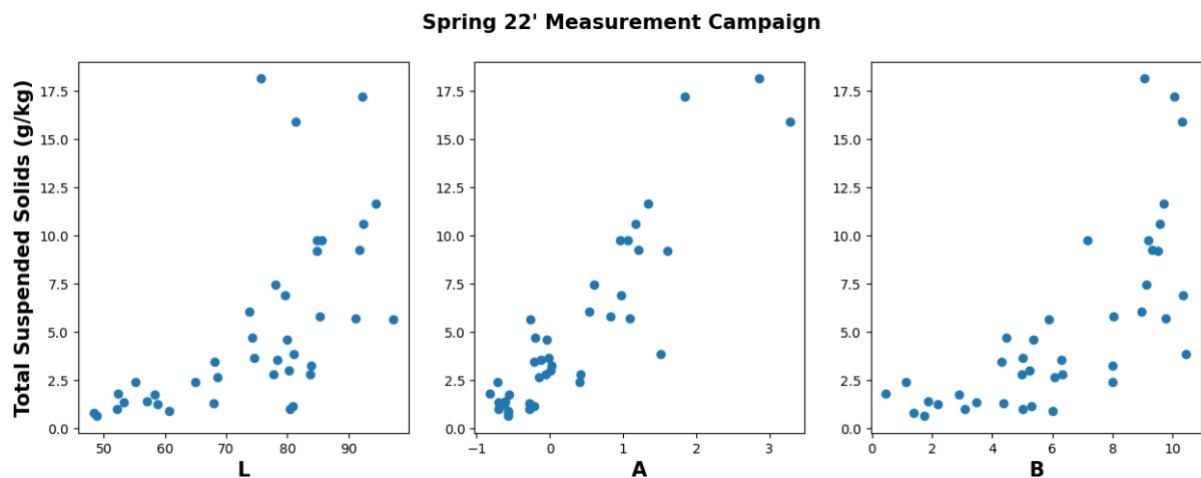


Figure 8 – Scatter plots for the 3 dimensions in the CIELAB colour scale vs TSS measurements. There appears to be an acceptable linear relationship between the red-green channel, A, and the TSS. R^2 between A and TSS is 0.814.

3.2 Spring Campaign 2023 (M2)

Figure 4 shows 3 select days from the campaign to demonstrate its performance. The data collected on the 8th and 17th of May both show positive correlation between the CIELAB A value from the camera (solid red line) and the reject grab samples TSS (blue dot), however the data collected on the 16th of May shows a negative correlation between the two.

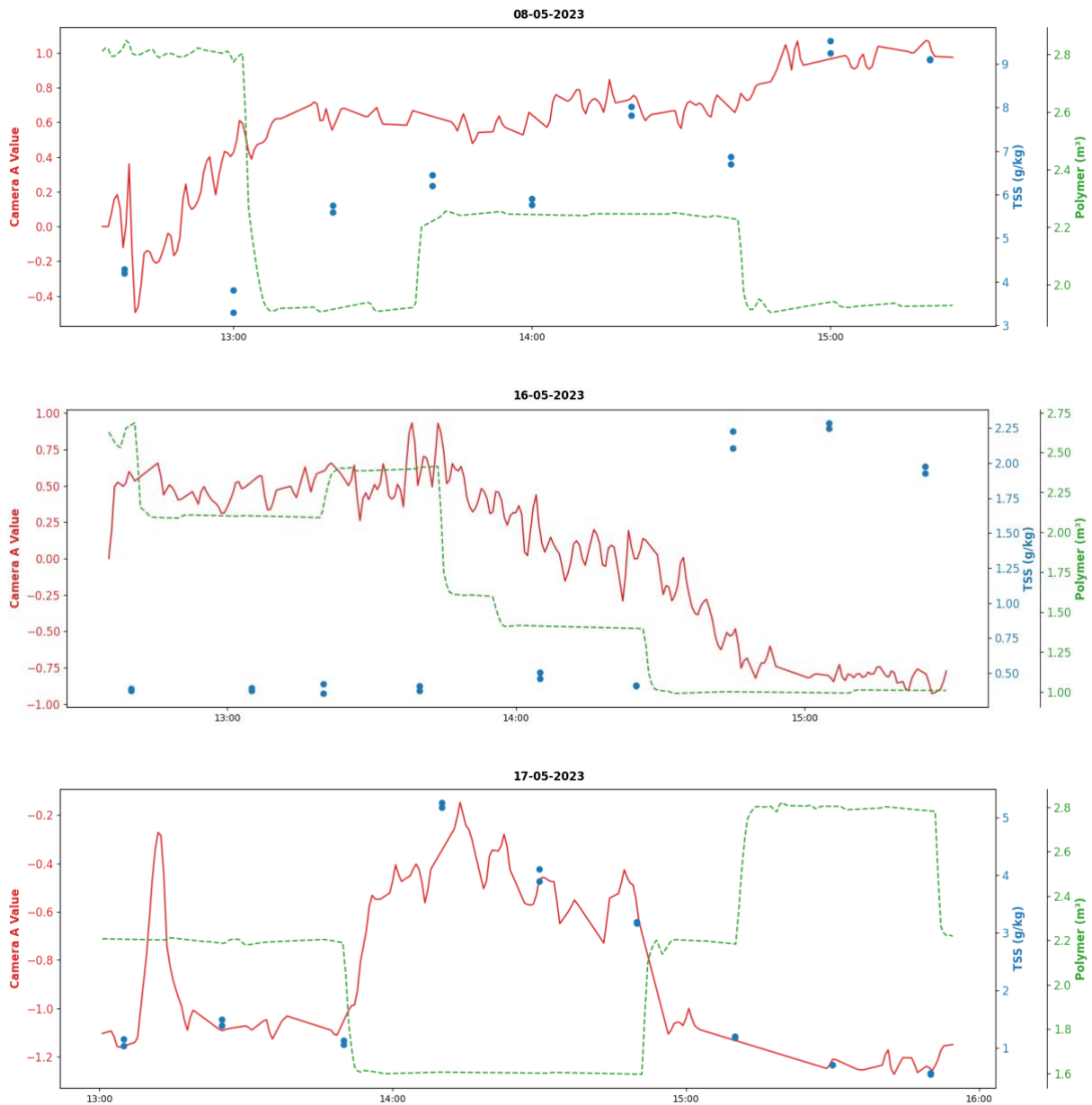


Figure 9 – Three select days of measurements with the camera and new degassing setup. The solid red line shows the CIELAB A value, the dashed green line shows the polymer dosage (m³) and the blue dots show TSS measurements from grab samples. The 8th and 17th display a positive correlation between the TSS measurements and the camera A value whereas the 16th displays a negative correlation. The data collected on the 16th was in general characterized by a considerably cleaner reject in terms of TSS than the 8th and 17th.

Figure 5 depicts a noisy camera signal which we suspect is due to the reject stream passing the camera exhibiting turbulent instead of laminar flow behaviour. This noisy behaviour was

associated with low ($R^2 < 0.55$) intraday correlations with TSS compared to other days where the intraday coefficient of determination was above 0.75 and approaching 0.98 as shown in table 1.

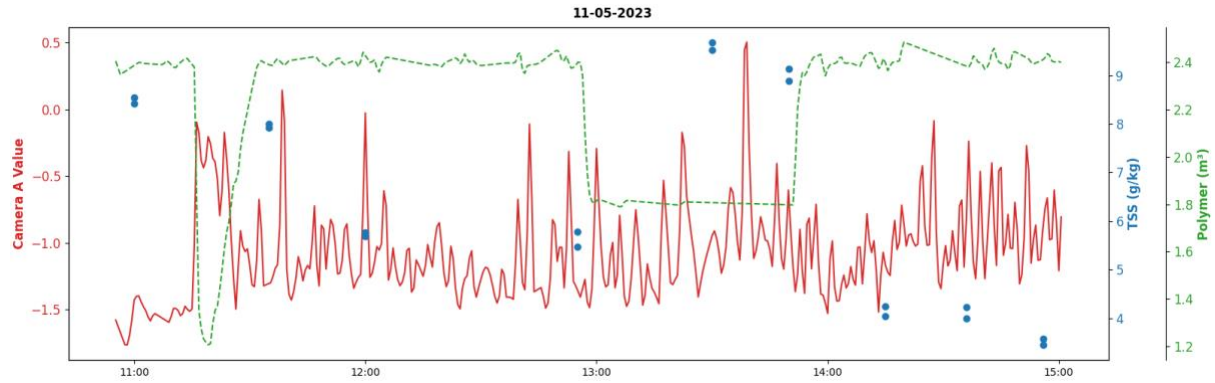


Figure 10 – 11th of May displayed a very noisy camera signal which we suspect is due to the reject stream passing the camera exhibiting turbulent instead of laminar flow behaviour.

Table 6 – Overview of coefficient of determination and model equations for intraday linear model fits.

Date	Coefficient of Determination (R^2)	Intraday Linear Model Fit
25/04/2023	0.765	$TSS = 5.72 * A + 6.1$
26/04/2023	0.831	$TSS = 6.45 * A + 5.06$
01/05/2023	0.895	$TSS = 8.88 * A + 11.74$
02/05/2023	0.534	$TSS = 2.07 * A + 1.76$
03/05/2023	0.769	$TSS = 3.35 * A + 5.33$
04/05/2023	0.782	$TSS = 5.86 * A + 9.37$
08/05/2023	0.940	$TSS = 9.13 * A + 0.44$
11/05/2023	0.499	$TSS = -3.57 * A + 2.57$
12/05/2023	0.825	$TSS = -6.54 * A + 2.63$
16/05/2023	0.829	$TSS = -1.33 * A + 1.05$
17/05/2023	0.979	$TSS = 4.71 * A + 6.42$
18/05/2023	0.965	$TSS = 6.89 * A + 5.89$
22/05/2023	0.785	$TSS = 6.03 * A + 4.03$

3.3 Detection of the no foam regime

Upon inspecting the images collected during the campaign it became evident that the images collected during the days with a negative correlation between the A value and the TSS measurements contained black reject, compared to the other days where the rejects observed demonstrated different shades of grey. Figure 6 shows 14 images collected during the campaign

sorted according to decreasing quality in terms of TSS, as well as their LAB colour values. The first two pictures appear to be textured compared to the next twelve, and these two samples correspond to a high reject quality with a TSS content below 0.4 g/kg. The samples appear black, however upon closer inspection they are transparent and merely reflect the background of the chamber, whereas the next 12 samples are foamy and do not reflect the background at all.

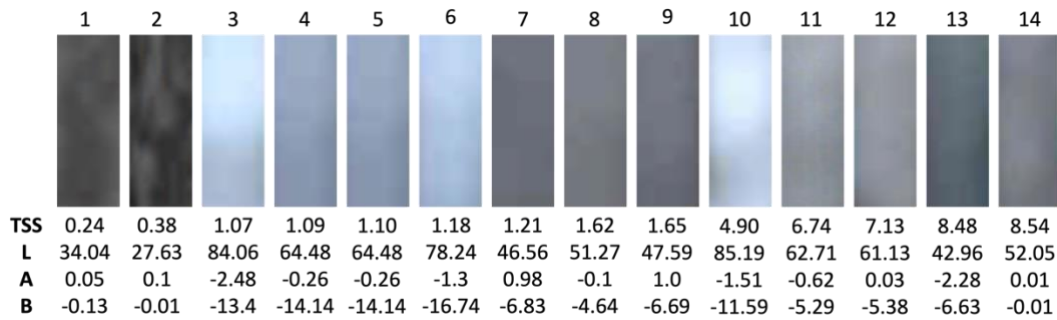


Figure 11 – 14 reject images from the campaign sorted by decreasing quality in terms of TSS (g/kg) as well as their LAB colour values.

Figure 7 shows photos from the old and new campaign to the left and right, respectively. The photos at the top display typical photos collected during each campaign, whereas the bottom photos display tap water and a clean reject left and right, respectively. The negative correlations between the CIELAB A value and the TSS appear when the TSS measurements are low, which corresponds to the transparent streams of liquid that reflect the background as seen by the tap water and low TSS reject reflecting the background colour in Figure 7.

During operation it is important to identify when the sensor provides reliable information, and in this case, it appears that for very clear rejects it is unreliable as the relationship switches from a positive correlation to a negative one. Based on the images and associated colour values from Figure 6, Figure 8 is created to try and identify this no-foam regime based on the other colour values in the CIELAB scale. The first and third subplot in figure 8 show that the outliers marked with an orange cross can be identified with a linear decision boundary. For instance, for the L and B values any value below 40 and above -2.5 respectively could identify the dark reject where there is no foam.

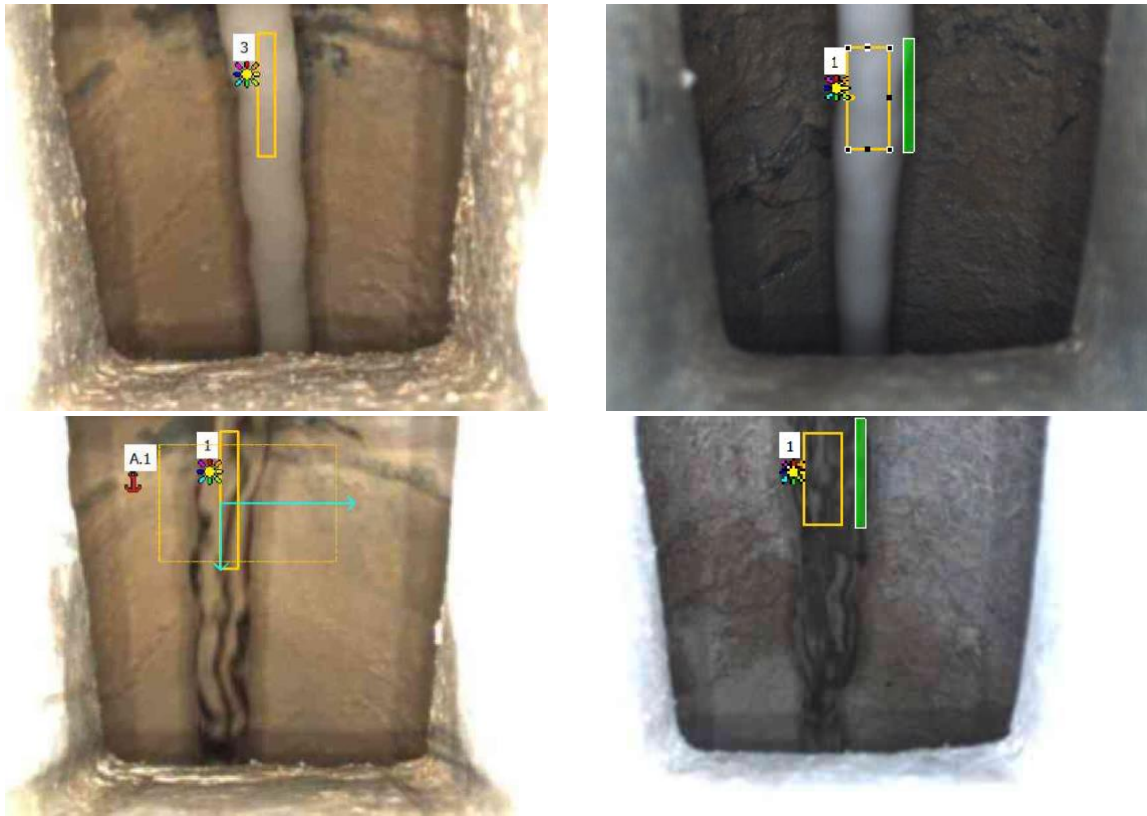


Figure 12 – Left: Old camera setup for validation of degassing procedure. Right: New camera setup for validating prior degassing procedure with a new degassing vessel. Top: Typical reject image with TSS ranging between 1-10 g/kg. Bottom: Water and low TSS reject (<0.4 g/kg) respectively.

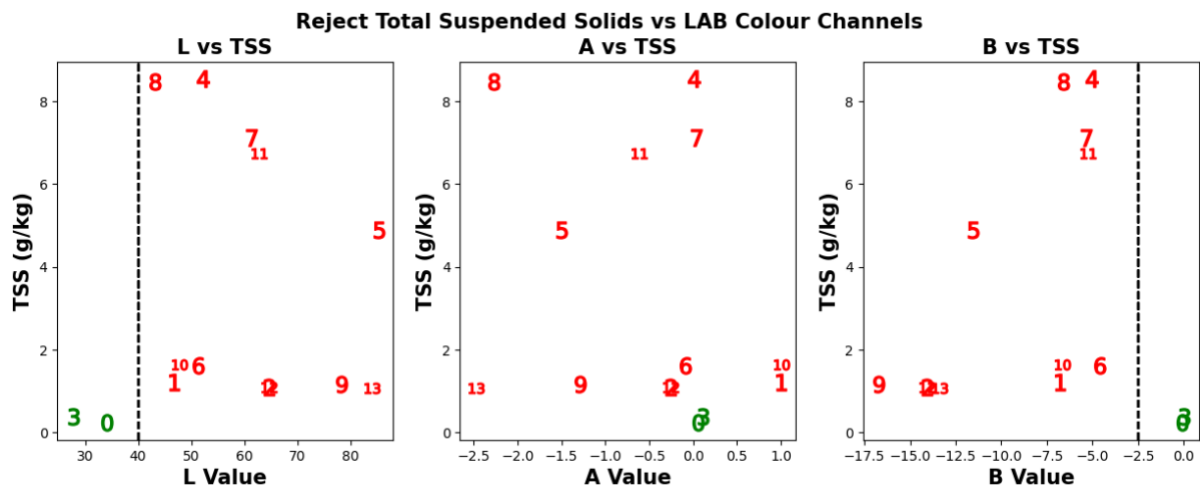


Figure 13 – LAB colour values vs TSS for a single day. The two points in the no foam regime are shown in green, with two examples of linear decision boundaries given as dashed black lines vertical lines at 40 and -2.5 for the L and B color scale respectively., i.e. the no foam regime appears to be detectable with the L and B channel, but not with the A channel.

3.4 Degassing Vessel Blockage

As the system runs continuously the degassing tank will gradually foul due to the accumulation of solids which eventually leads to blockage. Figure 9 shows approximately 3 days of operation where the vessel blocked in the last half of the period. The camera blockage can be easily identified as repeat camera measurements correspond to a blocked degassing vessel, i.e., $A_t = A_{t-1}$.

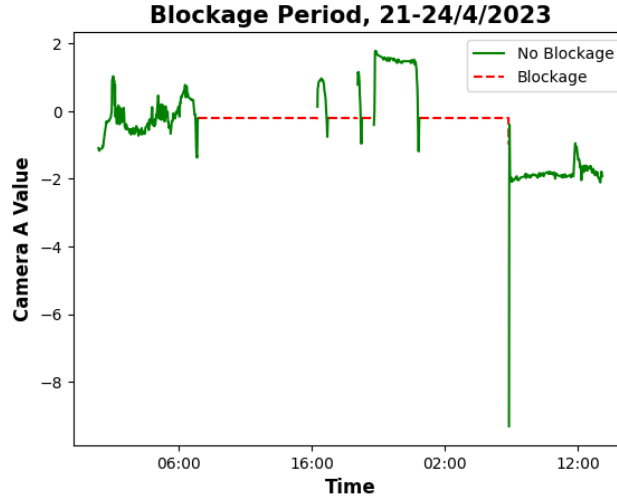


Figure 14 – Detection of camera blockage. The blockage in this period comes at the latter half where the signal is flat as marked in red. The stoppage can be detected by checking for repeat values where $A_t = A_{t-1}$

4. Discussion

Anecdotal, the days with poor intraday correlation coefficients (<0.7) are the ones where the static window used in M1 and M2 does not catch the liquid beam passing the camera adequately, which could be due to turbulence caused by the accumulation of solids in the pipe in the degassing vessel from which the liquid exits, or the absence of foam.

The no-foam regime which appears to occur when the TSS content is low presents a challenge in developing a reliable sensor system for monitoring TSS online. In the literature a few accounts are given of how the foaming potential of different sources of wastewater biosolids increases as the TSS content increases [11-12]. Jiang et al. demonstrated a negative relationship between the TSS content and the formation of foam for anaerobic digesters, however noted that this was contrary to previous work, speculating that the high TSS content of the investigated digestate (>30 g/kg) was the source of the differing dynamics [13]. Fryer et al. showed that for both a municipal and industrial WWTP the relationship between TSS and the foaming potential could be approximated with a positive second order polynomial, however the second order coefficient for the industrial WWTP was approximately 5 times larger than for the municipal one [14]. We speculate that this change in relationship between TSS and foam potential could be due to the presence of surfactants such as extra-polymeric substances (EPS) or additives that

promote the formation of foam, such as flocculants and coagulants. We postulate that the principal mechanism for inferring the TSS content of the reject via the camera is that the light from the blitz of the camera that is reflected by the foam is proportional to the TSS content as the suspended particles promote the formation of foam. This presents a unique opportunity for monitoring in situations where OBS devices are ill suited due to the presense of air, and in contrast with OBS devices which are typically in-line the camera is not submerged thereby reducing fouling and subsequent maintenance on the sensor. However, in this scenario the maintenance of the degassing vessel has to be accounted for in the overall maintenance comparison, and while the tank has to be cleaned approximately every 2-3 days which takes approximately 10-15 minutes the maintenance can be carried out by operators and does not require recalibration by technical staff.

The proposed solutions to identify the no-foam regime as well as the blockages should enable the low-cost camera solution for online monitoring of TSS, which in turn opens for developing control strategies, and fault detection systems that can notify operators if action is required, allowing them to direct their energy at more pressing tasks. By collecting high fidelity operational data both response curves and root cause analysis can be carried out to identify improved operational conditions as well as upstream conditions that result in challenging dewatering periods. Furthermore, the camera enables the collection of high fidelity data which can be used in conjunction with plantwide data to model and answer holistic questions such as how the upstream industrial production affects the downstream dewatering [15].

5. Conclusion

This paper showed the validation results of using a regular camera in combination with a degassing vessel to monitor the reject TSS content of a centrifugal decanter.

- Two validation campaigns were presented showing the ability of the camera to infer TSS with intraday correlation coefficients ranging from 0.53 to 0.98.
- Fault detection strategies were hypothesised to detect the no-foam regime corresponding to clean rejects, as well as blockage of the degassing vessel.
- The proposed setup can be used for collecting high fidelity full-scale operational data and developing control strategies.

6. Acknowledgements

Mr. Topalian would like to acknowledge Hiller GmbH, Novozymes and the Technical University of Denmark for their support during the project.

7. Data & Software Availability

The data and software developed in the frame of this project is confidential and will not be distributed.

8. References

- [1] Progress on wastewater treatment – Global status and acceleration needs for SDG indicator 6.3.1. United Nations Human Settlements Programme (UN-Habitat) and World Health Organization (WHO), Geneva, 2021.
- [2] O. Nowak, Optimizing the Use of Sludge Treatment Facilities at Municipal WWTPs, *Journal of Environmental Science and Health Part A*. 41 (2006) 9 1807-1817.
- [3] R. J. Wakeman, Separation technologies for sludge dewatering, *Journal of Hazardous Materials*. 144 (2007) 3 614-619.
- [4] Z. Yan, B. Örmeci, J. Zhang, Effect of Sludge Conditioning Temperature on the Thickening and Dewatering Performance of Polymers, *Journal of Residuals Science & Technology*. 13 (2016) 3 215-224.
- [5] A. Records and, K. Sutherland, *Decanter Centrifuge Handbook*, Elsevier: Amsterdam, The Netherlands, 2001.
- [6] D. S. Hansen, M. V. Bram, S. M. Ø. Lauridsen, Z. Yang, Online Quality Measurements of Total Suspended Solids for Offshore Reinjection: A Review Study, *Energies*. 14 (2021) 4 967.
- [7] A. F. B. Omar, M. Z. B. MatJafri, Turbidimeter Design and Analysis: A Review on Optical Fiber Sensors for the Measurement of Water Turbidity, *Sensors*. 9 (2009) 8311-8335.
- [8] I. Cáceres, J. M. Alsina, J. van der Zanden, D. A. van der A, J. S. Ribberink, A. Sánchez-Arcilla, The effect of air bubbles on optical backscatter sensor measurements under plunging breaking waves, *Coastal Engineering*. 159 (2020) 103721.
- [9] G. Olsson, ICA and me – a subjective review, *Water Research*. 46 (2012) 6 1585-1624.
- [10] C. Schneider, 2016, DE102015105988B3. <https://depatisnet.dpma.de/DepatisNet/depatisnet?window=1&space=menu&content=treffer&action=pdf&docid=DE102015105988B3&xxxfull=1>
- [11] L. L. Blackall, A. E. Habers, P. F. Greenfield, A. C. Hayward, Activated sludge foams: Effects of environmental variables on organism growth and foam formation, *Environmental Technology*. 12 (1991) 3 241-248. <https://doi.org/10.1080/09593339109385001>
- [12] D. B. Oerther, F. L. De Los Reyes III, M. F. De Los Reyes, L. Raskin, Quantifying Filamentous Microorganisms in Activated Sludge before, during, and after an Incident of Foaming by Oligonucleotide Probe Hybridizations and Antibody Staining, *Water Research*. 35 (2001) 14 3325-3336.
- [13] C. Jiang, R. Qi, L. Hao, S. J. McIlroy, P. H. Nielsen, Monitoring foaming potential in anaerobic digesters, *Waste Management*. 75 (2018) 280-288.
- [14] M. Fryer, E. O’Flaherty, N. F. Gray, Evaluating the Measurement of Activated Sludge Foam Potential, *Water*. 3 (2011), 424-444.
- [15] S. O. N. Topalian, P. Ramin, K. Kjellberg, C. K. Mbamba, D. J. Batstone, K. V. Gernaey, X. Flores-Alsina, A data analytics pipeline to optimize polymer dose strategy in a semi-continuous multi-feed dewatering system, *Journal of Water Process Engineering*. 55 (2023) 104048.

Appendix A – Intraday Experiments

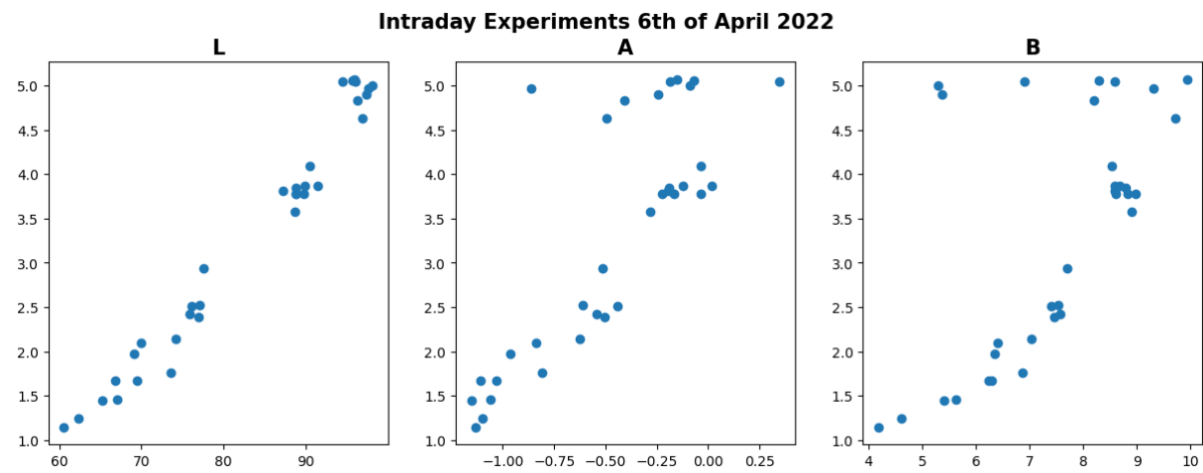


Figure 15 – Intraday CIELAB colour values vs TSS for the 6th of April 2022.

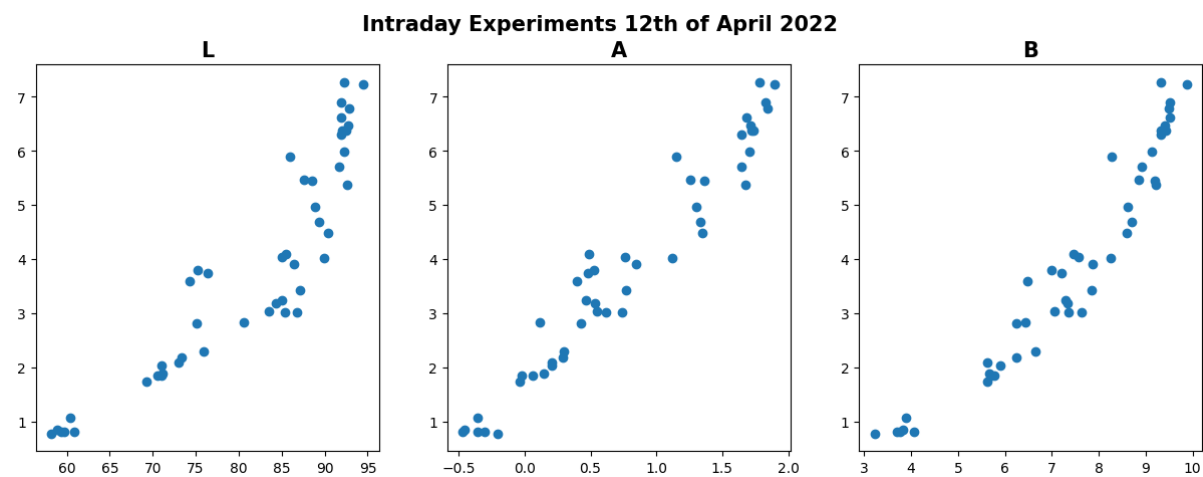


Figure 16 - Intraday CIELAB colour values vs TSS for the 12th of April 2022.

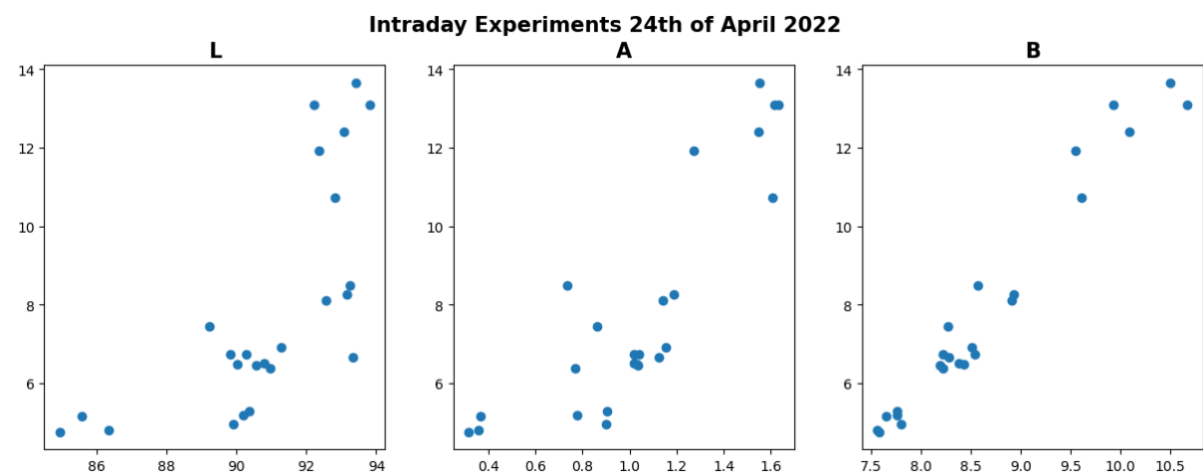


Figure 17 - Intraday CIELAB colour values vs TSS for the 24th of April 2022.

5. General discussion

In the first paper the approach of the data scientist was used to try and predict the polymer dosage for the industrial dewatering process. This was done due to the lack of an adequate turbidity signal, and the work assumed that the operators optimized the dosage of every batch, and that the polymer dosage would therefore be equivalent to the best possible polymer dosage in terms of separation efficiency. This assumption while naive at best still proved to link to some causal relationships with regards to the flow through the system that made sense from an engineer's intuition, and the impact of different production organisms and product codes could be analyzed despite the large delay between the production and the dewatering hall. Looking back, one thing that could have perhaps improved the model would be to know which operators had which shift as there is a suspicion that not all operators are created equally, and advanced analytics could help pinpoint which operators perform above average so that emphasis can be put on them as good examples.

For the second paper, the approach of the scientist, it was demonstrated that the operators frequently underdosed polymer in terms of achieving the best removal efficiency, which further supports the thought of including the operator schedule as a variable in analysis to see if this problem pertains to certain operators. In the second paper it was also evident that while techniques from computer vision can better distinguish between particle morphologies than traditional geometric properties, computer vision techniques alone were not sufficient to adequately predict the recovery efficiency in the machines.

One aspect that is not considered in this thesis is that the machines have a largely unobservable state in terms of the internal content of the machine, i.e., when the machine is separating successfully it is also assumed that the cake inside is a good consistency. This state is not reflected in the reject quality, as a phenomenon occurs where the reject quality can be high while no dewatering is performed, i.e., a wet cake is being produced. One way that this could be tackled is through the online monitoring of reject and cake, where the third paper addresses the former.

With online reject monitoring it would be possible to detect if one half of the outputs is of sufficient quality, however for determining the quality of the cake another sensor should be developed. With both in place it should be possible to gather high quality data of both reject and cake which would in turn motivate redoing the analysis from the first and second paper and predicting the removal efficiency or the state of the machine in terms of a metric combining both reject and cake quality. This would open up for repeating the analysis from the first paper to obtain more clear-cut results on which upstream conditions impact the outputs of the decanter, however it could also improve the accuracy of the framework developed in the second paper which could form the basis of a decision-support tool for simulating the effect of process settings such as polymer dosage on the recovery to provide operators with an educated guess and save time, so that their focus can be turned elsewhere. The downside of the work of the second paper is the fact that it is not clearly automatable when compared to the procedure developed in the first paper, however the framework from the second paper could allow diagnosis of the root cause in terms of which particles are problematic and provide a tool with greater explainability that could extrapolate to new products, or if conditions arise which are

not within the training data of the “big data” approach from the first paper. While each of the three approaches have relevant future work it is the belief of the author that the emphasis should be on improving the reject monitoring system as well as developing an online cake monitoring as well as potentially installing flow meters to close the mass balance around the unit. With a high-quality dataset an attempt could be made on creating a hybrid model based on the mechanistic models developed for centrifugal decanters treating minerals where the largest uncertainty for the sludge dewatering appears to be estimation of the input parameters such as particle size and weight.

5.1 Application in other domains and new horizons

While the work in this thesis focused on dewatering with centrifugal decanters, we believe that the tools here can be exported to other wastewater treatment processes that deal with physical-liquid separation such as other types of dewatering machines or clarifiers. One potential avenue for clarifier or machines where in-situ process measurements and samples are available would-be population balance models to determine the development and distribution of flocs over time, and potentially the development and distribution of bacterium types over type to better identify troublesome phenomena such as sludge bulking. Looking away from the field of water treatment a large overlap exists in terms of the separation technology used within the biotechnological industry, and we speculate that the same framework can be applied for similar product separation processes where measurements are available. The previously described avenue of population balance modelling could also be extended not only to the product separation of biotechnological products, but also to the upstream fermentations to monitor the state of the fermentations and how different cultures evolve if the subspecies are visually distinguishable.

One horizon that did not make it into the thesis was the use of Fourier transformed infrared (FTIR) spectroscopy to monitor extra polymeric substances (EPS) which could supplement the existing work with key information on surfactants that are believed to play a key role in flocculation mechanisms, and consequently impact dewatering. The use of spectroscopy has the potential to contribute with information regarding chemical composition that is not readily available from either the collected online data or the microscopy images, and the addition thereof could potentially increase the accuracy of future models.

6. Conclusion

In conclusion three papers were developed highlighting:

- Demonstrating the advantages of the emerging data scientist approach with linking upstream production and downstream production conditions.
- The interpretability that can be obtained for quantitative image analysis by coupling it with computer vision.
- The ability to monitor reject turbidity in the presence of foam where OBS devices typically fault.

Finally, it is recommended to continue the avenue of online monitoring of the reject as well as monitoring of the cake to close the mass balance around each unit and potentially developing a serial-hybrid model based on the centrifugal decanter models from the mineral dewatering literature.

7. References

- [1] G. Olsson, ICA and me – a subjective review. *Water Research*, 46 (2012) 6 1585-624. doi: 10.1016/j.watres.2011.12.054
- [2] P. Ginisty, R. Mailler, and V. Rocher, Sludge conditioning, thickening and dewatering optimization in a screw centrifuge decanter: Which means for which result?, *Journal of Environmental Management*. 280 (2021) 111745.
- [3] https://en.wikipedia.org/wiki/Decanter_centrifuge, visited 1/10/2023.

The end.