Aalborg Universitet



Machine Learning for Value Creation in the Water Sector

Hansen, Bolette Dybkjær

DOI (link to publication from Publisher):
10.54337/aau478418331

Publication date:
2022

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

Link to publication from Aalborg University

Citation for published version (APA):
Hansen, B. D. (2022). *Machine Learning for Value Creation in the Water Sector*. Aalborg Universitetsforlag. https://doi.org/10.54337/aau478418331

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 -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: December 06, 2025

MACHINE LEARNING FOR VALUE CREATION IN THE WATER SECTOR

BY
BOLETTE DYBKJÆR HANSEN

DISSERTATION SUBMITTED 2022



AALBORG UNIVERSITY
DENMARK

Machine Learning for Value Creation in the Water Sector

Ph.D. Dissertation

Bolette Dybkjær Hansen

Aalborg University
Department of Architecture, Design and Media Technology
Rendsburggade 14
9000 Aalborg

Dissertation submitted: March 2022

PhD supervisor: Professor Thomas B. Moeslund
Aalborg University

PhD committee: Associate Professor Jesper Rindom Jensen (chairman)
Aalborg University, Denmark

Professor Zoran Kapelan
Delft University of Technology, The Netherlands

Professor Anders Kofod-Petersen
Norwegian University of Science and Technology,
(NTNU) Norway

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Architecture, Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-927-1

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

© Copyright: Bolette Dybkjær Hansen

Printed in Denmark by Rosendahls, 2022

Curriculum Vitae

Bolette Dybkjær Hansen



Bolette D. Hansen received her masters degree in *Biomedical Engineering and Informatics* in 2017 from Aalborg University. Before starting her Ph.D. studies in 2019 she worked as a research assistant at Visual Analysis and Perception Lab (VAP-Lab) at Aalborg University.

Her main focus has been research in applied data science within a broad range on subject areas including water technology, medico technology, sports analysis, data science in business perspectives, machine learning and computer vision.

Abstract

Machine learning has contributed with significant value in several industries. However, despite the water sector being a data heavy intensive sector, it is still far behind in the implementation of machine learning. Several barriers for machine learning in the water sector exist, and often the decision makers in the water sector lack knowledge about machine learning, making it difficult to make optimal decisions regarding the investment.

The aim of this thesis is to investigate the potential for value creation in the water sector using machine learning and perform research in relevant use cases.

Therefore, several use cases were identified through meetings with experienced water professionals. The use cases were subsequently assessed according to their economic potential, the required investment and the risk related to development. Furthermore, the use cases were clustered according to area of the water sector, whether they represented value that could not be directly calculated as an economic benefit, and type of machine learning. Based on the analysis, four use cases were subject for research in this work. The four use cases were *sewer deterioration modeling*, *prediction of methane yield from biogas plants*, *fault detection in pumps*, and *Drift detection in Wastewater Treatment Plants (WWTP)*.

Sewer deterioration modeling entailed development of a Random Forest deterioration model, investigation of the potentials for optimizing the model by using logically grouped datasets, investigation of the features affecting the model, investigation of how the data affects the performance of the models and how the data affects the potential for forecasting pipe condition. The model obtained state-of-the-art performance, however, it was not possible to optimize it by utilization of logically grouped datasets. During the investigation of the features affecting the performance of it was observed that the feature importance varied between different utilities, and that the models were highly dependent on the pipe inspection strategy. The inspection strategy also affected the forecasts of the pipe conditions.

For predicting the methane yield from biogas plants, it was investigated if a hybrid model, consisting of a Gompertz model and a machine learning model, could obtain better performance when compared to one of the models. The results showed that for

predictions one day ahead, the hybrid model indeed performed better than each of the models individually.

For fault detection in pumping stations, it was investigated if Convolutional Neural Network (CNN) could improve the reconstructions of the energy in pumping stations compared to Multilayer Perception (MLP). However, despite both the CNN and MLP models performing well with their predictions being used for filling missing signals, the performance was not sufficient for fault detection. The primary reason for this was related to low resolution of the data.

For the last subject, drift detection in WWTPs, the application of methods developed in the literature to data from real operational plants was sought. The results showed that it was difficult to apply these methods to data from real WWTPs, as most of them were developed for highly controlled data such as simulated data and dry-weather data. Based on the findings from those methods, recommendations were made for bridging the gap between academia and practice.

From the experience obtained through research within the four subjects, the factors affecting the investment and risk related to development of machine learning solutions were discussed, and recommendations for minimizing the investment and the risk related to development of machine learning solutions were formulated. It is expected that these recommendations will contribute to future decision making.

Low data quality is a key barrier for machine learning in the water sector. However, despite the data being of low quality for machine learning purposes, the data quality might be sufficient for solutions based on other types of AI, statistics and visualisations. Despite this barrier, several drivers for machine learning in the water sector exist. For instance, it is expected that the data quality will increase due to increased focus. Furthermore, the need for reaching the United Nations Sustainable Development Goals (SDGs), increased awareness of the environment among citizen are among the drivers for more machine learning in the water sector.

Resumé

Machine learning har bidraget til stor værdiskabelse i mange forskellige sektorer, men på trods af at vandsektoren er en datatung industri, er anvendelsen af machine learning langt bagud sammenlignet med eksempelvis produktionssektoren. En udfordring i forbindelse med inklusion af machine learning i vandsektoren er, at beslutningstagerne ofte mangler erfaring med machine learning, hvilket gør det sværere at træffe en velbegrundet beslutning.

Formålet med dette arbejde er at undersøge potentialet for værdiskabelse i vandsektoren ved brug af machine learning samt at udføre forskning indenfor relevante use cases.

Derfor blev der først identificeret en lang række potentielle use cases gennem møder med fageksperter indenfor vandsektoren. Use casene blev herefter systematisk vurderet udfra deres økonomiske potentialer, den forventede investering og risikoen forbundet til udvikling af produktet. Derudover blev use casene inddelt efter, hvilket område indenfor vandsektoren de tilhørte, om de repræsenterede en værdi, som ikke kunne beregnes som et økonomisk potentiale samt typen af machine learning der skulle anvendes til use casen. På baggrund af analysen blev der forsket indenfor følgende fire områder: *Tilstandsmodellering af kloakledninger*, *forudsigelse af metanudbyttet fra biogasanlæg*, *detektion af fejl i pumpestationer* samt *detektion af sensordrift i rensningsanlæg*.

Arbejdet med tilstandsmodellering af kloakledninger indebar udvikling af en random forest tilstandsmodel, forskning i optimering af modellen ved anvendelse af logisk opdelte datasæt, forskning i hvilke parametre der er betydende for modellens performance, og hvordan datagrundlaget påvirker modellernes ydeevne samt undersøgelse af, hvordan historisk data bedst muligt anvendes til at fremskrive ledningstilstanden. Den udviklede model opnåede state-of-the-art performance, men det lykkedes ikke at optimere modellen yderligere ved brug af logisk opdelte datasæt. I forbindelse med undersøgelserne af hvilke parametre der er betydende for modellernes performance, viste det sig at der var en stor variation i vigtigheden af parameterne for de forskellige forsyninger, og at modellerne var stærkt afhængige af den anvendte inspektions strategi. Den anvendte inspektionsstrategien påvirkede og fremskrivningen af ledningstilstanden.

Til forudsigelse af metanudbyttet fra biogasanlæg blev det undersøgt om en hybrid-

model, bestående af en gompertz-model og en machine learning-model, kunne levere bedre resultater sammenlignet de to modeltyper hver for sig. Resultaterne viste, at hybridmodellen var bedre til at forudsige metan udbyttet en dag ud i fremtiden, end de to modeller var hver for sig.

Til fejldetektion i pumpestationer blev det undersøgt, om Convolutional Neural Network (CNN) kunne levere bedre rekonstruktioner af energiforbruget i pumpestationer sammenlignet med Multilayer Perception (MLP). På trods af at både CNN og MLP-modellerne opnåede god performance, hvis forudsigelserne skulle bruges til at udfylde manglende data, var rekonstruktionerne ikke tilstrækkelige til fejldetektion. Den primære årsag til dette var, at opløsning af dataene var for lav.

I forbindelse med detektion af sensordrift i sensorer på rensningsanlæg, blev det afdækket, hvorvidt tidligere publicerede metoder kunne anvendes på data fra operationelle rensningsanlæg. Resultaterne viste, at det var svært at anvende metoderne på data fra operationelle rensningsanlæg, da de fleste af dem var baseret på data med langt højere datakvalitet end det, der indsamles på de fleste operationelle rensningsanlæg. Eksempelvis er mange af metoderne udviklet til simuleret data, hvor fejltyper kan isoleres, og hvor spildevandssammensætningen m.v. i forvejen er kendt. På baggrund af resultaterne blev der formuleret anbefalinger til, hvordan forskning og praksis kan nærme sig hinanden.

På baggrund af erfaringerne fra forskningen i de fire emner, blev der lavet en ny vurdering af, hvilke faktorer der er betydende for størrelsen af investeringen og risikoen relateret til udvikling af machine learning-løsninger. Derudover blev de forskellige faktorer diskuteret, og der blev formuleret anbefalinger til, hvordan disse kan minimeres. Det forventes, at disse anbefalinger vil bidrage til fremtidig beslutningstagning i krydsfeltet mellem vandsektoren og machine learning.

Lav datakvalitet er en væsentlig barriere for machine learning i vandsektoren, men selvom data ofte er af for ringe kvalitet i machine learning-sammenhæng, kan datakvaliteten godt være tilstrækkelig høj til anvendelse i løsninger, der er baseret på statistik, mindre komplicerede AI teknikker og visualisering. På trods af denne barriere er der adskillige drivere for machine learning i vandsektoren. Eksempelvis forventes det, at datakvaliteten vil stige som følge af øget fokus. Ydermere er behovet for at nå FN's verdensmål for bæredygtig udvikling og øget opmærksomhed på miljø blandt borgere nogen af de faktorer, der trækker i retning af mere machine learning i vandsektoren.

Contents

Curriculum Vitae iii

Abstract v

Resumé vii

Thesis Details xiii

Preface xv

I Overview of work 1

1 Introduction 3
1.1 Scope of thesis . . . 5
1.2 Thesis structure . . . 5
References . . . 6

2 Potential of Machine Learning in the Water Sector 11
2.1 Potential use cases . . . 12
2.2 Assessment of use cases . . . 12
2.3 Selection of use cases . . . 16
2.4 Contributions . . . 20
References . . . 20

3 Sewer Deterioration Modeling 23
3.1 Sewer deterioration modeling . . . 24
3.2 Optimization of sewer deterioration models . . . 28
3.3 Features affecting the performance . . . 31
3.4 Forecasting of pipe conditions . . . 39
References . . . 43

4 Forecasting of Methane Yield from Biogas Plants **47**
4.1 Existing approaches . . . 48
4.2 Model for forecasting the methane yield . . . 49
4.3 Contributions . . . 52
References . . . 52

5 Fault Detection in Pumps **55**
5.1 Existing approaches . . . 56
5.2 Fault detection using CNNs . . . 57
5.3 Results . . . 61
5.4 Contributions . . . 62
References . . . 63

6 Drift detection in sensors at Wastewater Treatment Plants **65**
6.1 Existing approaches . . . 67
6.2 Drift detection in operating WWTPs . . . 69
6.3 Contributions . . . 77
References . . . 77

7 Value Creation **81**
7.1 Analysis of the method for assessment of use cases . . . 81
7.2 Machine learning in the water sector . . . 85
7.3 Drivers for digitalization, AI and machine learning in the water sector . . . 87
7.4 Contributions . . . 88
References . . . 88

8 Conclusion **91**

II Papers **93**

A General Sewer Deterioration Model Using Random Forest **95**
A.1 Introduction . . . 97
A.2 Method . . . 99
A.3 Results . . . 105
A.4 Discussion . . . 106
A.5 Conclusion . . . 112
References . . . 112

B Sewer Deterioration Modeling: The Effect of Training a Random Forest Model on Logically Selected Data-groups **115**
B.1 Introduction . . . 117
B.2 Method . . . 119
B.3 Results . . . 122
B.4 Discussion . . . 124
B.5 Conclusion . . . 126
References . . . 126

C Comprehensive Feature Analysis for Sewer Deterioration Modeling **129**
C.1 Introduction . . . 131
C.2 Materials and Methods . . . 134
C.3 Results . . . 141
C.4 Discussion . . . 147
C.5 Conclusion . . . 151
References . . . 152

D Prediction of the Methane Production in Biogas Plants Using a Combined Gompertz and Machine Learning Model **155**
D.1 Introduction . . . 157
D.2 Method . . . 159
D.3 Results . . . 162
D.4 Discussion . . . 164
D.5 Conclusion . . . 167
References . . . 167

E Data-Driven Drift Detection in Real Process Tanks: Bridging the Gap between Academia and Practice **169**
E.1 Introduction . . . 171
E.2 Materials and Methods . . . 174
E.3 Results . . . 177
E.4 Discussion . . . 187
E.5 Perspectives and Recommendations . . . 189
E.6 Conclusions . . . 194
References . . . 195

III Appendices **197**

F Lookup Tables for Assessment of Use Cases **199**
F.A Lookup tables . . . 201

G Supplementary to paper C, Correlation between topographically connected pipes **205**
G.A Correlation between topographically connected pipes 207

Thesis Details

Thesis Title:	Machine Learning for Value Creation in the Water Sector
Ph.D. Student:	Bolette Dybkjær Hansen
Supervisors:	Prof. Thomas Baltzer Moeslund, Aalborg University
	Ph.D. David Getreuer Jensen, EnviDan A/S

The main body of this thesis consists of the following papers:

[A] Bolette Dybkjær Hansen, David Getreuer Jensen, Søren Højmark Rasmussen, Jamshid Tamouk, Mads Uggerby and Thomas BaltzerMoeslund, "General Sewer Deterioration Model Using Random Forest", *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence(SSCI), Xiamen, China, 6–9 December 2019*, pp. 834–841, 2019.

[B] Bolette D. Hansen, Søren H. Rasmussen, Thomas B. Moeslund, Mads Uggerby and David G. Jensen, "Sewer Deterioration Modeling: The Effect of Training a Random Forest Model on Logically Selected Data-groups", *Procedia Comput. Sci.*, Vol. 176 pp. 291–299, 2020.

[C] Bolette D. Hansen, Søren H. Rasmussen, Mads Uggerby, Thomas B. Moeslund and David G. Jensen, "Comprehensive Feature Analysis for Sewer Deterioration Modeling", *Water*, Vol. 13, 819, 2021.

[D] Bolette D. Hansen, Jamshid Tamouk, Christian A. Tidmarsh, Rasmus Johansen, Thomas B. Moeslund, and David G. Jensen, "Prediction of the Methane Production in Biogas Plants Using a Combined Gompertz and Machine Learning Model", *Procedia Comput. Sci.*, Vol. 176 pp. 291–299, 2020.

[E] Bolette D. Hansen, Thomas B. Hansen, Thomas B. Moeslund and David G. Jensen, "Data-Driven Drift Detection in Real Process Tanks: Bridging the Gap between Academia and Practice", *Water*, Vol. 14, 926, 2022.

In addition to the publications listed above, the following publications have been co-authored by the Ph.D. student, but are not related to the thesis.

- Noor Ul Huda, Bolette Dybkjær Hansen, Rikke Gade, Thomas B. Moeslund, "The Effect of a Diverse Dataset for Transfer Learning in Thermal Person Detection", *Sensors*, Vol. 20, 7, 2020.
- Noor Ul Huda, Bolette Dybkjær Hansen, Rikke Gade, Thomas B. Moeslund, "Occupancy Analysis of Soccer Fields Using Wide-Angle Lens", *13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Vol. 43, pp. 302-313, 2017.
- Anne Krogh Nøhr, Louise Pedersen Pilgaard, Bolette Dybkjær Hansen, Rasmus Nedergaard, Heidi Haavik, Rene Lindstrøm, Maciej Plocharski, Lasse Riis Østergaard, "Semi-automatic method for intervertebral kinematics measurement in the cervical spine", *20th Scandinavian Conference, SCIA, 12-14 June 2017, Tromsø, Norway, Proceedings, Part II*, Vol. 43, pp. 302-313, 2017.
- Kim Munck, Bolette Dybkjær Hansen, Nina Jacobsen, Louise Pedersen Pilgaard, Samuel Schmidt, Kasper Sørensen, Johannes Jan Struijk, "Body Surface Mapping of the Mechanical Cardiac Activity", *Computing in Cardiology*, Vol. 43, pp. 661-664, 2016.
- Johannes J Struijk, Kim Munck, Bolette D Hansen, Nina Jacobsen, Louise P Pilgaard, Kasper Sørensen, Samuel E Schmidt, "Heart-valve Sounds Obtained with a Laser Doppler Vibrometer", *Computing in Cardiology*, Vol. 43, pp. 197-199, 2016.

Preface

This Industrial Ph.D. thesis is the result of a close collaboration between EnviDan A/S and Aalborg University.

EnviDan A/S is a consulting engineering company working within the water sector. Inspired by the value gained by introducing machine learning in other data heavy industries, EnviDan and Aalborg University joined forces for the purpose of investigating how machine learning can create value in the water sector. Consequently, an industrial Ph.D. study was defined. An industrial Ph.D. study is a Ph.D. study where the Ph.D. student is employed in a company while enrolled at a university. The Ph.D. fellowship was partly funded by Innovation Fund Denmark.

I would like to thank my two supervisors Thomas B. Moeslund and David G. Jensen for providing excellent supervison, showing great confidences in my work, and for providing a large degree of freedom during the project. I would also like to thank my colleges at EnviDan who have been very helpful and contributed to this work with expert knowledge within several different areas. Especially i would like to thank Søren H. Rasmussen for his assistance during the project. Another thanks goes to my colleges at Visual Analysis and Perception Lab at Aalborg University for being part of an including research environment and for helping keep me up the motivation during the corona lock-downs.

Finally, I would like to thank Mads Uggerby and EnviDan for allocating resources for the project as well as Innovation Fund Denmark for supporting the project financially.

Bolette Dybkjær Hansen
Aalborg University, March 25, 2022

Part I

Overview of work

Chapter 1

Introduction

The value gained from digital inventions is increasing exponentially [23], and digital innovation is expected to add 14 % to the global economy by 2030 compared to 2017 [1]. This has also been reflected in the European Union's budget, which has allocated billions of euros for the digital transformation through different programs [7]. One such program is Horizon Europe, worth €95,52 billion[1], which targets research in the green and digital transition, health and resilience. 35 % of the funding in Horizon Europe should go to support the digital transition. Another program, Digital Europe, is worth €7.59 billion[1] and has the purpose of bridging digital technologies to citizens, businesses and public administration. It is worth mentioning that these programs are coherent with other programs and funding, meaning that the real amounts to be spent on the areas are significantly larger [7].

The maturity of the digital transformation depends on the specific industry. While manufacturing is a leading industry in the digital transformation [18], the water sector is far behind despite being a very data heavy industry [36].

Recently, there has been an increased focus on digitization of the water sector to meet the United Nations Sustainable Developments Goals (SDGs) [23]. However, the water sector is a complex industry with many causalities in the variables influencing the decision processes [25]. Despite progress within, e.g. environmental decision support systems, there is still a long way before research can deliver systems which smoothly integrate several knowledge types and reasoning methods [5]. Several reasons for this have been recognized in the literature, such as economic, technical, regulatory, and human factors [36].

An example of an economic factor is the willingness to pay for the product; compared to other industries, the water sector has not been forced to optimize their solutions through competition [36] but through local legislation and taxes for discharges, as is the

[1]Prices are inflation adjusted for November 2020.

case in Denmark [22, 28]. Furthermore, the water prices are between 0.5 % and 1 % of the oil prices. Thereby, data-driven solutions within, for instance, leak detection in oil pipes are far ahead compared to solutions for leak detection in water pipes. However, recently water has been considered a more valuable resource, and more research has been put into optimization in the water sector [36].

The technical reasons include, for instance, the fact that the processes in the water sector are fairly stable, and thereby could be operated without automation. However, increasing complexity in, e.g., wastewater treatment and recovery systems, means that the human ability to handle and predict disturbances in the systems becomes a limiting factor. Another technical reason is that historically utilities have been required to meet specified limits for emissions, which can be achieved by rule-based solutions, ensuring no need for developing the most optimal solutions. At Wastewater Treatment Plants (WWTPs), missing reliability of more complicated sensors, such as ammonia, nitrate, etc., has prevented well-developed control algorithms from being applied. Furthermore, a sufficient level of instrumentation for control and automation is often not included in the design of the plants but added as supplementary afterwards [36]. Despite a movement towards including these perspectives in modern designs, the life cycle of a wastewater treatment plant is several decades. Thereby this problem will be present for several decades in the future.

A regulatory factor often present in Europe is that only laboratory analysis of manually collected samples are accepted by regulatory agencies.

Human factors include operators changing set points to increase the safety margins for the quality of treated water. Increased safety margins typically entails increased usage of energy and chemicals, but historically, the operators have seldom been rewarded for reducing the usage of energy and chemicals [36]. However, today the focus on climate is increasing, and goals for energy and climate neutrality have been defined. For example, in Denmark the water sector should be climate and energy neutral in 2030 [2]. Furthermore, other factors, such as job protection, lack of education in operational performance, and overdesign of the systems, are relevant. The reason for overdesign of systems being relevant is that it entails a lower sensitivity to optimal control to meet the requirements for emissions [36].

Despite these limiting factors, an exponential increase in applications of machine learning and artificial intelligence has been seen within environmental science and water management [30]. The applications are within a large number of different areas, including burst detection in clean water distribution systems [35], clean water supply [19], drinking water quality [4, 6, 34], inland water quality assessments [4, 8, 25, 27, 31, 33], sewer management [12–15, 20, 32], hydrology [39, 40][18,19], rainfall [3, 29], flood detection and management [16, 26], wastewater treatment management and control [5, 11, 17, 21, 24, 37, 38], and climate research [23].

A key challenge for machine learning in the water sector is that despite several learning based solutions have been developed in academia, only few have made it into

real world implementations [5, 10, 11].

For instance, operation of WWTPs is complicated by the fact that it needs to be both environmental, economic and socially sustainable. Decision Support Systems (DSS) encountering the sustainability aspects might be more complicated than traditional DSS, though that does not make them more reliable. The reusability of the DSS systems in WWTPs varies between studies, and the fact that many solutions are site-specific is one of the major challenges. Additionally, the datasets used are often kept private. [21].

There is still a huge potential for new machine learning applications [24], and despite machine learning being widely applied within the water sector, a large potential remains [9]. Within water, environmental, and resource management the need of Big Data analysis is inevitable. However, it is even more important to ensure correct implementation and proper usage and planning of the resources available [9]. In the future, Big Data will change environmental and water research design, performance and analysis fundamentally [30], and these potentials apply several levels of digitization, from automatic meter reading to high level digitization methods such as advanced data analytic.

1.1 Scope of thesis

The aim of this thesis is to investigate the potential for value creation in the water sector using machine learning and perform research within relevant areas. This is done by identification and assessment of relevant use cases and subsequently perform research and development within a number of these cases that a) has a potential for value creation, b) can contribute within different areas of the sector and c) can bring new methodological knowledge to the commercial part of the water sector. Based on the experience obtained through investigation of the different use cases, recommendations for future work within the sector are given.

1.2 Thesis structure

The body of this thesis consists of eight chapters: Chapter 2 is an investigation of the potential for machine learning in the water sector from a practical perspective. Based on the findings in Chapter 2, Chapters 3, 4, 5, and 6 present the research conducted within four main themes: sewer deterioration modeling, forecasting of methane yield from biogas plants, fault detection in pumping stations, and drift detection in WWTPs, respectively. Chapter 7 contains a follow-up on the analysis presented in Chapter 2 based on the experiences gained from the work within the four themes. Furthermore, Chapter 7 provides general perspectives on machine learning in the water sector. An overview of the main content of the thesis, related papers, and one appendix can be seen in Figure 1.1.

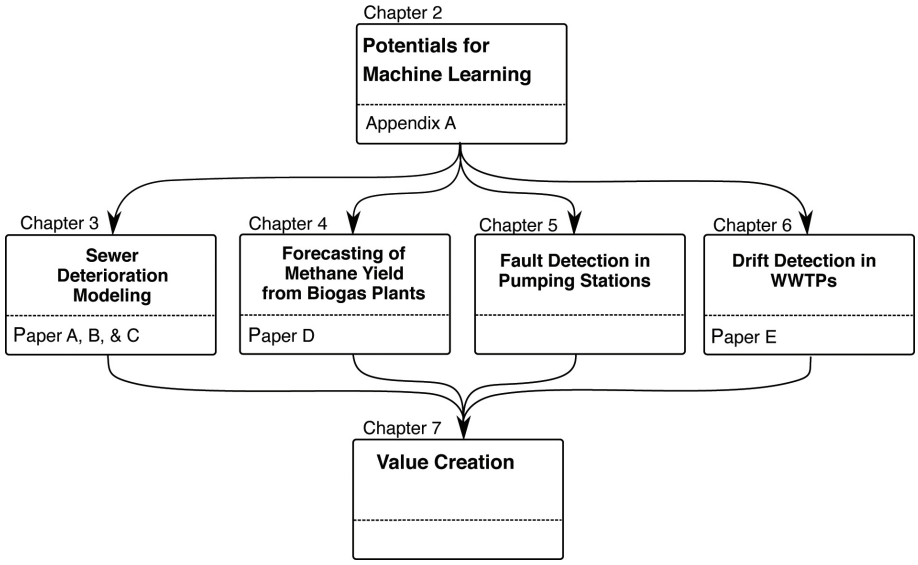


Fig. 1.1: Overview of the content of the thesis. In the figure, each box refers to a chapter. As indicated in the figure, some of the chapters present research published in the papers which can be found in Part II, while one of the chapters refers to data available in an appendix.

References

[1] "Sizing the prize: What's the real value of ai for your business and how can you capitalise?https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf," 2017. [Online]. Available: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

[2] "Regeringens klimapartnerskaber - Affald og vand, cirkulær økonimi," Mar. 2020. [Online]. Available: https://kefm.dk/media/6651/klimpartnerskab_afrapportering-for-affald-vand-og-cirkulaer-oekonomi.pdf

[3] S. Aftab, M. Ahmad, N. Hameed, M. Salman, I. Ali, and Z. Nawaz, "Rainfall prediction using data mining techniques: A systematic literature review," vol. 9, no. 5. [Online]. Available: http://thesai.org/Publications/ViewPaper?Volume=9&Issue=5&Code=ijacsa&SerialNo=18

[4] Y. Chen, L. Song, Y. Liu, L. Yang, and D. Li, "A review of the artificial neural network models for water quality prediction," vol. 10, no. 17, p. 5776. [Online]. Available: https://www.mdpi.com/2076-3417/10/17/5776

[5] L. Corominas, M. Garrido-Baserba, K. Villez, G. Olsson, U. Cortés, and M. Poch, "Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques," vol. 106, pp. 89–103. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364815217302359

[6] E. M. Dogo, N. I. Nwulu, B. Twala, and C. Aigbavboa, "A survey of machine learning methods applied to anomaly detection on drinking-water quality data," vol. 16, no. 3, pp. 235–248. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1573062X.2019.1637002

[7] European Commission. Directorate General for the Budget., *The EU's 2021-2027 long-term budget & NextGenerationEU: facts and figures.* Publications Office. [Online]. Available: https://data.europa.eu/doi/10.2761/808559

[8] M. Gholizadeh, A. Melesse, and L. Reddi, "A comprehensive review on water quality parameters estimation using remote sensing techniques," vol. 16, no. 8, p. 1298. [Online]. Available: http://www.mdpi.com/1424-8220/16/8/1298

[9] J. Gohil, J. Patel, J. Chopra, K. Chhaya, J. Taravia, and M. Shah, "Advent of big data technology in environment and water management sector." [Online]. Available: https://link.springer.com/10.1007/s11356-021-14017-y

[10] A. Hadjimichael, J. Comas, and L. Corominas, "Do machine learning methods used in data mining enhance the potential of decision support systems? a review for the urban water sector," vol. 29, no. 6, pp. 747–756. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/AIC-160714

[11] H. Haimi, M. Mulas, F. Corona, and R. Vahala, "Data-derived soft-sensors for biological wastewater treatment plants: An overview," vol. 47, pp. 88–107. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364815213001308

[12] B. D. Hansen, S. H. Rasmussen, T. B. Moeslund, M. Uggerby, and D. G. Jensen, "Sewer deterioration modeling: The effect of training a random forest model on logically selected data-groups," vol. 176, pp. 291–299. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S187705092031855X

[13] B. D. Hansen, S. H. Rasmussen, M. Uggerby, T. B. Moeslund, and D. G. Jensen, "Comprehensive feature analysis for sewer deterioration modeling," vol. 13, no. 6, p. 819. [Online]. Available: https://www.mdpi.com/2073-4441/13/6/819

[14] B. D. Hansen, D. Getreuer Jensen, S. H. Rasmussen, J. Tamouk, M. Uggerby, and T. B. Moeslund, "General sewer deterioration model using random forest," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 834–841. [Online]. Available: https://ieeexplore.ieee.org/document/9002727/

[15] J. B. Haurum and T. B. Moeslund, "A survey on image-based automation of CCTV and SSET sewer inspections," vol. 111, p. 103061. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0926580519311227

[16] U. Iqbal, P. Perez, W. Li, and J. Barthelemy, "How computer vision can facilitate flood management: A systematic review," vol. 53, p. 102030. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2212420920315326

[17] J. Jawad, A. H. Hawari, and S. Javaid Zaidi, "Artificial neural network modeling of wastewater treatment and desalination using membrane processes: A review," vol. 419, p. 129540. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S138589472101127X

[18] I. S. Khan, M. O. Ahmad, and J. Majava, "Industry 4.0 and sustainable development: A systematic mapping of triple bottom line, circular economy and sustainable business models perspectives," vol. 297, p. 126655. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652621008751

[19] L. Li, S. Rong, R. Wang, and S. Yu, "Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review," vol. 405, p. 126673. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1385894720328011

[20] M. Malek Mohammadi, M. Najafi, S. Kermanshachi, V. Kaushal, and R. Serajiantehrani, "Factors influencing the condition of sewer pipes: State-of-the-art review," vol. 11, no. 4, p. 03120002. [Online]. Available: http://ascelibrary.org/doi/10.1061/%28ASCE%29PS.1949-1204.0000483

[21] G. Mannina, T. F. Rebouças, A. Cosenza, M. Sànchez-Marrè, and K. Gibert, "Decision support systems (DSS) for wastewater treatment plants – a review of the state of the art," vol. 290, p. 121814. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960852419310442

[22] *Bekendtgørelse om spildevandstilladelser m.v. efter miljøbeskyttelseslovens kapitel 3 og 4*, Miljøministeriet, Jun. 2021. [Online]. Available: https://www.retsinformation.dk/eli/accn/B20210139305#idfe0caa41-8c37-47d9-b466-884ab594feaf

[23] M. E. Mondejar, R. Avtar, H. L. B. Diaz, R. K. Dubey, J. Esteban, A. Gómez-Morales, B. Hallam, N. T. Mbungu, C. C. Okolo, K. A. Prasad, Q. She, and S. Garcia-Segura, "Digitalization to achieve sustainable development goals: Steps towards a smart green planet," vol. 794, p. 148539. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0048969721036111

[24] K. B. Newhart, R. W. Holloway, A. S. Hering, and T. Y. Cath, "Data-driven performance analyses of wastewater treatment plants: A review," vol. 157, pp. 498–513. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0043135419302490

[25] J. Ponce Romero, S. Hallett, and S. Jude, "Leveraging big data tools and technologies: Addressing the challenges of the water quality sector," vol. 9, no. 12, p. 2160. [Online]. Available: http://www.mdpi.com/2071-1050/9/12/2160

[26] S. Rehman, M. Sahana, H. Hong, H. Sajjad, and B. B. Ahmed, "A systematic review on approaches and methods used for flood vulnerability assessment: framework for future research," vol. 96, no. 2, pp. 975–998. [Online]. Available: http://link.springer.com/10.1007/s11069-018-03567-z

[27] V. Sagan, K. T. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B. A. Greeling, S. Maalouf, and C. Adams, "Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing," vol. 205, p. 103187. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0012825220302336

[28] *Bekendtgørelse af lov om afgift af spildevand*, Skatteministeriet, Apr. 2020. [Online]. Available: https://www.retsinformation.dk/eli/lta/2020/478

[29] Z. Sokol, J. Szturc, J. Orellana-Alvear, J. Popová, A. Jurczyk, and R. Célleri, "The role of weather radar in rainfall estimation and its application in meteorological and hydrological modelling—a review," vol. 13, no. 3, p. 351. [Online]. Available: https://www.mdpi.com/2072-4292/13/3/351

[30] A. Y. Sun and B. R. Scanlon, "How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions," vol. 14, no. 7, p. 073001. [Online]. Available: https://iopscience.iop.org/article/10.1088/1748-9326/ab1b7d

[31] S. N. Topp, T. M. Pavelsky, D. Jensen, M. Simard, and M. R. V. Ross, "Research trends in the use of remote sensing for inland water quality science: Moving towards multidisciplinary applications," vol. 12, no. 1, p. 169. [Online]. Available: https://www.mdpi.com/2073-4441/12/1/169

[32] F. Tscheikner-Gratl, N. Caradot, F. Cherqui, J. P. Leitão, M. Ahmadi, J. G. Langeveld, Y. Le Gat, L. Scholten, B. Roghani, J. P. Rodríguez, M. Lepot, B. Stegeman, A. Heinrichsen, I. Kropp, K. Kerres, M. d. C. Almeida, P. M. Bach, M. Moy de Vitry, A. Sá Marques, N. E. Simões, P. Rouault, N. Hernandez, A. Torres, C. Werey, B. Rulleau, and F. Clemens, "Sewer asset management – state of the art and research needs," vol. 16, no. 9, pp. 662–675. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1573062X.2020.1713382

[33] N. Wagle, T. D. Acharya, and D. H. Lee, "Comprehensive review on application of machine learning algorithms for water quality parameter estimation using remote sensing data," vol. 32, no. 11, p. 3879. [Online]. Available: http://myukk.org/SM2017/article.php?ss=2953

[34] W. Wu, G. C. Dandy, and H. R. Maier, "Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling," vol. 54, pp. 108–127. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364815213003198

[35] Y. Wu and S. Liu, "A review of data-driven approaches for burst detection in water distribution systems," vol. 14, no. 9, pp. 972–983. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1573062X.2017.1279191

[36] Z. Yuan, G. Olsson, R. Cardell-Oliver, K. van Schagen, A. Marchi, A. Deletic, C. Urich, W. Rauch, Y. Liu, and G. Jiang, "Sweating the assets – the role of instrumentation, control and automation in urban water systems," vol. 155, pp. 381–402. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0043135419301599

[37] W. Zhang, N. B. Tooker, and A. V. Mueller, "Enabling wastewater treatment process automation: leveraging innovations in real-time sensing, data analysis, and online controls," vol. 6, no. 11, pp. 2973–2992. [Online]. Available: http://xlink.rsc.org/?DOI=D0EW00394H

[38] L. Zhao, T. Dai, Z. Qiao, P. Sun, J. Hao, and Y. Yang, "Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse," vol. 133, pp. 169–182. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957582019318403

[39] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: A review," vol. 598, p. 126266. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022169421003139

[40] M. Zounemat-Kermani, E. Matta, A. Cominola, X. Xia, Q. Zhang, Q. Liang, and R. Hinkelmann, "Neurocomputing in surface water hydrology and hydraulics: A review of two decades retrospective, current status and future prospects," vol. 588, p. 125085. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S002216942030545X

Chapter 2

Potential of Machine Learning in the Water Sector

As described in Chapter 1, the water sector is a data heavy industry which suggests a large potential for machine learning solutions. Generally, there has been a movement towards data-driven solutions in research, but out of 340 published papers on data-driven solutions in WWTPs only 16 % were implemented as software tools [1]. Even when full-scale WWTPs were used in studies on soft sensors, only a small minority of the studies reported a practical implementation of the methods [3]. Likewise, despite several decision support systems having been published in academia, only few have made it into actual frameworks [2].

Reasons for the solutions not being implemented in practice include: lack of association between computer engineering and water engineering, limited practical experience among the academics using the machine learning methods, challenges with complexity due to the intricate issues in water and wastewater systems, and the need for simpler tools and user-friendly interfaces [2, 4].

Experienced water professionals have in-depth insight into the needs of the water sector, including the processes and solutions which have a potential for optimization, and where data are available. However, there is often a lack of knowledge about how machine learning can be applied to the data to make new products and services. On the other hand, data scientists have a lot of knowledge about how data can be used but little knowledge about the water sector. Insight into both the water sector and data science are needed to make machine learning solutions which can create value in the water sector. Furthermore, the decision makers in the water sector are usually experienced water professionals who have little or no experience with machine learning. This makes it challenging for them to make qualified decisions on how to optimally invest in machine learning solutions.

Therefore, there is a need for identification of use cases for machine learning, as well as there is a need for a systematic way of assessing the use cases. This assessment should be intelligible for the decision makers and enable an enlightened discussion of which use cases represent the best investments. Thus, our first task has been to identify and assess use cases for machine learning in the water sector.

2.1 Potential use cases

The potential use cases were identified in the first half of 2019. For identifying the potential use cases, resources in EnviDan were available, which in 2019 included approximately 200 consulting engineers, biologists, software developers and economists, working within different areas of the water sector. To identify as many and relevant use cases as possible, 1½ - 2 hour meetings where held with the heads of business and development at EnviDan or engineers with corresponding insights in the specific business area of EnviDan. The business areas were clean water, sewers, climate, nature and streams, wastewater treatment, energy, informatics, and economics. The meetings contained an introduction to machine learning and a brainstorm session where potential use cases were formulated. As there are several experienced water professionals who are not titled head of business or development but have great knowledge about the water sector and can contribute with potential use cases, a five-minute presentation of machine learning and the collected use cases was given to all the employees in EnviDan. The presentation was followed by a post on the company intranet where everyone was urged to contribute with use cases. Furthermore, a coffee meeting was held with an interested customer.

The collection resulted in 55 use cases after transcription. Of these, 49 were sufficiently specific for further consideration.

To gain an overview of the ideas, a sunburst diagram showing the use cases based on the required machine learning approach was made. A modified version of this diagram is presented in Figure 2.1, as it is not possible to reveal the exact use cases because they are proprietary knowledge.

2.2 Assessment of use cases

Typically, companies use a business model to evaluate the potential of a new product or service before initiating the development process. This is also the case for EnviDan. When an employee in EnviDan has an idea for a product or service, the employee must fill out a schedule concerning ten different factors which can impact the idea. However, it would be too cumbersome and time consuming for the decision makers to go through several pages with descriptions for each of the 49 potential use cases for machine learning. Instead, the schedule for assessing ideas in EnviDan was analyzed, and based

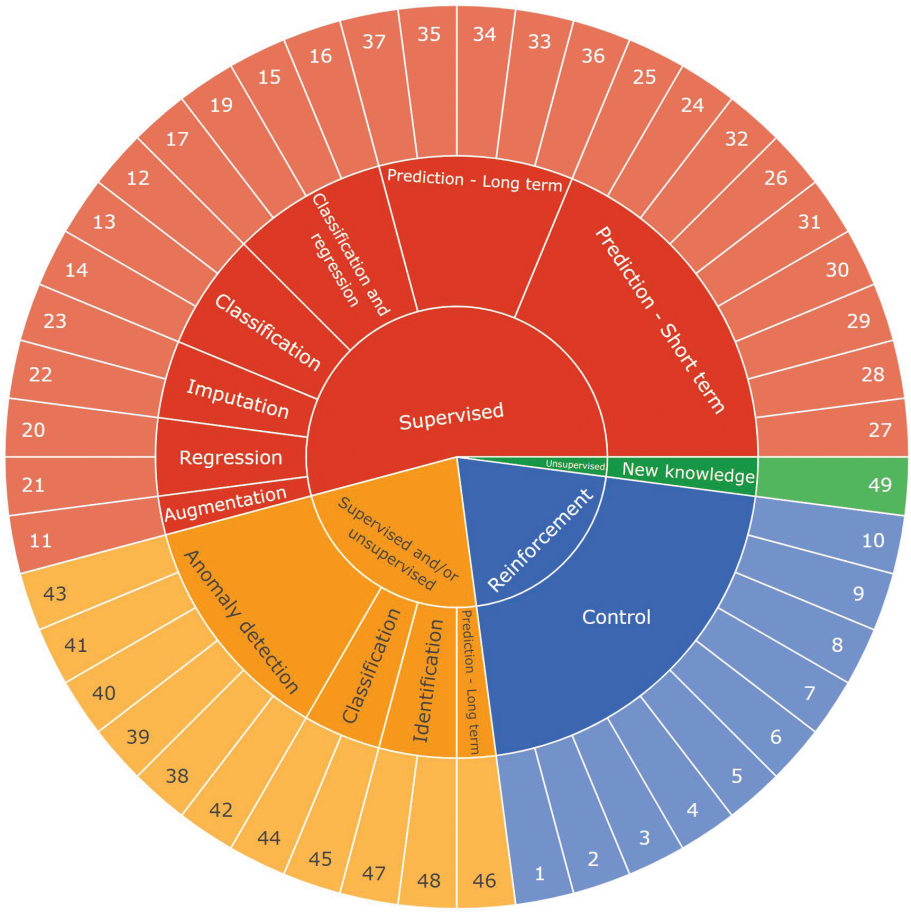


Fig. 2.1: Sunburst diagram showing the distribution of the ideas according to machine learning methods and applications. The names of the potential use cases have been replaced by numbers, as the specific use cases are proprietary knowledge

on this analysis, three keystones were extracted. These keystones were the economic potential, the required investment and the risk related to development. Furthermore, legislation hindering the use case and value, which could not be transformed to an economic potential, were considered in the business model. Based on this, use cases which could not be made within the framework of the law were excluded, and subsequently models for economic potential, investment and risk were made for the remaining cases.

2.2.1 Model for potential

The potential of a product depends on the customers willingness to pay for the product, the resale potential, and the potential savings. Therefore, the potential was calculated as shown in equation 2.1:

$$Potential = willingness_to_pay \cdot n_customers + savings \tag{2.1}$$

For each business area in EnviDan, the head of the business area, or a substitute with similar experience, assessed the use cases according to willingness to pay, number of customers over five years and the expected savings over five years. The specification of five years was inspired from EnviDan's business model. As it was impossible to give precise estimates of these parameters, checklists with intervals were used. The checklists can be found in Appendix F Table F.1, F.2 and F.3.

2.2.2 Model for investment

Two factors influence the investment related to a product: the required working hours and the material prices. For simplicity, material prices were excluded from this analysis as the main expense in data science is working hours, especially if the data have already been acquired or the material needed for acquisition is already available, which was the case for most of the proposed use cases. Based on this, the required investment was considered equal to the time spent on the project.

Estimating the working hours related to a data science project is associated with high uncertainties due to unforeseen factors related to, e.g., data quality, data accessibility and experience by the developers. Therefore, a model for the time duration is deemed to be imprecise. To accommodate the large uncertainties related to data science projects the decision was made to not make exact time estimates. Therefore, the time duration related to each of the use cases was assessed on a scale from zero to one.

The time duration related to a project was considered as the time duration needed for development and time duration needed for adjustment to different customers, as illustrated in Equation 2.2.

$$time = \alpha_1 \cdot time_{dev} + \alpha_2 \cdot time_{adj} \tag{2.2}$$

In Equation 2.2, the α_1 and α_2 are constants used for tuning the weight between the time for development and time for adjustment. Details on the α values can be found in Appendix F Table F.12. The time for development was considered to depend on the data accessibility and the complexity of the solution, as illustrated in Equation 2.3.

$$time_{dev} = \alpha_3 \cdot data_{access} + \alpha_4 \cdot complexity \tag{2.3}$$

In Equation 2.3, $data_{access}$ is the accessibility of the data, and *complexity* is the complexity of the solution. α_3 and α_4 are constants used for tuning. Details on the

factors affecting the data accessibility can be seen in Appendix F Table F.4. The factors influencing the complexity of the solution considered here are the number of input parameters, number of output parameters, the correlation between input and output, the required machine learning method, and whether more complex parameters were used as input. The complexity was calculated, as illustrated in Equation 2.4.

$$complexity = \alpha_5 \cdot n_{in} + \alpha_6 \cdot n_{out} + \alpha_7 \cdot corr_{in_out} + \alpha_8 \cdot type_{in} + \alpha_9 \cdot method \quad (2.4)$$

In Equation 2.4, n_{in} refers to the number of input parameters, n_{out} refers to the number of output parameters, $corr_{in_out}$ refers to the correlation between input and output, $type_{in}$ refers to the type of input parameters, $method$ refers to the required machine learning method and α_5 to α_9 are constants used for tuning. Details on the parameters can be found in Appendix F Tables F.5, F.6, F.7, F.8, F.9, and F.12.

The time duration for adjustment to different customers was estimated based on the adjustment needed per sale, as shown in Equation 2.5.

$$time_{adj} = adj_{amount} \cdot n_customers \quad (2.5)$$

In Equation 2.5, adj_{amount} refers to the amount of adjustment required per customer. Details on the parameters can be seen in Tables F.10 and F.2.

2.2.3 Model for risk

The risk related to investing in a use case was considered as the time spent on development and the likelihood of the solution not reaching the minimal viable product (MVP), or that the time for development would exceed the expectations, which has been expressed in Equation 2.6.

$$Risk = Time_{dev} \cdot (\alpha_{10} \cdot Likelihood_{MVP_not_feasible} + \alpha_{11} \cdot Likelihood_{time_overrun}) \quad (2.6)$$

In Equation 2.6, $Likelihood_{MVP_not_feasible}$ is the likelihood that the MVP is not feasible, and $Likelihood_{time_overrun}$ is the likelihood for exceeding the expected time duration. α_{10} and α_{11} are parameters for tuning of the $Likelihood_{MVP_not_feasible}$ and $Likelihood_{time_overrun}$ and can be found in Appendix F Table F.12. A key hindering for obtaining a MVP is lack of data quality. However, a fair evaluation of the data quality is not feasible for a large number of use cases, as it would require in depth knowledge of the data. Therefore, the decision was made to focus on the correlation between input and output of the model as this is a more narrow scope than data quality. Another important factor is the minimum model performance required for reaching the MVP. The formula for $Likelihood_{MVP_not_feasible}$ is shown in Equation 2.7.

$$Likelihood_{MVP_not_feasible} = Corrin_out \cdot min_req \tag{2.7}$$

In Equation 2.7, min_req is the minimum requirements in terms of precision which need to be met for the solution to create value. Details can be found in Appendix F Table F.11.

Finally, the likelihood of a task to exceed the development time depends on the complexity of the algorithm, as complex solutions often entail a higher likelihood of unforeseen elements. Furthermore, there is a higher risk for exceeding the time limitations if the MVP requires a high performance than if a low performance is acceptable.

The formula for $Likelihood_{time_overrun}$ is shown in Equation 2.7.

$$Likelihood_{time_overrun} = complexity \cdot min_req \tag{2.8}$$

2.2.4 Selection of weights of the different parameters

In the formulas for calculating the potential and risk related to the use cases, the parameters were weighted according to each other with α values. Selection of these values was based on how important the different parameters were. For instance, when considering the complexity of a solution, the correlation between input and output is considered more important than the number of input parameters. Therefore, α_7 was assigned a higher value than α_5. The α values in this evaluation were based on EnviDan's experience level regarding machine learning in the spring of 2019. If this analysis were to be remade in 2022, several parameters would need to be changed as EnviDan's experience level within machine learning has increased.

Furthermore, the time required for development is highly dependent on factors such as the skill levels of the person developing the model, access to sparring and supervision, data accessibility and licenses. These factors were considered when assigning values in the lookup tables. For instance, EnviDan had some experience with supervised learning and no experience with reinforcement learning in 2019. For this reason, reinforcement learning would assign a higher value to the method parameter than supervised learning would.

2.3 Selection of use cases

The results of the assessment are shown in Figures 2.2, 2.3, and 2.4. In the figures, all use cases are plotted according to the economic potential and the investment. Furthermore, the risk is indicated as *blob size* so that ideas with low risk are represented by large blobs. The blobs in the three figures are color-coded according to area, machine learning approach, and whether they represent a value which could not directly be measured as an economic benefit, such as customer retention. In the figures, the economic potential

of the use cases looks quantified. The reason for this is that checklists, and thereby all use cases within the same interval, were assigned the same value. This was also the case for the models for investment and risk, though because the investment was evaluated on several more parameters than the economic potential, it did not result in a quantification of the investment in the figures.

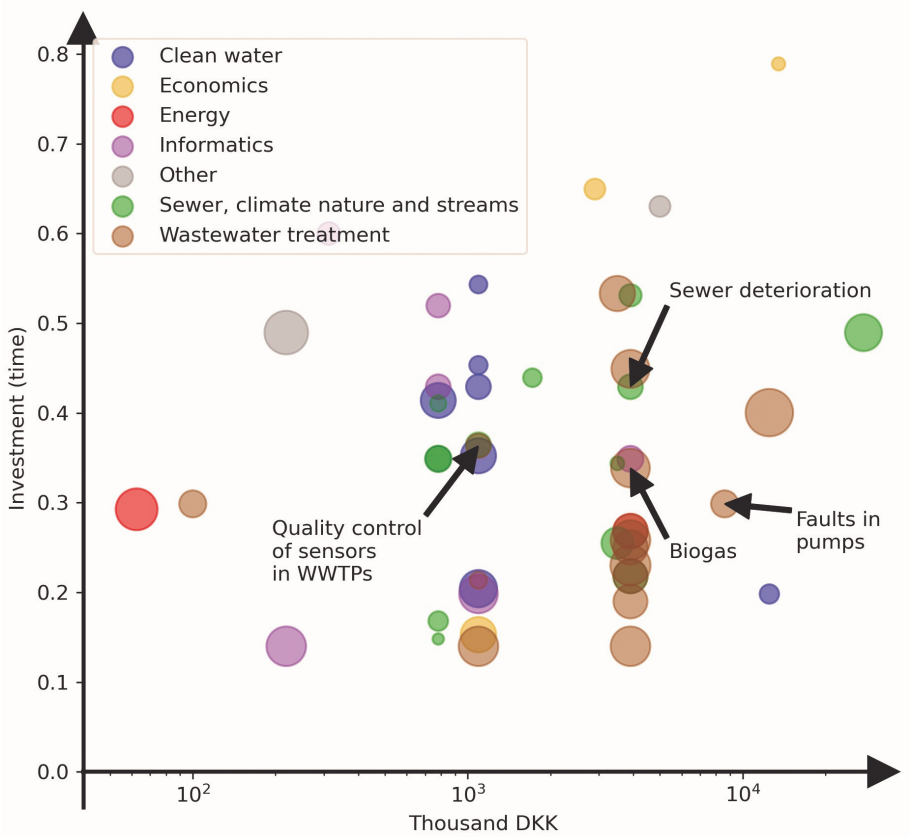


Fig. 2.2: Results of the assessment color-coded according to subject area. The first axis shows the economic potential, and the second axis shows the investment as time duration. The time duration is assessed on a scale from 0 to 1, as the time estimates represent the use cases according to each other rather than the exact time duration. The sizes of the blobs are negatively correlated with the risk, e.g., the larger the blob is, the smaller the risk is. Use cases which were selected for research in this work are marked with arrows.

The results of the assessment were presented for a group of decision makers in EnviDan, followed by a discussion of which use cases were relevant for development in general and which were relevant for research in work.

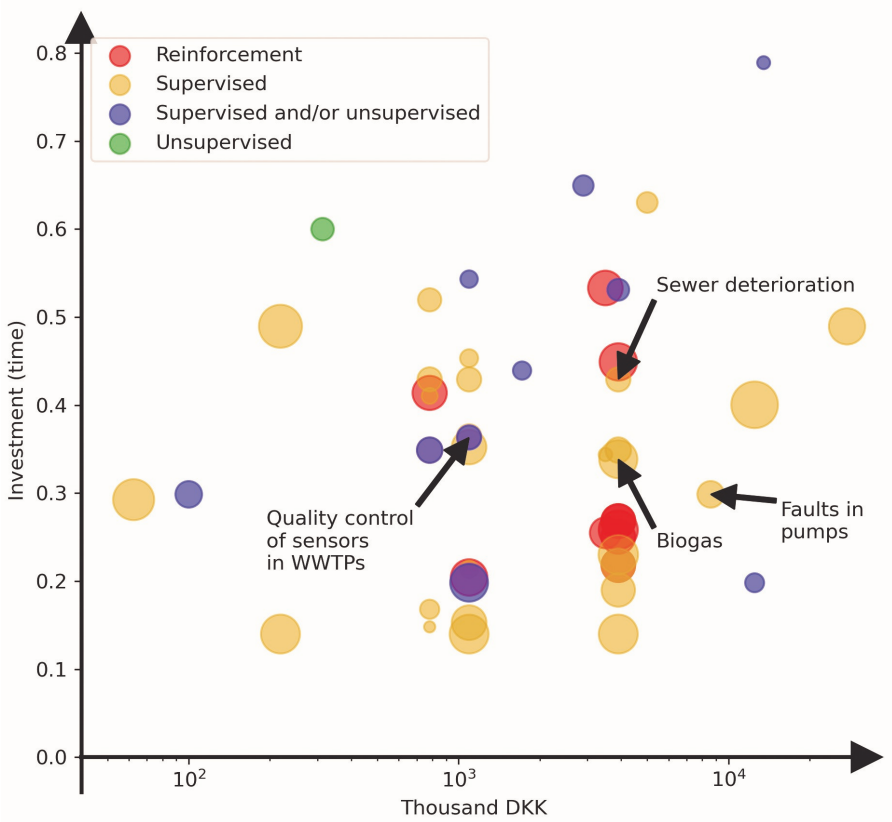


Fig. 2.3: Results of the assessment color-coded according machine learning type. The first axis shows the economic potential, and the second axis shows the investment as working hours. The sizes of the blobs are negatively correlated with the risk, e.g., the larger the blob is, the smaller the risk is. Use cases which were selected for research in this work are marked with arrows.

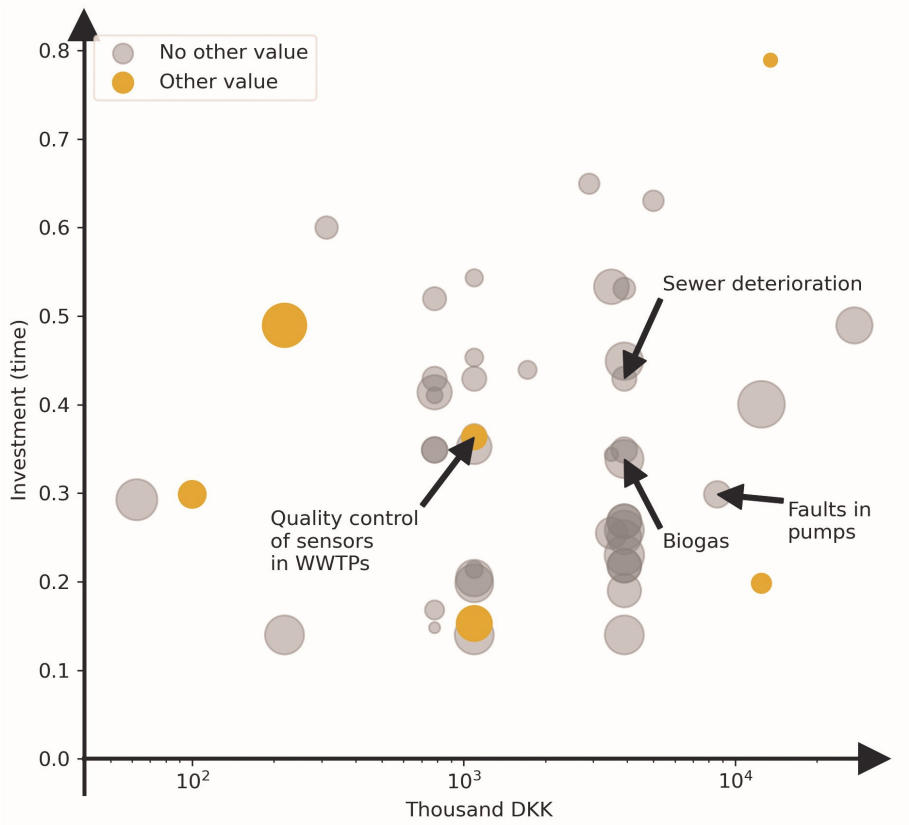


Fig. 2.4: Results of the assessment color-coded according to whether the use case represents value which cannot be measured as an economic potential. The first axis shows the economic potential, and the second axis shows the investment as working hours. The sizes of the blobs are negatively correlated with the risk, e.g., the larger the blob is, the smaller the risk is. Use cases which were selected for research in this work are marked with arrows.

As some of the intentions with this work were to both increase knowledge of machine learning in EnviDan generally and within the different businesses of EnviDan, it was sought to select use cases within different subject areas and use cases which required different machine learning techniques. Initially three use cases were selected. These were *sewer deterioration modeling focusing on ageing curves*, *fault detection focusing on a general system for implementation in EnviDans softwear products*, and *control of air pumps in WWTPs*. The machine learning method required for development within these three use cases were supervised learning, supervised and/or unsupervised learning, and reinforcement learning. However, due to time constrains and challenges related to data quality, the actual working areas in this work were *sewer deterioration modeling*, *prediction of methane yield from biogas plants*, *fault detection in pumps*, and *Drift detection in WWTPs*, which was a research-relevant sub-task within *Quality control of sensors in WWTPs*. The use cases developed in this project are marked with arrows in Figures 2.2, 2.3, and 2.4.

2.4 Contributions

- An overview of potential use cases for machine learning for value creation in the water sector provided and analyzed from a business perspective.
- The analysis proved to be an efficient tool providing a common basis for discussion between experienced water professionals and data scientists. The systematic way of assessing each use case entailed that there was an argument behind the assessment of each use case, making it easy to explain why one case was more complex than another.
- A new method for making an intuitive visualization of different use cases was provided. This is a significant contribution for EnviDan as the method has subsequently been used for evaluation of different solutions by other engineers in EnviDan.
- The details in the developed method are adjusted for EnviDan. However, it is suggested that the method can be adjusted for companies other than EnviDan and other subject areas than machine learning in the water sector.

References

[1] L. Corominas, M. Garrido-Baserba, K. Villez, G. Olsson, U. Cortés, and M. Poch, "Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques," vol. 106, pp. 89–103. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364815217302359

[2] A. Hadjimichael, J. Comas, and L. Corominas, “Do machine learning methods used in data mining enhance the potential of decision support systems? a review for the urban water sector,” vol. 29, no. 6, pp. 747–756. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/AIC-160714

[3] H. Haimi, M. Mulas, F. Corona, and R. Vahala, “Data-derived soft-sensors for biological wastewater treatment plants: An overview,” vol. 47, pp. 88–107. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364815213001308

[4] G. Mannina, T. F. Rebouças, A. Cosenza, M. Sànchez-Marrè, and K. Gibert, “Decision support systems (DSS) for wastewater treatment plants – a review of the state of the art,” vol. 290, p. 121814. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960852419310442

Chapter 3

Sewer Deterioration Modeling

The urban drainage system is essential for modern society, and reduced functionality of sewers can result in exfiltration and flooding, which can affect several externalties such as public health, property, the environment and traffic disruption [28, 30]. Furthermore, breakdowns in the system can entail significant damage to the surrounding infrastructure such as roads and buildings [17]

Due to the sewers being underground, they are in general very hard to monitor. Today, monitoring is done through Closed-Circuit Television (CCTV) inspection, where an operator sends a remote controlled CCTV robot into the sewer and manually annotates all observations [30]. This is a labor-intensive process, so utilities only have enough resources to inspect a small percentage of their sewer pipes every year. Therefore, the utilities need to prioritize which pipes to inspect.

Historically, risk-based rehabilitation has been area-based in Denmark. In area-based rehabilitation, all the pipes in a given sewer network or part of a sewer network are inspected based on the experience of the operators and pipe age. Subsequently, it is decided if the inspected network should be rehabilitated.

When a network is selected for rehabilitation all the pipes are rehabilitated or replaced, despite that some of them could easily be functional for several more years. This is done to ensure that all the pipes can last until next time the network is inspected. To extend the pipes' lifetime and thereby save resources and economic costs, there is a movement toward risk-based planning of CCTV-inspection and rehabilitation of individual pipes [16].

Pipe level-based asset management entails new requirements to software-based asset management systems as the utilities now need to keep track of tens of thousands individual pipes compared to a limited number of areas, which was the case before. To meet these needs, several risk-based asset management systems have been developed [1, 12–14], typically consisting of a deterioration model and a consequence model.

The deterioration model predicts the condition of a pipe or the likelihood of a pipe to be in a given condition, and the consequence model estimates the severity of a pipe failure [30].

3.1 Sewer deterioration modeling

This section contains an overview of the methods previously used for sewer deterioration modeling and how a general model was made for several Danish utilities.

3.1.1 Existing approaches

Sewer deterioration is affected by several different factors. When developing deterioration models, the researchers typically select input parameters based on expert knowledge about which factors affect the condition. Table 3.1 shows the factors which have previously been considered relevant in the literature, including physical, environmental, operational, and constructional factors.

Table 3.1: Factors affecting the deterioration of sewers. The overall taxonomy is based on [3] and the details are a combination of [3, 23, 24].

Factors influencing the deterioration of sewers			
Physical	*Environmental*	*Operational*	*Constructional*
Age	Infiltration/exfiltration	Sediment level	Standard of
Size	Groundwater level	Maintenance and	workmanship
Shape	Presence of trees	repair strategies	Installation method
Length	Traffic and surface loading	Sewage characteristics	
Depth	Soil/backfill type	Yearly sewage flow	
Material	Precipitation		
Type			
Slope			
Joint type and material			

It is worth noticing that the factors listed in the table are not exclusively the factors which influence the deterioration. For instance the land use (e.g., city or industry) influences the load, hydraulics, and sewage characteristics. Furthermore, other factors such as number of road grates or buildings might be relevant from a model perspective as it indirectly contains information on factors such as sewage characteristics and hydraulics.

Several different methods for sewer deterioration modeling have been presented in the literature. An overview of the methods can be seen in Table 3.2.

Sewer deterioration models can either be on a network level or a pipe level. On a network level Markov Chain and Survival Analysis are the most reliable, whereas on a

Table 3.2: Overview of methods for sewer deterioration modeling based on [3, 19, 26]

Sewer deterioration models		
Deterministic	*Statistical*	*Artificial Intelligence*
Rule-based Simulation	Discriminant Analysis	Support Vector Machine
	Markov Chain	Decision Trees
	Semi Markov Chain	Random Forest
	Cohort Survival	Bayesian Networks
	Regression	Artificial Neural Networks
	- Logistic	Fuzzy Logic
	- Binary	Evidential Reasoning
	- Linear/Multi Linear	
	- Exponential	
	- Ordinal	

pipe level machine learning models and statistical regression models are best suited for detecting pipes in critical condition. [30]

On a pipe level the deterioration models either predict the probability of a sewer to be in a certain condition or the exact condition of the sewer [17]. In this work, focus has been on models which predict the exact condition of the sewers.

There is no consensus in the definition of when a pipe is in a certain condition, as it typically follows local standards such as the European standard [23], German standard [20], Norwegian standard [29], and the Danish standard [17]. Furthermore, some authors combine multiple condition classes into fewer classes [21] or define the condition classes based on specific needs for a utility [16]. The performance of the models are generally low, which has motivated researchers to make binary deterioration models or to evaluate the performance of the models by combining the classes, which also makes them easier to compare to each other [17, 29]. Another challenge is a lack of consensus in the published models regarding the balance between sensitivity and specificity. However, one work [23] fixed the sensitivity to 0.80 based on inputs from a utility. The performance of the models are highly dependent on the used dataset and how the condition is defined. Furthermore, due to privacy issues the datasets are usually not publicly available, making it difficult to compare the models.

Most of the deterioration models have been developed based on data from a single city [2, 4, 8, 18, 20, 22, 29, 31], and in a few cases data from two cities [11] or an area [23]. However, as not all utilities are in possession of sufficiently large datasets, it would be favorable if the models were not dependent on individual utilities.

3.1.2 Suggested model

In this section an overview of our work of developing a general sewer deterioration model, which was trained and tested on pipe data from several different utilities, is presented. Further details can be found in Paper A.

CCTV inspection reports for pipes and pipe sections were available from 35 different Danish utilities. The condition of the pipes was evaluated using an adjusted version of the physical index (TS), which is a Danish standard for measuring the condition of the pipes. The TS is measured on a continuous scale from 0-10. For each of the inspected pipes, 47 parameters were extracted from EnviDan's data portal, GIS, and other online databases. An overview of the parameters can be found in Paper A Tables A.1 and A.2. In total, all parameters were available for 146,856 of the inspection reports.

Three versions of the dataset were used: one with the full dataset, one without geographical information and one where pipes were removed from the dataset to obtain an approximately equal distribution of pipes within the different TS.

The dataset was divided into a training set containing 90 % of the data points and a test set containing the remaining 10 % of the data points. 15 different machine learning algorithms and ensemble of the best performing algorithms were tested. Based on these tests and the literature the decision was made to use Random Forest for the model.

As Random Forest is known for introducing a bias towards the average performance, we found the linear transformation which described the bias when predicting the TS of the training data, as shown in Figure 3.1, and utilized it on the predictions of the testset.

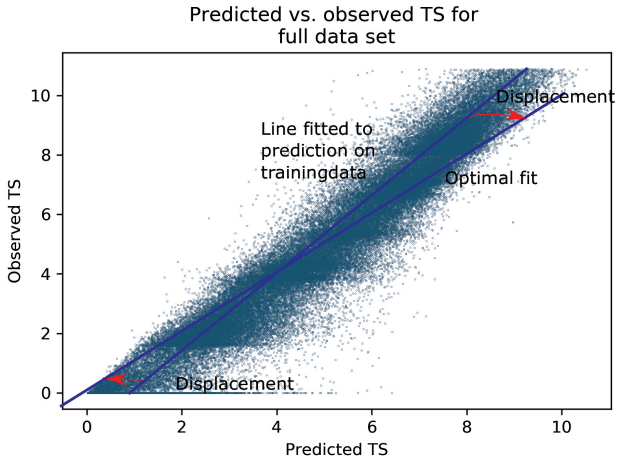


Fig. 3.1: Illustration of the bias in the predictions for the training data. The figure is adapted from Hansen et al. [17].

To compare the performance of the algorithms to performances seen in the literature, the TS was divided into respectively five and two condition states as shown in Figure 3.2.

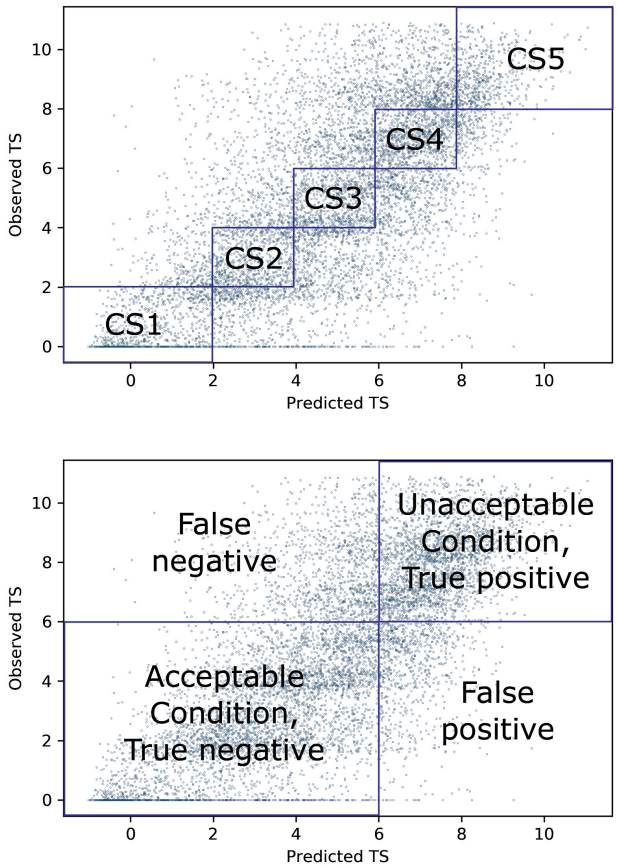


Fig. 3.2: Illustration of how the TS was divided into five and two condition states. The figure is adapted from Hansen et al. [17]

For a utility, it can be beneficial to define the sensitivity of the model and, with inspiration from Laakso et al. [23], the decision was made to fix the sensitivity to 0.80 for the binary evaluation. This was done by moving the threshold for when to categorize a pipe to be in bad condition, as shown in Figure 3.3.

The results showed that the model was able to obtain state-of-the-art performance.

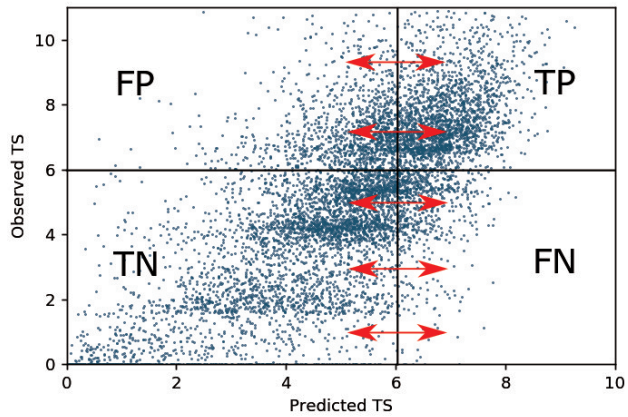


Fig. 3.3: The sensitivity can be adjusted by moving the threshold for when a pipe is predicted to be in good or bad condition as shown by the red arrows. The figure is adapted from Hansen et al. [17].

The performance of our model compared to other models presented in the literature can be found in Table 3.3. Thereby a model was simultaneously developed for several different utilities. This is beneficial compared to developing a model for each individual utility, which does not necessarily have a sufficiently high number of inspected pipes.

3.1.3 Contributions

- We made a model with state-of-the-art performance when applied to pipes from several different utilities. This is contrary to existing models which have typically been applied to pipes in a single or few cities.
- The model setup was implemented in an asset management tool for risk assessment in EnviDan.

3.2 Optimization of sewer deterioration models

Despite the model presented in Section 3.1.2 obtained state-of-the-art performance, it was expected that the performance could be further improved by combining the model with expert knowledge on the factors related to deterioration.

Table 3.3: Sensitivity and specificity for our model and models presented in the literature. The table is modified from [17].

	Sensitivity	**Specificity**	**Size of test set (accept/unaccept)**
Our model			
full data set	0.80	0.75	(9233/5453) pipes
AED-data* set	0.80	0.76	(5252/3532) pipes
AED-data set without geo. info	0.80	0.74	(5290/3494) pipes
Harvey and McBean 2014	0.82	0.73	(318/38) pipes
Fuchs-Hanusch 2015	0.55**	0.73**	
Kabir 2018			
Cementitous	0.64	0.87	(1698/69) pipes
Clay	0.75	0.86	(414/68) pipes
Metallic	0.66	0.97	(157/3) pipes
Plastic	0.50	0.98	(129/4) pipes
Average	0.63	0.90	
Laakso2018			
RF model Original	0.50	0.80	∼112 km pipes***
RF model FNR fixed on 0.20	0.80	0.47	∼112 km pipes***

*AED-dataset refers to approximately equally distributed dataset. **Read from graph. ***This number has been calculated from the information that 30 % of 1241 km pipes have been inspected, and 30 % of these were used as test data.

3.2.1 Existing approaches

As described in Section 3.1.1, several different methods for sewer deterioration modeling are present in the literature. From other domains the combination of expert knowledge and machine learning resulted in better solutions than only using machine learning. For example, a combination of expert knowledge and machine learning has proven efficient for classifying livestock herd types [6]. Furthermore, water professionals at EnviDan suggested that some factors, such as pipe material, affected the deterioration of the pipes to a degree where it was hardly fair to consider one model for all the pipes. Therefore, a water engineer with more than 20 years experience within urban drainage systems was asked to identify the most important variable for sewer deterioration, and, in case the amount of data was sufficient, sub-variables. The water engineer stated that material type was the most important variable. For concrete and plastic pipes

the amount of data was considered sufficient for utilization of sub-groups. The most important variable within these two data groups were content type and road type, respectively. An overview of the most important parameters and the number of pipes within each of the data groups can be found in Figure 3.4.

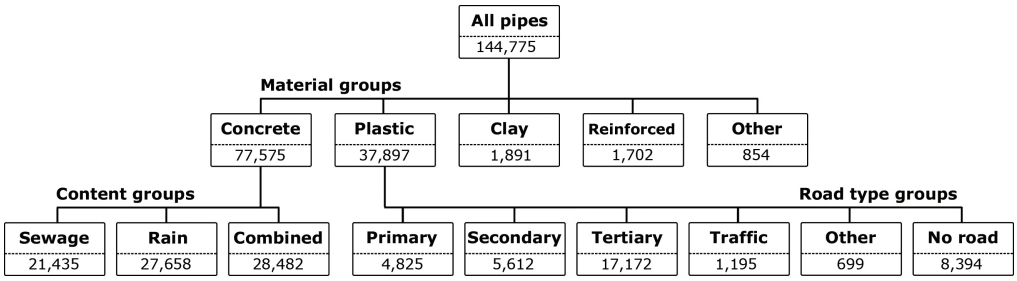


Fig. 3.4: Most important parameters related to sewer deterioration and the number of pipes in each of the data group. The figure is modified from [15].

Training models for specific material types is a reasonable consideration for optimizing the performance of the models and is essential for some statistical models [22, 25]. However, this leaves the question of whether it can contribute to machine learning models.

3.2.2 Training on logically grouped datasets

This section is a presentation of the work documented in Paper B. In this work the model presented in Section 3.1.2 is trained on the data groups presented in Figure 3.4.

The results of training a model on the material-specific data groups is shown in Figure 3.5, where it is compared to training a model on all the data. As seen in the figure, the distribution of the predictions of the general model, when considering the different material groups, and the material specific models are very similar.

To ensure a fair comparison between the general model and the material-specific models, each model was trained 10 times, and the mean and standard deviation of the sensitivity, specificity and precision was found for each material group, when the sensitivities were fixed to 0.80. The results are presented in Table 3.4

The results showed that training the model on material-specific data groups did not increase the performance of the model. For material groups with many data points the standard deviation was 0.00-0.02 for the sensitivity, specificity and precision whereas the standard deviation of the precision reached up to $\pm$ 0.10 for small data groups. Similar results were found when training the model on concrete pipes used for sewage, rain water, and combined water, respectively. For the plastic pipes it was not possible to make a fair evaluation of the road-based subgroups. This was due to a lack of plastic pipes being in bad condition when dividing the dataset into six different road types.

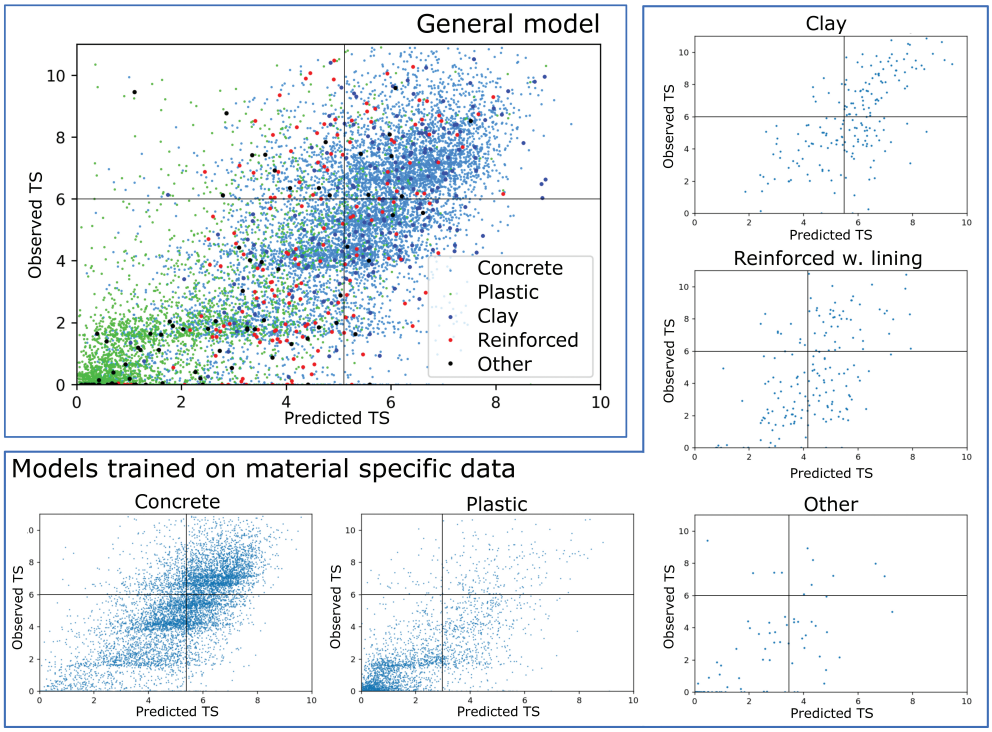


Fig. 3.5: Plots of the results obtained using the general model and the material specific models. The figure is slightly modified from [15].

3.2.3 Contributions

- We showed that the performance of the Random Forest model presented in Section 3.1.2 did not improve when training on logically grouped data. This was a question raised by water professionals, and had, to the best of our knowledge, not previously been investigated in the literature.

3.3 Features affecting the performance

The fact that the model performance did not increase when using logically grouped datasets motivated for an investigation of which features contributed to the performance. Furthermore, feedback from EnviDan's customers showed that the performance of the models varied between the utilities. For this reason, the decision was made to make an analysis of which features affected the model performance for different utilities.

Table 3.4: Sensitivity, specificity and precision for the general model when considering each data group individually and for the models trained on the material-specific data sets. The table is modified from Hansen et al. [15].

	General model			Models trained on material-specific data		
	Sensitivity	↑Specificity	↑Precision	Sensitivity	↑Specificity	↑Precision
All pipes	0.80 ± 0.00	0.78 ± 0.01	0.62 ± 0.01	0.80 ± 0.00	0.73 ±0.00	0.57 ± 0.00
Concrete	0.80 ± 0.00	0.69 ± 0.02	0.65 ± 0.01	0.80 ± 0.00	0.69 ± 0.01	0.65 ± 0.01
Plastic	0.80 ± 0.00	0.82 ± 0.00	0.24 ± 0.02	0.80 ± 0.01	0.83 ± 0.02	0.25 ± 0.02
Clay	0.80 ± 0.01	0.60 ± 0.01	0.66 ± 0.07	0.81 ± 0.02	0.59 ± 0.11	0.65 ± 0.07
Reinforced w. lining	0.80 ± 0.01	0.52 ± 0.04	0.44 ± 0.05	0.82 ± 0.03	0.55 ± 0.03	0.45 ± 0.04
Other material	0.82 ± 0.05	0.76 ± 0.06	0.44 ± 0.08	0.80 ± 0.05	0.80 ± 0.09	0.36 ± 0.10

3.3.1 Features which have previously shown to be important

Sewer deterioration is affected by several factors. However, if a feature describing each of these factors was extracted, not all of the extracted features would contribute to the model performance. Furthermore, there is a lack of consensus regarding which factors contribute to the performance of the models. In 2020 Mohammadi et al. [27] reviewed 24 papers on sewer deterioration models and the corresponding predictor variables. 19 variables were considered, and in 19 of the papers it was stated whether the variables contributed to the performance. The considered parameters were age, material, diameter, depth, length, slope, sewer type, location, up invert, down invert soil type, bedding type, groundwater, corrosivity, road type, number of trees, traffic, flow and hydraulic. In addition to the 19 variables, some of the papers also used other predictor variables. An overview of how many times the different parameters were included and in how many percentage of the cases it contributed to the performance is presented in Table 3.5.

Table 3.5: Overview of how often a parameter was included and how many percentage of the times the parameter contributed to the performance in 19 models reviewed by Mohammadi et al. [27].

	Age	Material	Diameter	Depth	Length	Slope	Sewer type	Location	Up invert	Down invert	Soil type	Bedding type	Groundwater	Corrosivity	Road type	No. Trees	Traffic	Flow	Hydraulics
Times used	18	15	17	16	11	12	6	5	1	1	5	2	3	2	5	5	1	3	2
Percentage of times contributing	78	67	71	44	91	42	83	40	0	0	20	100	100	50	40	60	100	67	100

As seen in the table, there was a large discrepancy regarding which variables contributed to the performance. However, in most cases age, material, diameter, depth and length contributed to the predictive performance. It is suggested that factors such as the variations in available predictor variables, definition of target variables, and distribution of the dataset can impact which predictor variables are found to contribute

to model performance. Furthermore, several different approaches for evaluation of the feature importance have been used in the literature. For instance, Laakso et al. [23] used the Boruta algorithm, Davis et al. [10] used backward selection and Yin [32] used backward variable elimination. Caravalho et al. [9] tested eight different methods for evaluation of the feature importance and obtained very different results for the different methods. For instance, if analysing the feature importance included stepwise removal of features from the dataset, the remaining features contributed more, due to less redundancy in the features. This redundancy is not encountered for when, for instance, using the build-in Random Forest method, and thereby features with high redundancy will be less important when using this method.

To obtain a better insight in how the feature importance varies across different utilities, the decision was made to make a comprehensive feature analysis.

3.3.2 Comprehensive feature analysis

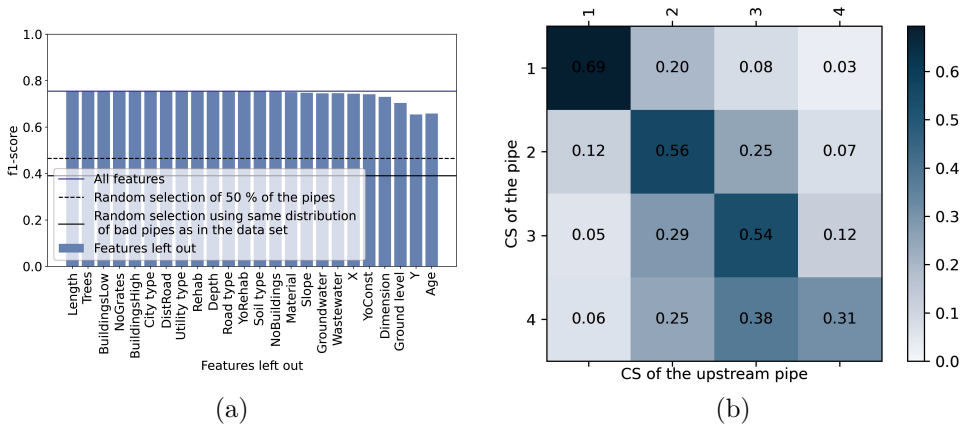
This section is a presentation of the work presented in Paper C.

Inspection reports from 35 different utilities were accessed. Contrary to previous work, the pipes were categorised into four condition states. The definition of the condition states was made with inspiration from a utility, and 24 feature groups were extracted for each pipe. The reason for using feature groups instead of features is that presence of some of the features would exclude presence of other features. For example, if the pipe is made from concrete, it cannot also be made from plastic. Thereby all features related to material were collected in a feature group called material.

All inspected pipes with the 24 features available were used for making a backward step analysis. In the backward step analysis the least contributing feature, step-wise, was removed. The performance was found as the $F1_{score}$ when considering pipes in condition states one and two as being in good condition and pipes in condition states three and four as being in bad condition. The first results of the analysis showed that features directly related to the geographical position of the pipes contributed much to the model performance. This can be seen in Figure 3.6a. Further investigations showed a large correlation between the condition state of topographically connected pipes, as seen in Figure 3.6b.

Further details on the correlation between defect types in topographically connected pipes can be found in Paper C.

The large correlation between topographically connected pipes represents a challenge in the development of deterioration models: Historically, the sewer pipes have been inspected area-wise. Furthermore, the data is typically randomly divided into a training and a test set. Thereby the training and test sets come from similar areas while the models are typically used in areas which have not been inspected, as illustrated in Figure 3.7. This is problematic as sewers in the same area typically share several known and unknown features such as the constructor and quality of the installation, year laid,



(a) (b)

Fig. 3.6: (a) Backward step analysis for the baseline. The charts indicate the performance of the model when a given feature and all the features to the right were used for training. The dashed line indicates the performance when using a random selection of 50% of the pipes, and the solid line shows the performance when using a similar distribution of bad pipes as present in the dataset. (b) The Normalized confusion matrix for the Condition States (CS) of typographically connected pipes. The figure is adapted from Hansen et al. [16].

age at the time of inspection, terrain, etc. However, despite removing location specific parameters is a step in the right direction, it does not solve the problem. This is because several factors, such as year laid and age, are still shared for the inspected area, whereby the pipes from the same areas still are more similar than pipes in areas which have not been inspected. As a result, the performance of the models is higher when testing them on a randomly divided dataset, than if, e.g., testing on data from another utility. Prospectively it is suggested to include a certain number of randomly selected pipes in the inspection plan.

After removing features related to the geographical position of the pipes the backward step analysis was remade. This can be seen in Figure C.3.

Subsequently, a backward step analysis for each utility with more than 100 pipes in bad condition was made, and it was discovered that for some utilities, some of the features could not be accessed. Consequently, a decision was made to exclude a feature for the utilities if it was not available for at least 20 % of the inspected pipes. The results from the utility specific feature analysis can be seen in Table 3.6.

The results showed that there was a high variance regarding which features contributed to the performance and in the number of features which contributed to the performance. A part of the large variations regarding the feature importance for the different utilities could be explained by redundancy between the features. For instance, there is a large degree of redundancy between age, year of construction and year of rehabilitation. For 79 % of the utilities, either age, year of construction, or year of

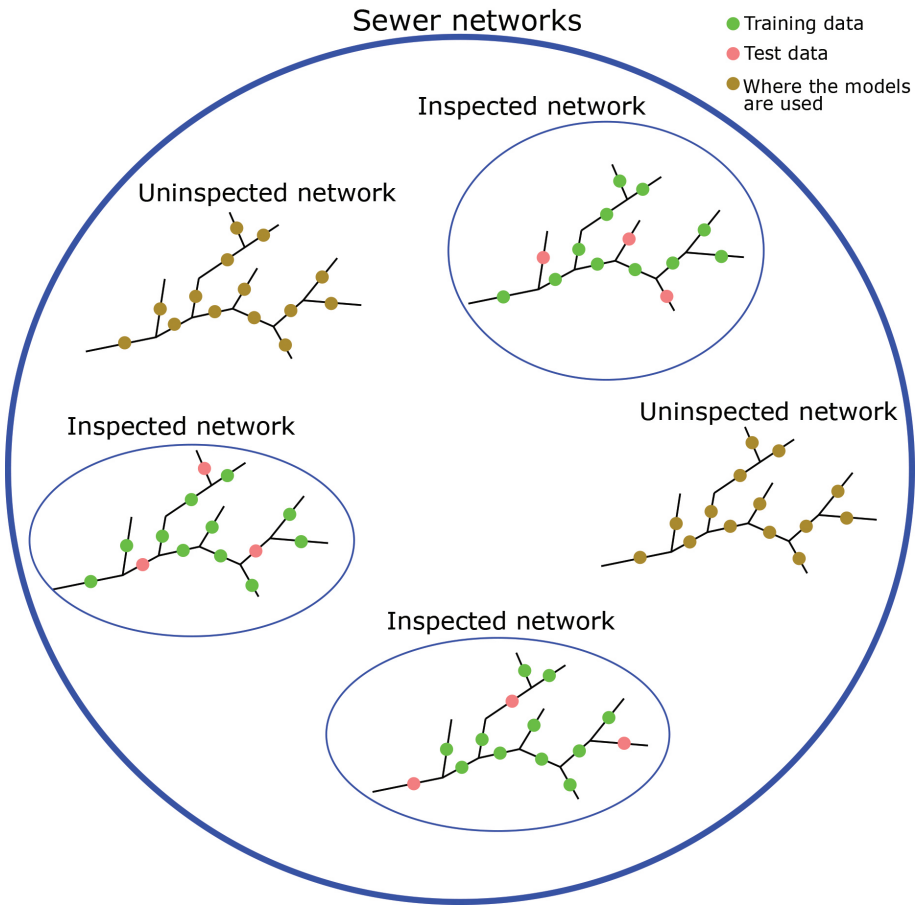


Fig. 3.7: Illustration of a typical historical inspection strategy and how this affects the training data, test data and where the models are typically used.

Table 3.6: Overview of which of the parameters contributed to the performance of the deterioration model for each utility. "•" indicates that the parameter contributed, "o" indicates that the parameter was available but did not contribute to the performance and blank space indicates that the parameter was not available for the utility. To the right the performance of the model is shown. In the bottom of the table the total number of times that a parameter contributed to the performance and the number of times the parameter was the best, which means that it was the last remaining parameter in the backward step analysis, is shown. The table is adapted from [16].

Utility	Ground Level	Age	Groundwater	Wastewater	Length	Dimension	YoConst	YoRehab	Soil Type	Slope	Depth	NoBuildings	NoGrates	Material	DistRoad	Trees	Road Type	Rehab	City Type	Buildings Low	Buildings High	No. Features Relevant	Performance
Utility 14	•	•	o	o	o	•	•	o	o	o	•	o	•	o	•	o	o	o	o	o	o	7	0.78
Utility 1	•	•	•	o	•	•	o	o	•	o	o	o	•	•	o	o	o	o	o	o	o	8	0.71
Utility 24	•	•	•	o	o	•	o	o	•		•	o	•	•	o	o	o	o	o	o	o	8	0.76
Utility 6	•	•	•	•	o	o	•	o	o	o	•	o	o	o	o	•	o	o	o	o	o	7	0.71
Utility 16		•	•	•	•	o	•	•	o	o		•	o	o	•	o	o	•	o	o	o	9	0.71
Utility 12	•	o	•	•	•	•	o	•	•	o	o	•	o	o	o	o	o	o	•	o	o	9	0.69
Utility 28	•	•	•	o	•	•	•	o	o		o	o	o	o	o	o	o	o	o	o	o	6	0.77
Utility 5	•	•	•	•	o	•	•	o	•	•	o	•	o	o	o	o	•	o	o	o	o	10	0.78
Utility 29	•	o	•	•	•	•	•	•	•	•	o	o	•	•	o	•	o	o	o	o	o	12	0.76
Utility 3	•	o	•	o	o	•	•	o	o		o	o	o	o	o	o	o	o	o	o	o	4	0.77
Utility 9	•	•	•	•	•	o	•	o	o	o		o	o	o	o	o	o	o	o	o	o	6	0.77
Utility 2	o	•	•	•	o	o	o	•	o	o	o	o	o	o	o	o	o	o	o	o	o	4	0.82
Utility 25	•	•	•	o	o	•	o	•	o	•	o	o	o	o	o	o	o	o	o	o	o	6	0.77
Utility 18	•	•	o	•	•	•	o	o	•	o	•	o	o	•	o	o	•	o	o	o	o	9	0.79
Utility 27	o	o	•	•	•	o	o	•	o		o	o	o	o	o	o	•	o	o	o	o	5	0.66
Utility 19	o	o	•	o	•	o	•	•	o	o	o	•	•	o	o	•	o	o	o	o	o	7	0.71
Utility 4	•	•	o	•	o	•	o	o	•	•	o	o	•	o	o	•	o	o	o	o	o	8	0.65
Utility 20	•	•	o	•	•	•	•	•	o	•	o	o	o	o	o	o	o	o	o	o	o	8	0.76
Utility 21	•	•	o	•	•	•	•	o	o		o	o	o	o	o	o	o	o	o	o	o	6	0.73
Utility 30	•	•		o	•	•	o	o	o	o	o			o		•	o	o	o			5	0.70
Utility 11	o	•	•	•	•	o	•	o	o		•	o	o	o	o	o	•	o	o	o	o	7	0.69
Utility 31	•	•	•	o	•	o	o	o	•		o	o	o	•	•	o	o	o	o	o	o	7	0.77
Utility 22	•	•	•	o	•	•	o	o	o		•			o		o	o	o				6	0.84
Utility 32	o	o	•	•	o	o	o	•	o	o	o	o	o	o	o	o	o	o	o	o	o	3	0.71
Utility 8	•	o	o	•	o	•	o	•	•	o	o	•	o	o	•	o	o	o	o	o	o	7	0.74
Utility 23	•	•	•	•	•	o	o	o	•	•	•	o	o	o	o	o	o	•	o	o	o	9	0.80
Utility 17	o	o	o	o	o	o	•	o	o	o	o	o	o	o	o	o	o	o	o	o	o	1	0.81
Utility 33	•	•		•	o	o	o	•	o		•			o		o	o	o				5	0.56
Utility 7	o	•	•	•	•	o	o	•	•	•	o			•		o	o	o				8	0.71
Utility 26	o	•	o	o	o	o	o	o	•	o	o	o	o	o	o	o	o	•	o	o	o	3	0.75
Utility 15	o	o	o	•	o	•	•	o	o	o	o	o	o	o	•	o	o	o	•	o	o	5	0.80
Utility 13	o	o	o	•	o	o	•	•	o	o	•	•	o	o	o	o	o	o	o	o	o	5	0.86
Utility 10	•	o	o	o	•	o	•	o	o		o	o	o	o	o	o	•	o	o	o	o	4	0.85
Total	22	22	20	20	18	17	16	13	12	7	9	6	6	6	5	5	5	3	2	0	0		
Times best	0	8	6	0	0	1	13	5	0	0	0	0	0	0	0	0	0	0	0	0	0		

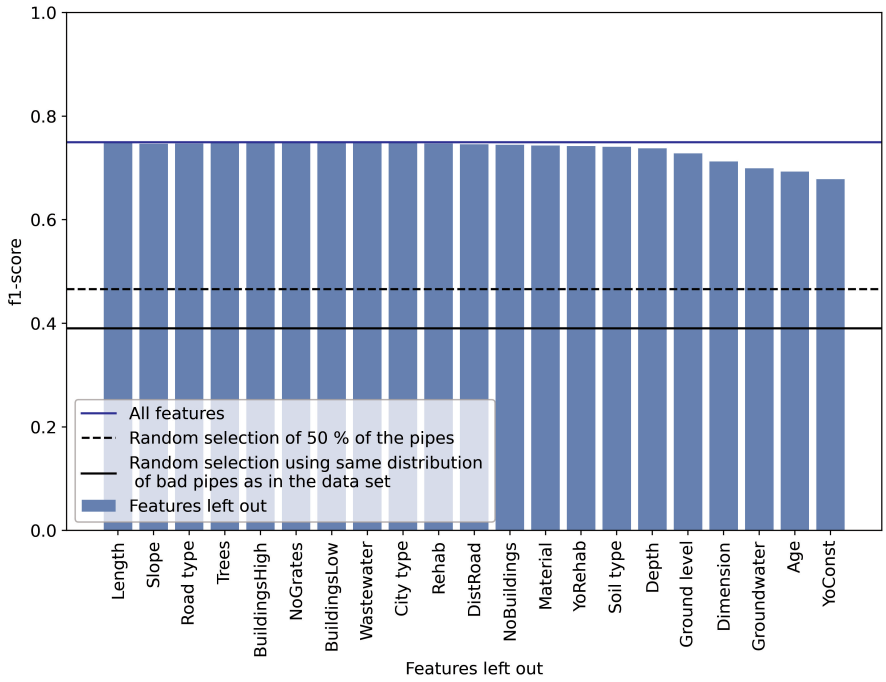


Fig. 3.8: Backward step analysis for the baseline without the geographical information. The charts show the performance of the model when a given feature and all the features to the right were used for training. The dashed line shows the performance when using a random selection of 50% of the pipes, and the solid line shows the performance when using a similar distribution of bad pipes as present in the dataset. The figure is adapted from Hansen et al. [16].

rehabilitation was the most important parameter. It might be very small factors which entails that one feature is removed before another in the step analysis because no or little additional information is added if more than one of the features is used. If only one of the three features was available, it might have contributed to the performance or be the most important feature for a larger number of the utilities. Another explanation for some of the variations could be local variations between the utilities and the data available for the utilities.

A comparison of how often the different features contributed in our analysis and in the review of Mohammadi et al. is presented in Table 3.7. Further details on the analysis can be found in Paper C.

Table 3.7: Comparison of how often a parameter contributed to the performance in our analysis, and how often it contributed in the studies reviewed by Mohammadi et al. [27].

Predictor Variables	Results		Mohammadi et al.	
	Times Present	Percent of Times Found Relevant	Times Present	Percent of Times Found Relevant
Ground level	27	78	-	-
Age	28	71	18	78
Groundwater	26	73	3	100
Wastewater	28	61	6	83
Length	28	57	11	91
Dimension	28	57	17	71
Year of construction	28	46	-	-
Year of rehabilitation	28	39	-	-
Soil type	28	36	5	20
Slope	19	32	12	42
Depth	26	31	16	44
No. buildings	25	20	-	-
No. grates	25	24	-	-
Material	28	21	15	67
Dist. to road center	25	16	-	-
Dist. to trees	28	18	-	-
Road types	28	14	5	40
Rehabilitation type	28	7	-	-
City type	26	4	-	-
Building low	25	0	-	-
Buildings high	25	0	-	-
Location	-	-	5	40
Up-invert	-	-	1	0
Down-invert	-	-	1	0
Bedding type	-	-	2	100
Corrosivity	-	-	2	50
Number of trees	-	-	5	60
Traffic	-	-	1	1
Flow	-	-	3	67
Hydrohalic	-	-	2	100
Location	-	-	5	40
Up-invert	-	-	1	0

3.3.3 Contributions

- The systematic feature analysis provided in this work gave insight regarding the number of features which contribute to the performance of deterioration models and which features are the most relevant to use. Thereby researchers and model developers can save time for feature selection and extraction.
- A challenge with the traditional inspection strategies has been elaborated and recommendations for future inspection strategies have been suggested.

3.4 Forecasting of pipe conditions

Forecasting pipe conditions is useful for long-term planning [30]. Input from water professionals working with EnviDan's Asset Management Tool suggested to make these forecasts for individual pipes. This work was carried out in 2019 and is based on the model presented in Section 3.1.2.

3.4.1 Existing approaches

For prediction of current pipe condition, machine learning models perform better than statistical models. However, for forecasting the pipe condition statistical models, Gompertz models for example, are better than machine learning models such as Random Forest [7]. This is because machine learning models are fully data-driven, and thereby the model can predict the condition to improve over time, whereas the statistical models are bound by statistical relations. However, more research is needed within machine learning models for future predictions [7].

An untouched issue regarding machine learning models for ad hoc predictions is that the data used for training and testing is historical data collected over several years, which means that making a prediction of the current condition state already implies usage of future predictions, as the test data rarely is up to date.

Based on this, the decision was made to investigate how the ageing curves of the model described in section 3.1.2 would look like if the age was set to 0-100 years. Since this work was conducted, in 2019 Balekelayi and Tesfamariam [5] proposed a Bayesian statistics based model for predicting the future conditions of individual pipes.

3.4.2 Machine learning based ageing curves

Taking basis in the model presented in Section 3.1.2, ageing curves were made for a number of pipes by changing the age of the pipe and plotting the curve according to the year it was installed and the observations performed. Examples of ageing curves can be seen in Figure 3.9.

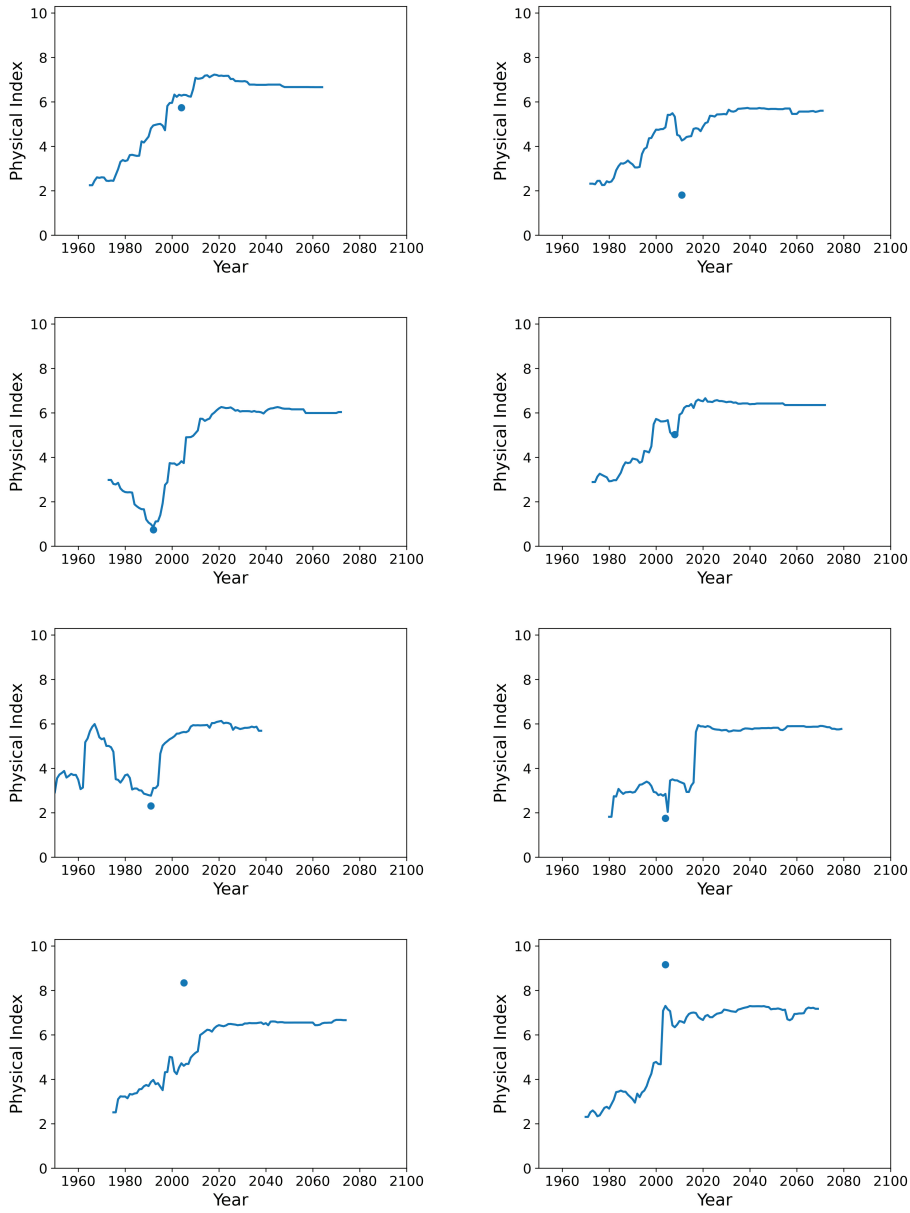


Fig. 3.9: Examples of machine learning based ageing curves for 100 years. The dot in the graphs indicates the observation of the pipe.

In Figure 3.9 it can be seen that the ageing curves are sensitive to the year of inspection and that the the condition is not bound to decrease over time. This underlines the findings in the literature which state that machine learning models are better for immediate predictions than for future predictions. However, it also shows a challenge in the current way of evaluating the deterioration models. Typically, the models are tested on a random selection of the inspected pipes. As the data have been collected over more than 20 years, a large part of the pipe inspections used for both training and testing are several years old. However, the models are applied to predict the current condition state of the pipes. To obtain a fair evaluation of the models they should be tested on data from the newest inspections.

3.4.3 Statistical ageing curves

The large adaptability to age at the time of inspection for the machine learning model entails that it might be beneficial to use statistical ageing curves. Statistical ageing curves for the different material types are shown in Figure 3.10. The curves were made by finding the average condition of the pipes in a given age.

As can be seen in the figure, there is a large uncertainty in the statistical curves. However, several observations can be made. Generally, there is a tendency that the curves bend after 40-60 years. This could indicate that the pipes in bad condition have been rehabilitated, for which reason only the pipes in a sufficiently good condition are left, as also described by Tscheikner-gratl et al. [30]. Similar patterns can be seen in the data for sewer inspection presented by Caradot et al. [7]. To reduce the impact of this, it is suggested that ageing curves are made by fitting a function to pipes younger than 40-60 years. Another observation is the large number of clay pipes which are zero years. This is most likely due to registration errors in the databases as clay pipes were outdated at the time where CCTV inspection robots existed. Likewise for the plastic pipes, it might be beneficial to consider only younger pipes, as recent plastic pipes may have a better quality than the early plastic pipes, and plastic pipes above 60 years could be registration errors.

3.4.4 Contributions

- It was shown that Random Forest is very sensitive to the age at the time of inspection.
- Recommendations for how to test model performance were made to minimize the difference between the development environment and the production environment.
- Recommendations for how to make ageing curves were made and included in EnviDan's asset management software.

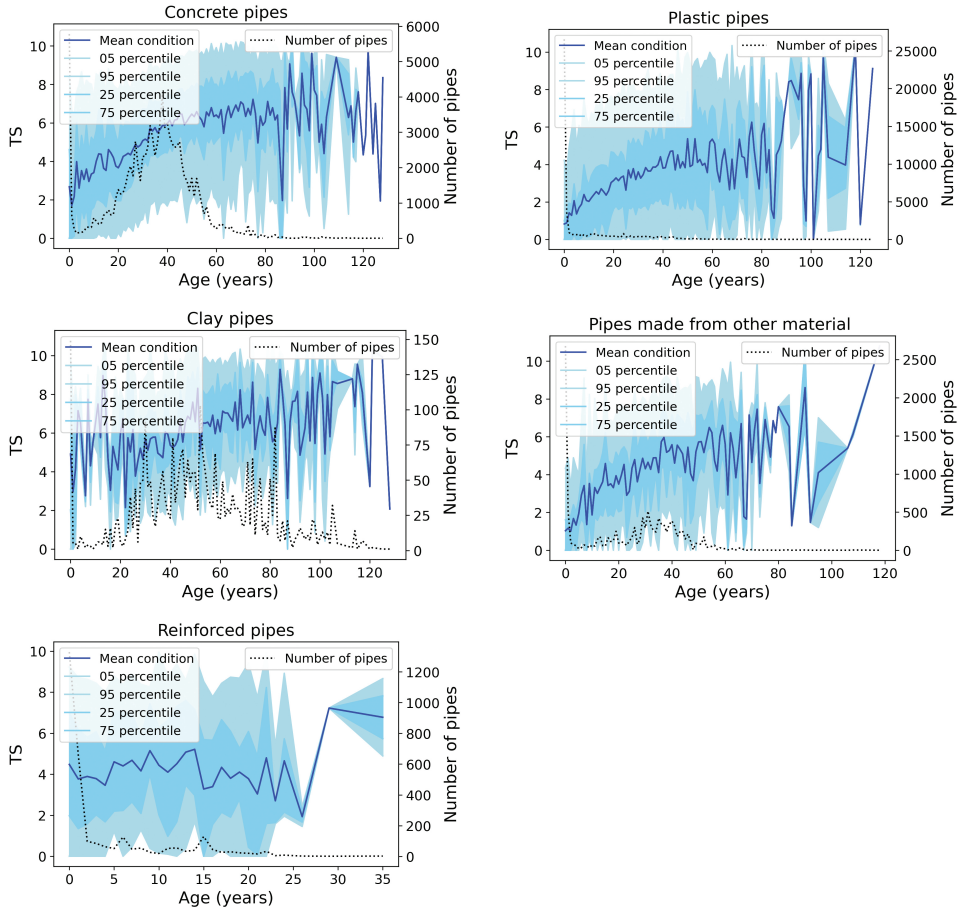


Fig. 3.10: Statistical ageing curves for pipes made from concrete, plastic, clay, other materials, and pipes reinforced with lining. Furthermore, the number of pipes with a given age is shown.

References

[1] A. Altarabsheh, M. Ventresca, and A. Kandil, "New approach for critical pipe prioritization in wastewater asset management planning," vol. 32, no. 5, p. 04018044. [Online]. Available: http://ascelibrary.org/doi/10.1061/%28ASCE%29CP.1943-5487.0000784

[2] A. G. Altarabsheh, "Managing urban wastewater system using complex adaptive system approach," Ph.D. dissertation, PURDUE UNIVERSITY GRADUATE SCHOOL, 2015.

[3] E. V. Ana and W. Bauwens, "Modeling the structural deterioration of urban drainage pipes: the state-of-the-art in statistical methods," vol. 7, no. 1, pp. 47–59. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/15730620903447597

[4] K. Baah, B. Dubey, R. Harvey, and E. McBean, "A risk-based approach to sanitary sewer pipe asset management," *Science of The Total Environment*, vol. 505, pp. 1011–1017, feb 2015.

[5] N. Balekelayi and S. Tesfamariam, "Statistical inference of sewer pipe deterioration using bayesian geoadditive regression model," vol. 25, no. 3, p. 04019021. [Online]. Available: http://ascelibrary.org/doi/10.1061/%28ASCE%29IS.1943-555X.0000500

[6] J. Brock, M. Lange, J. A. Tratalos, S. J. More, D. A. Graham, M. Guelbenzu-Gonzalo, and H.-H. Thulke, "Combining expert knowledge and machine-learning to classify herd types in livestock systems," *Scientific Reports*, vol. 11, no. 1, p. 2989, Dec. 2021. [Online]. Available: http://www.nature.com/articles/s41598-021-82373-3

[7] N. Caradot, M. Riechel, M. Fesneau, N. Hernandez, A. Torres, H. Sonnenberg, E. Eckert, N. Lengemann, J. Waschnewski, and P. Rouault, "Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in berlin, germany," vol. 20, no. 5, pp. 1131–1147. [Online]. Available: https://iwaponline.com/jh/article/20/5/1131/40825/Practical-benchmarking-of-statistical-and-machine

[8] N. Caradot, H. Sonnenberg, I. Kropp, A. Ringe, S. Denhez, A. Hartmann, and P. Rouault, "The relevance of sewer deterioration modelling to support asset management strategies," *Urban Water Journal*, vol. 14, no. 10, pp. 1007–1015, may 2017.

[9] G. Carvalho, C. Amado, R. S. Brito, S. T. Coelho, and J. P. Leitão, "Analysing the importance of variables for sewer failure prediction," vol. 15, no. 4, pp. 338–345. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1573062X.2018.1459748

[10] J. P. Davies, B. Clarke, J. Whiter, R. Cunningham, and A. Leidi, "The structural condition of rigid sewer pipes : a statistical investigation," vol. 3, pp. 277–286.

[11] M. Elmasry, A. Hawari, and T. Zayed, "Defect based deterioration model for sewer pipelines using bayesian belief networks," vol. 44, no. 9, pp. 675–690. [Online]. Available: http://www.nrcresearchpress.com/doi/10.1139/cjce-2016-0592

[12] ——, "An economic loss model for failure of sewer pipelines," vol. 14, no. 10, pp. 1312–1323. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15732479.2018.1433693

[13] M. Elmasry, T. Zayed, and A. Hawari, "Multi-objective optimization model for inspection scheduling of sewer pipelines," vol. 145, no. 2, p. 04018129. [Online]. Available: http://ascelibrary.org/doi/10.1061/%28ASCE%29CO.1943-7862.0001599

[14] S. M. Ghavami, Z. Borzooei, and J. Maleki, "An effective approach for assessing risk of failure in urban sewer pipelines using a combination of GIS and AHP-DEA," vol. 133, pp. 275–285. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957582019313473

[15] B. D. Hansen, S. H. Rasmussen, T. B. Moeslund, M. Uggerby, and D. G. Jensen, "Sewer deterioration modeling: The effect of training a random forest model on logically selected data-groups," vol. 176, pp. 291–299. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S187705092031855X

[16] B. D. Hansen, S. H. Rasmussen, M. Uggerby, T. B. Moeslund, and D. G. Jensen, "Comprehensive feature analysis for sewer deterioration modeling," vol. 13, no. 6, p. 819. [Online]. Available: https://www.mdpi.com/2073-4441/13/6/819

[17] B. D. Hansen, D. Getreuer Jensen, S. H. Rasmussen, J. Tamouk, M. Uggerby, and T. B. Moeslund, "General sewer deterioration model using random forest," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 834–841. [Online]. Available: https://ieeexplore.ieee.org/document/9002727/

[18] R. R. Harvey and E. A. McBean, "Predicting the structural condition of individual sanitary sewer pipes with random forests," vol. 41, no. 4, pp. 294–303. [Online]. Available: http://www.nrcresearchpress.com/doi/10.1139/cjce-2013-0431

[19] A. Hawari, F. Alkadour, M. Elmasry, and T. Zayed, "A state of the art review on condition assessment models developed for sewer pipelines," vol. 93, p. 103721. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0952197620301391

[20] ——, "Simulation-based condition assessment model for sewer pipelines," *Journal of Performance of Constructed Facilities*, vol. 31, no. 1, p. 04016066, feb 2017.

[21] N. Hernández, N. Caradot, H. Sonnenberg, P. Rouault, and A. Torres, "Optimizing SVM models as predicting tools for sewer pipes conditions in the two main cities in colombia for different sewer asset management purposes," *Structure and Infrastructure Engineering*, vol. 17, no. 2, pp. 156–169, Mar. 2020. [Online]. Available: https://doi.org/10.1080/15732479.2020.1733029

[22] G. Kabir, N. B. C. Balek, and S. Tesfamariam, "Sewer structural condition prediction integrating bayesian model averaging with logistic regression," vol. 32, no. 3, p. 04018019. [Online]. Available: http://ascelibrary.org/doi/10.1061/%28ASCE%29CF.1943-5509.0001162

[23] T. Laakso, T. Kokkonen, I. Mellin, and R. Vahala, "Sewer condition prediction and analysis of explanatory factors," vol. 10, no. 9, p. 1239. [Online]. Available: http://www.mdpi.com/2073-4441/10/9/1239

[24] J. Lee, C. Y. Park, S. Baek, S. H. Han, and S. Yun, "Risk-based prioritization of sewer pipe inspection from infrastructure asset management perspective," *Sustainability*, vol. 13, no. 13, p. 7213, Jun. 2021. [Online]. Available: https://doi.org/10.3390/su13137213

[25] P. Lin, X. Yuan, and E. Tovilla, “Integrative modeling of performance deterioration and maintenance effectiveness for infrastructure assets with missing condition data,” vol. 34, no. 8, pp. 677–695. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12452

[26] Malek Mohammadi, Najafi, Kaushal, Serajiantehrani, Salehabadi, and Ashoori, “Sewer pipes condition prediction models: A state-of-the-art review,” vol. 4, no. 4, p. 64. [Online]. Available: https://www.mdpi.com/2412-3811/4/4/64

[27] M. Malek Mohammadi, M. Najafi, S. Kermanshachi, V. Kaushal, and R. Serajiantehrani, “Factors influencing the condition of sewer pipes: State-of-the-art review,” vol. 11, no. 4, p. 03120002. [Online]. Available: http://ascelibrary.org/doi/10.1061/%28ASCE%29PS.1949-1204.0000483

[28] D. Marlow, L. Pearson, D. H. MacDonald, S. Whitten, and S. Burn, “A framework for considering externalities in urban water asset management,” vol. 64, no. 11, pp. 2199–2206. [Online]. Available: https://iwaponline.com/wst/article/64/11/2199/17133/A-framework-for-considering-externalities-in-urban

[29] M. M. Rokstad and R. M. Ugarelli, “Evaluating the role of deterioration models for condition assessment of sewers,” vol. 17, no. 5, pp. 789–804. [Online]. Available: https://iwaponline.com/jh/article/17/5/789-804/3510

[30] F. Tscheikner-Gratl, N. Caradot, F. Cherqui, J. P. Leitão, M. Ahmadi, J. G. Langeveld, Y. Le Gat, L. Scholten, B. Roghani, J. P. Rodríguez, M. Lepot, B. Stegeman, A. Heinrichsen, I. Kropp, K. Kerres, M. d. C. Almeida, P. M. Bach, M. Moy de Vitry, A. Sá Marques, N. E. Simões, P. Rouault, N. Hernandez, A. Torres, C. Werey, B. Rulleau, and F. Clemens, “Sewer asset management – state of the art and research needs,” vol. 16, no. 9, pp. 662–675. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1573062X.2020.1713382

[31] F. Tscheikner-Gratl, R. Sitzenfrei, W. Rauch, and M. Kleidorfer, “Integrated rehabilitation planning of urban infrastructure systems using a street section priority model,” *Urban Water Journal*, vol. 13, no. 1, pp. 28–40, jul 2015.

[32] X. Yin, Y. Chen, A. Bouferguene, and M. Al-Hussein, “Data-driven bi-level sewer pipe deterioration model: Design and analysis,” vol. 116, p. 103181. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0926580519312713

Chapter 4

Forecasting of Methane Yield from Biogas Plants

Increasing energy demands and the United Nations' (UN) goal of keeping the global temperature increase below 1.5 degrees results in more focus on renewable energy. The biogas industry produces energy from biological waste products and thereby reduces emissions and pollution from the suppliers of the biogas plants. In the period from 2007-2017 the biogas industry was almost tripled in Europe [18].

In biogas plants, organic materials such as industrial waste, sludge, food, lignocellulosic materials, and animal manure are transformed to biogas by anaerobic digestion [6].

The digestion time of the feedstocks depends on the feedstock's composition. Carbohydrates, fat and protein are easy to digest whereas lignocellulosic materials are harder to digest [6].

The digestion process is affected by several additional parameters such as organic loading rate, pH, temperature, carbon/nutrient rate, the present enzymes and microorganisms, added substrates and the type of fatty acids in the digester. The system is very sensitive to over- and underrepresentation of some of the parameters, which, in some cases, can lead to failure of the system [10, 16].

Co-digestion is often preferable over mono-digestion of the feedstocks as mono-digestion of some feedstocks can bring the system out of balance. Furthermore, research has shown that co-digestion can increase the performance by 25-400 % [10], though co-digestion complicates the digestion process even more. Additionally, different plants have different goals regarding the production. In some cases the goal is to produce as much methane as possible, while the goal for other plants is to obtain as stable a production as possible and meet the market demand while avoiding overproduction. Overproduction can be handled by utilization of inhibiting chemicals or by burning surplus biogas, but neither of the solutions are optimal [12].

Mathematical models can give essential knowledge to keep the balance in the plant and thereby avoid failures [10]. However, for plants aiming for a stable production, there is a need for optimization of the models used for forecasting the methane yield. Furthermore, having precise models is an important step towards automatic control of the feed to the plants.

4.1 Existing approaches

The IWA Anaerobic Digestion Model No. 1 (ADM1), which models the whole process at the plant, was published in 2002 [3]. This model can simulate general tendencies for the different parameters in the plant but lacks the capability of simulating immediate variations [9]. After the publication of ADM1, it was enhanced by several plugins. However, in 2015 Batstone et al. [4] justified that an ADM2 could be developed. Further research is required to obtain a uniform procedure for handling the mechanisms and challenges [4]. After 2015, the ADM1 has been optimized further and other models have been developed [2].

Another model type often used for forecasting the biogas production is the Gompertz model as this model type can give accurate estimates of the methane production [14]. Contrary to the ADM1 model, this model type only focuses on forecasting the biogas production. Similarly to the ADM1 model it needs precise calibration.

When using a Gompertz model, a Gompertz Function needs to be made for each feedstock. Several studies have focused on experimentally finding the kinetic parameters which best describe the methane production for a wide range of feedstocks [13–15]. If the kinematic parameters are not available in the literature, they can be found experimentally, though this often takes several months for which reason the parameters are typically based on expert knowledge. Even if the kinematic parameters are available, local variations in the composition of the feedstocks, co-digestion, and other parameters affecting the digestion entail that the parameters found in experimental setups might not reflect real case scenarios [12].

To optimise and forecast the methane production, machine learning models can be used [12], and there is a rising trend in usage of machine learning for prediction of biogas production [7]. Cruz et al. [7] reviewed 32 papers on machine learning and anaerobic digestion which were published in the period from 2010 - 2021. The most popular method was Artificial Neural Network (33 %) followed by Hybrid Models, which is a model that consists of at least two different models (15 %), Adaptive Network-based Fuzzy Inference System (ANFIS) (10 %), Support Vector Machine (8 %), Random Forest (RF) (8 %), Genetic algorithm (8 %), and Particle Swam Optimization (4 %). The remaining 14 % encountered either Deep Learning, Extreme Gradient Boosting (XGBoost), K-Nearest Neighbours (KNN), Ant Colony Optimization, Multi-class Logistic Regression (MLR), Ensamble of Neural Networks, or Logistic Regression [7].

The performance obtained in different studies is hard to compare, as there is a huge variation in the different cases which varies from biogas plants at wastewater treatment facilities [1] to agricultural plants [5], industrial-scale co-digestion facilities [8], and laboratory scale experiments [17]. Some authors forecast the methane production using daily inputs while others forecast the biogas production for a set of feedstocks [7]. Here it is worth noting that the phrase *biogas* covers both methane, CO_2, and other gasses, while methane is the desired outcome.

The large variations in the different solutions entail that methods need to be compared to other methods using the same dataset to ensure a fair evaluation. Cruz et al. [7] reviewed seven comparative studies which forecast either the biogas production or the methane production, but there was no consensus regarding which method was the best. RF, XGBoost, and KNN were all the best or one of the best in two cases, while recurrent neural networks, MLR, and ANFIS were the best or one of the best methods in one of the cases [7]. We have previously shown that the performance of RF is sensitive to the amount of training data, and if the training data are limited there can be high standard deviations of the performance [11]. This is most likely also the case for other learning algorithms.

One challenge with machine learning models is that they are often considered as black boxes, whereas the statistical models are much easier to interpret. Furthermore the machine learning models require a sufficiently large historical dataset for the biogas plant, which is not the case for, e.g., a Gompertz model.

Generally there are pros and cons with both expert based models and machine learning models. However, there is a lack of studies which combine the models to obtain more precise and intelligible models.

4.2 Model for forecasting the methane yield

This section is a presentation of the work documented in Paper D.

Three models for biogas forecasting the methane yield in an industrial biogas plant were made: a Gompertz model, a machine learning model, and a hybrid model.

The biogas plant's capacity was approximately 220,000 ton biomass per year with a yield of 10 mio. Normal cubic Meter (NM) methane per year. Consecutive daily measures of the gas production and information of used feedstocks were available for 818 days.

The main feedstocks were seaweed, manure, pectin, and eulat. In total, 18 different feedstocks were used, though the used feedstocks changed over time depending on availability of the different waste products.

4.2.1 Gompertz model

For each feedstock a Gompertz function was set up. The functions were based on empirical data, if available. If empirical information on the feedstock was not available, experts within biogas production set up the functions based on empirical data from similar feedstocks. A plot of the Gompertz functions can be seen in Figure 4.1.

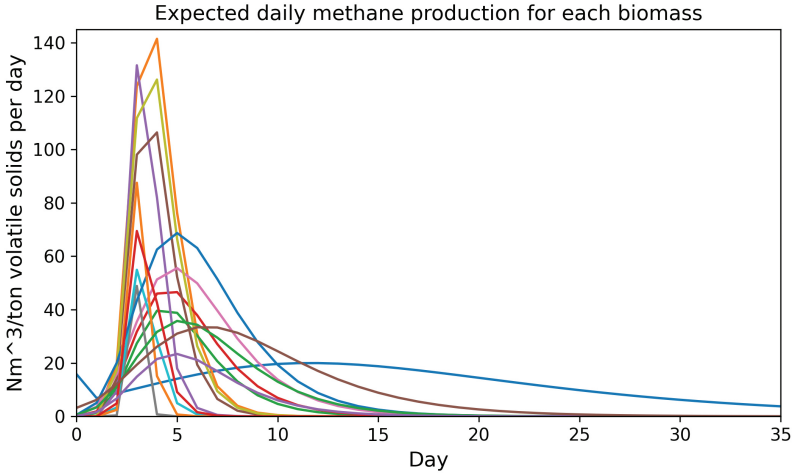


Fig. 4.1: Gompertz functions for the different feedstocks. Due to proprietary information it is not possible to provide the names of the feedstocks in the figure. The figure is slightly modified from Hansen et al. [12].

4.2.2 Machine learning model

The machine learning model was trained to forecast the biogas yield one day ahead. For development of the machine learning model, all data points were normalized, and features such the methane production from the previous six days and mean and standard deviation of the methane production from the 9 previous days were found for each data point.

Based on initial test of 15 machine learning algorithms, the seven best performing algorithms were used for model development. These were uniform kNN, distance kNN, multi-layer perception (MLP), Random Forest, recursive feature elimination with linear ridge, recursive feature elimination with gradient boosting, and AdaBoost with decision trees.

The first 430 data points were used for training the model. The data were divided into 25 folds of which 23 were used for training, one was used for validation and the last one was used for testing.

For each fold, an ensemble of the three best models was made and compared to the best of the models. If the ensemble model was better than the best of the models and if the performance exceeded a threshold, it was saved. Otherwise, if the best model was better than the ensemble and exceeded the threshold this model was saved.

For forecasting the methane yield on the test set, the mean prediction of the saved models was used.

4.2.3 Hybrid model

The hybrid model consisted of the Gompertz model and a machine learning model. The machine learning model used in the hybrid model was trained in the same way as the machine learning model used alone, but instead of predicting the methane production it was trained to predict the error of the Gompertz model.

4.2.4 Results

The last 350 data points were used for testing the different models. The results can be seen in Figure 4.2. A bias in the forecasts was observed. The Mean Absolute Percentage Error (MAPE) before and after accounting for the bias can be seen in Table 4.1.

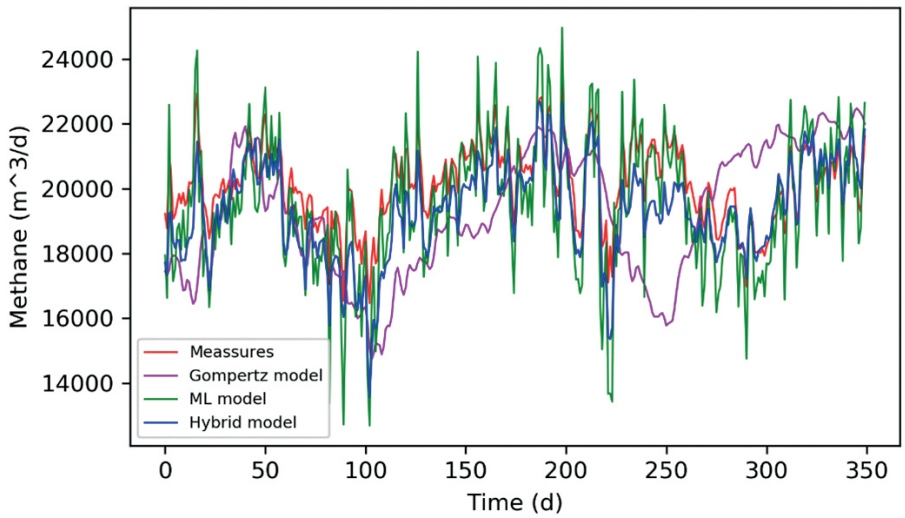


Fig. 4.2: Plot of the measured methane production, Gompertz predictions, machine learning predictions and hybrid predictions.

The results showed that the hybrid model performed better than both the Gompertz model and a machine learning model. In fact, the hybrid model reduced the error of

Table 4.1: Mean absolute percentage error (MAPE) for the different methods before and after bias adjustment. The table is modified form Hansen et al. [12].

Model	↓ MAPE	↓ MAPE after adjustment
Gompertz model	9.61 %	9.00 %
ML model	4.84 %	3.78 %
Hybrid model	**4.52 %**	**3.06 %**

the Gompertz model by 53 % when not compensating for the bias and by 66 % when compensating for the bias. In this case the machine learning model and the hybrid model were trained to forecast the production one day ahead. In the future this could be extended to multiple days.

4.3 Contributions

- It was shown that a hybrid model consisting of a Gompertz model and a machine learning model could forecast the methane production one day ahead better than a Gompertz model and a machine learning model individually. This indicates that there is a potential for hybrid models within forecasting of methane production, and further research within the subject is recommended.

References

[1] H. Akbaş, B. Bilgen, and A. M. Turhan, "An integrated prediction and optimization model of biogas production system at a wastewater treatment facility," vol. 196, pp. 566–576. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960852415011219

[2] J. A. Arzate, M. Kirstein, F. C. Ertem, E. Kielhorn, H. Ramirez Malule, P. Neubauer, M. N. Cruz-Bournazou, and S. Junne, "Anaerobic digestion model (AM2) for the description of biogas processes at dynamic feedstock loading rates," vol. 89, no. 5, pp. 686–695. [Online]. Available: http://doi.wiley.com/10.1002/cite.201600176

[3] D. J. Batstone, J. Keller, I. Angelidaki, S. V. Kalyuzhnyi, S. G. Pavlostathis, A. Rozzi, W. T. M. Sanders, H. Siegrist, and V. A. Vavilin, "The IWA anaerobic digestion model no 1 (ADM1)," vol. 45, no. 10, pp. 65–73.

[4] D. J. Batstone, D. Puyol, X. Flores-Alsina, and J. Rodríguez, "Mathematical modelling of anaerobic digestion processes: applications and future needs," vol. 14, no. 4, pp. 595–613. [Online]. Available: http://link.springer.com/10.1007/s11157-015-9376-4

[5] T. Beltramo, M. Klocke, and B. Hitzmann, "Prediction of the biogas production using GA and ACO input features selection method for ANN model," vol. 6, no. 3, pp. 349–356. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214317318302270

[6] L. M. Colla, A. C. F. Margarites, A. Decesaro, F. G. Magro, N. Kreling, A. Rempel, and T. S. Machado, "Waste biomass and blended bioresources in biogas production," in *Biofuel and Biorefinery Technologies*. Springer International Publishing, 2019, pp. 1–23. [Online]. Available: https://doi.org/10.1007/978-3-030-10516-7_1

[7] I. A. Cruz, W. Chuenchart, F. Long, K. Surendra, L. R. S. Andrade, M. Bilal, H. Liu, R. T. Figueiredo, S. K. Khanal, and L. F. R. Ferreira, "Application of machine learning in anaerobic digestion: Perspectives and challenges," *Bioresource Technology*, vol. 345, p. 126433, Feb. 2022. [Online]. Available: https://doi.org/10.1016/j.biortech.2021.126433

[8] D. De Clercq, Z. Wen, F. Fei, L. Caicedo, K. Yuan, and R. Shang, "Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion," vol. 712, p. 134574. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0048969719345656

[9] K. Derbal, M. Bencheikh-lehocine, F. Cecchi, A.-H. Meniai, and P. Pavan, "Application of the IWA ADM1 model to simulate anaerobic co-digestion of organic waste with waste activated sludge in mesophilic condition," vol. 100, no. 4, pp. 1539–1543. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960852408006366

[10] K. Hagos, J. Zong, D. Li, C. Liu, and X. Lu, "Anaerobic co-digestion process for biogas production: Progress, challenges and perspectives," vol. 76, pp. 1485–1496. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032116309388

[11] B. D. Hansen, S. H. Rasmussen, T. B. Moeslund, M. Uggerby, and D. G. Jensen, "Sewer deterioration modeling: The effect of training a random forest model on logically selected data-groups," vol. 176, pp. 291–299. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S187705092031855X

[12] B. D. Hansen, J. Tamouk, C. A. Tidmarsh, R. Johansen, T. B. Moeslund, and D. G. Jensen, "Prediction of the methane production in biogas plants using a combined gompertz and machine learning model," in *Computational Science and Its Applications – ICCSA 2020*. Springer International Publishing, 2020, pp. 734–745. [Online]. Available: https://doi.org/10.1007/978-3-030-58799-4_53

[13] V. C. Hernández-Fydrych, G. Benítez-Olivares, M. A. Meraz-Rodríguez, M. L. Salazar-Peláez, and M. C. Fajardo-Ortiz, "Methane production kinetics of pretreated slaughterhouse wastewater," vol. 130, p. 105385. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0961953419303344

[14] V. Ripoll, C. Agabo-García, M. Perez, and R. Solera, "Improvement of biomethane potential of sewage sludge anaerobic co-digestion by addition of "sherry-wine" distillery wastewater," vol. 251, p. 119667. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652619345378

[15] L. A. d. Santos, R. B. Valença, L. C. S. d. Silva, S. H. d. B. Holanda, A. F. V. d. Silva, J. F. T. Jucá, and A. F. M. S. Santos, "Methane generation potential through anaerobic digestion of fruit waste," vol. 256, p. 120389. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652620304364

[16] T. Scapini, A. F. Camargo, F. S. Stefanski, N. Klanovicz, R. Pollon, J. Zanivan, G. Fongaro, and H. Treichel, "Enzyme-mediated enhanced biogas yield," in *Biofuel and*

Biorefinery Technologies. Springer International Publishing, 2019, pp. 45–68. [Online]. Available: https://doi.org/10.1007/978-3-030-10516-7_3

[17] F. Tufaner and Y. Demirci, "Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models," vol. 22, no. 3, pp. 713–724. [Online]. Available: http://link.springer.com/10.1007/s10098-020-01816-z

[18] S. Xue, J. Song, X. Wang, Z. Shang, C. Sheng, C. Li, Y. Zhu, and J. Liu, "A systematic comparison of biogas development and related policies between china and europe and corresponding insights," vol. 117, p. 109474. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032119306823

Chapter 5

Fault Detection in Pumps

Intruding water in the sewer system is a problem for utilities as it reduces the capacity for sewage and rain water, which can led to flooding. Furthermore, intruding water is lead to the wastewater treatment plant (WWTP) and treated like sewage water. This is problematic because the intruding water takes capacity at the WWTP, and its treatment process is energy-intensive. To repair the sewers, one must identify where the intruding water is coming from. Identification of damaged sewers is typically done by CCTV inspection, but, as this is labor heavy, other methods need to be utilized to identify where the intruding water enters the system.

The amount of intruding water can be estimated if the flow in the sewer is known. The flow can be found using sensors; however, this is an expensive solution. A much cheaper solution is to use the existing pumps in the system. Approximately three-fourths of the pumps in the Danish sewage systems do not measure the flow. Therefore, to include as many pumping stations as possible, the flow can be estimated using the energy usage [5, 8]. The flow at each pumping station can subsequently be divided into rainwater, sewage, and drainage water and intruding water.

However, for many utilities it is sufficient to know how much energy is used to pump sewage, rainwater, and intruding water and drain water, respectively.

In Denmark, the hourly energy usage of pumping stations is easily accessible and systematized for all the Danish utilities. This means several utilities can benefit from a previously described solution, which is also one of EnviDan's products. A challenge observed in the product is that the data is often faulty. Large errors such as longer periods of flatlines can easily be detected, but for smaller errors a Multilayer Perception (MLP) model is used to replace untrustworthy predictions. This MLP model is trained to predict the energy usage of each of the utilities' pumping stations based on the energy usage from the remaining stations. This can be done because the pumping stations are positioned in the same utility, and thereby it is expected that they generally are exposed

to similar weather patterns. However, the MLP model has not been tested thoroughly. Furthermore, it was found relevant to investigate if other methods would be better for the purpose, which was the purpose of this work.

5.1 Existing approaches

Faults in data is a common problem and the reason for which anomaly and fault detection have been hot topics for decades, with several surveys having been published during the last two decades [12]. The techniques used include supervised learning, where labeled data is available, semi-supervised learning, one-class learning, if only labeled data is available for the normal class, and unsupervised, if no labeled data is available [3].

In addition to fault detection, machine learning has been used for maintenance management such as failure mode analysis, condition monitoring, and downtime minimization in several industrial applications. For failure mode analysis, neural network is the most used approach to determine the cause of failures on machines and equipment [2].

Especially within energy usage in buildings, anomaly detection has been popular. Himeur et al. [7] reviewed several methods which have been applied to this. An overview of the methods is presented in Table 5.1. The review showed that the unsupervised approaches are easy to apply as they do not require annotated datasets, though these approaches are only able to detect excessive energy consumption. Generally, using Convolutional Neural Networks (CNN) for supervised anomaly detection has proven efficient, however, they require an annotated dataset [7].

Fan et al. [6] investigated the potential of Autoencoders for anomaly detection in energy data from buildings. Due to the challenge with missing information about the anomalies, Fan et al. used the premise that generally 5 % of the data would be disrupted and concluded that their Autoencoder could successfully detect anomalies [6].

Since 2015 there has been a significant increase in the number of publications using deep learning for machine fault diagnosis. In this period there was also an increase in publications using traditional machine learning. However, the increase in publications on traditional machine learning applications are less significant than the increase in deep learning applications, and in 2017 more papers were published using deep learning than traditional machine learning techniques [9]. Azadeh et al. [1] used SVM and neural network for condition monitoring of centrifugal pumps. Zhou et al. proposed a combination of a noise adaptive Kalman Filter and a Neural Network for fault detection in oil pumps [11].

Several deep learning methods have been applied to hydraulic systems for fault detection. The methods used include Stacked Autoencoders, Deep Belief Network, CNNs, Recurent Neural Network, and Generative Adversarial Network [4].

As previously described, deep learning methods have generally been shown efficient in hydraulic systems, and CNNs have been shown efficient for supervised anomaly detection in energy consumption in buildings. The data available for this work is unlabeled, and

Table 5.1: Overview of methods used for anomaly detection in energy consumption in buildings. The table is based on Himeur et al. [7].

Supervised Detection	**Unsupervised Learning**	**Ensemble**
Neural Networks	*Clustering*	*Boosting*
- Deep Autoencoder	- k-Means	- Adaboost
- Convolutional Neural Network	- c-Means	- Gradient Boosting Machine
- Recurrent Neural Network	- Entropy-based	- Gradient Tree Boosting
- Deep Delief Network	- Mutual kNN	*Bagging*
- Generative Adversarial Network	*One-class learning*	- Bootstrap Aggregation
- Extreme Learning Machines	- One-class Support Vector Machine	- Multiveiw Stack Ensemble
- Multilayer Perception	- One-class Neural Network	- Random Forest
- Radial Basis Function Neural Network	- One-class Convolutional Neural Network	- Feature Bagging
Traditional Classification	- One-class Random Forest	**Feature Extraction**
- k-Nearest Neighbour	*Dimensionality Reduction*	*Distance-based*
- Support Vector Machine	- Principal Component Analysis	- Distance-based Outlier Detection
- Decision Trees	- Linear Discriminant Analysis	- Resolution-based Outlier Factor
- Logistic Regression	- Quadratic Discriminant Analysis	- Isolated Forest
Regression	- Multiple Discriminant Analysis	*Time-series Analysis*
- Support Vector Regression		- Short-term Time-series
- Autoregressive	**Hybrid Learning**	- Rule-based
- Autoregressive Integrated Moving Average	- Semi-Support Vector Machine	*Density-based*
Probabilistic Models	- DAE-kNNG[1]	- Density-based Spatial Clustering
- Bayesian Networks		- Local Outlier Factor
- Naive Bayes	**Other Techniques**	- Local Density Cluster-based Outlier Factor
- Statistical Models	- Compressive Sensing	*Graph-based*
	- Visualization	- Parallel Graph-based Outlier Detection
		- Graph-based Abnormaly Detection

[1]Deep Autoencoder k-Nearest Neighbour Graphs

thereby unsupervised learning and one-class learning can be applied. For this reason, the decision was made to investigate if a CNN Autoencoder could obtain better performance than a MLP model.

5.2 Fault detection using CNNs

In the work presented in this section it was investigated if CNNs could be used to optimize fault detection in pumping stations compared to MLP.

5.2.1 Data

For this study, the hourly energy consumption for 161 pumping stations operated by a Danish utility were available. 56 of the pumping stations were excluded due to fac-

tors such as longer periods with missing data, abnormal operation, and low energy usage. The remaining 105 stations included both start/stop pumps, frequency modulated pumps and alternating pumps. A challenge with the start/stop pumps is that if the flow to the pumps is low, the pumps will only be activated a few times an hour and, in some cases, remain inactive. To compensate for this, a moving average filter or a sum of the energy usage per day could be considered, though this would entail that a large part of the signal would be lost. For this reason, this work solely looks at pumping stations with frequency modulated pumps.

No information on the pump type was available in the achieved data. Therefore, to distinguish between frequency modulated pumps and start/stop pumps, four features were found. The features were the max amplitude, mean amplitude, ratio between high and low frequencies and the standard deviation of the normalized data. These features were subsequently normalized, and Principal Component Analysis (PCA) was performed. To investigate if the different pump types could be clustered, plots of the energy usage for each of the pumping stations were manually inspected for several different time slots. The reason that one time slot was not sufficient was that the patterns of start/stop pumps are similar to frequency modulated pumps during periods with high load. Five frequency modulated pumps were identified. Figure 5.1 shows a plot of the first two principal components, and whether the different pumps were categorized as start/stop pumps of frequency modulated pumps.

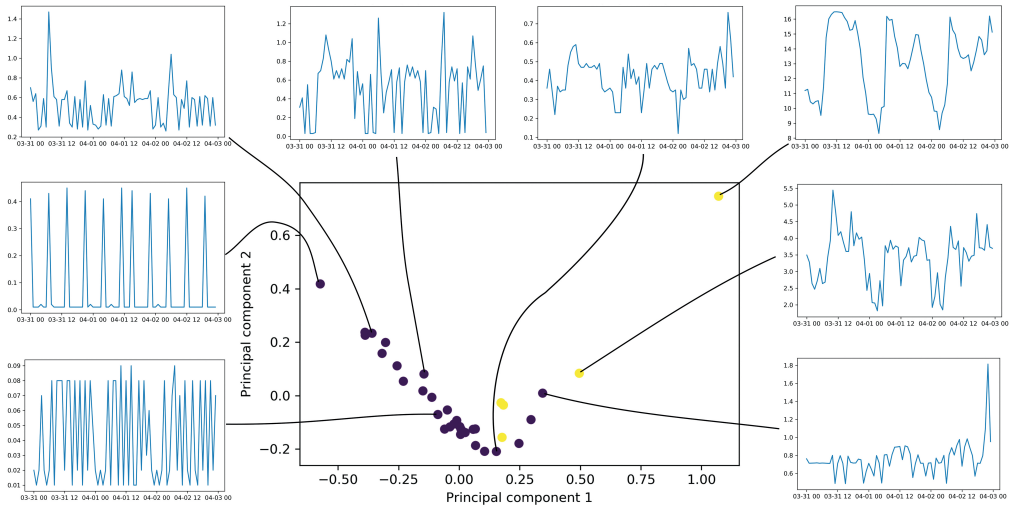


Fig. 5.1: The pumps plotted according to principal component one and two. A yellow dot indicates that the pump was used for the analysis, and a blue dot indicates that the pump was excluded. Examples of three days of data are provided for eight of the pumps.

From Figure 5.1 it can be seen that the frequency modulated pumps and some of the start/stop pumps are distinguishable from the majority of the start/stop pumps. The reason that some of the start/stop pumps are positioned among the frequency modulated pumps is that if a start/stop pump has a high load, the hourly signal would be more similar to the hourly signal from frequency modulated pumps, as the stochastic part of the energy usage is neglected when the pump is active full-time or turned on and off several times per hour. From a practical perspective, it would be fine to include these pumps together with the remaining pumps for the analysis. However, for this work only the frequency modulated pumps were considered. This is because start/stop pumps are turned on and off several times, and it is much more energy consuming to turn a pump on than to keep it running. For instance, if there is a constant large flow and the pump in running constantly, it can use less energy than if the flow is low enough for the pump to turn on and off several times. Therefore, there are some non-linearities in the energy usage of start/stop pumps compared to frequency modulated pumps.

5.2.2 Preprocessing

Three ways of normalizing the signals were tested. First, the signal was normalized according to min and max values, but due to peaks, most likely related to rain, a large amount of the signal would be within a few percentages of the scale. Secondly, to compensate for this, the signal was normalized according to the density function. The challenge with this approach was that it complicated comparison to the original signal. For instance, if a prediction of the signal was 5 % off in the normalized data this could correspond to 10 % or 2 % in reality. Therefore, the choice was made to normalize the data linearly up to the 95 % percentiles. Values above the 95 % percentile were set to one. Hereby the peaks in the signal were kept while the variations in the primary part of signal were not severely limited. The energy usage was measured for 538 days. The first 443 days were used for training, and the remaining 95 days were used for testing.

5.2.3 Models

Two model types were trained. The first was a MLP model. The second was a 1D CNN Autoencoder which reconstructed the signal of the pumps all at once.

MLP model

The MLP model was trained to predict the hourly energy usage of each pump based on the energy usage of the remaining pumps. The model used the energy usage from one hour to predict the energy usage in the remaining pump for the same hour. Thereby this model did not encounter temporal variations. The MLP model used the standard settings in the python library scikit-learn version 0.24.2 [10]. A MLP model was trained

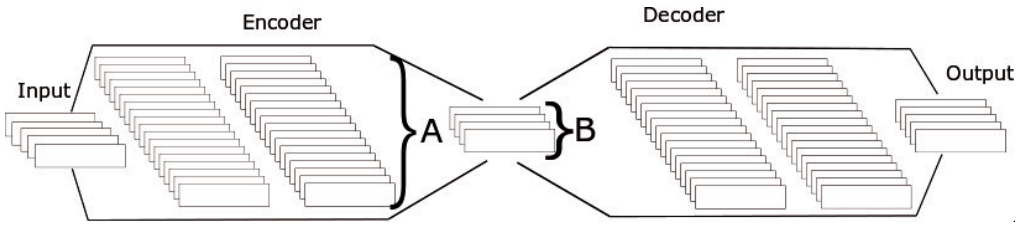


Fig. 5.2: An overview over the used 1D Autoencoder CNN. In the figure, A marks the encoder and B marks the width of the bottle neck.

for each of the CNN models. To avoid sensitivity to local minima each model was trained and tested five times.

CNN Autoencoder[1]

For the 1D CNN the learning rate was set to 0.0005, the batch size was set to 32 and the kernel had a length of four. The training of the CNN was done through 250 epochs. The encoder and decoder were mirrored versions of each other and had a depth of three convolutional layers. Batch normalization was applied after each convolutional layer. Figure 5.2 shows an overview of the used Autoencoder. In the figure, A illustrates the encoder width, which also corresponds to the decoder width. B illustrates the width of the bottleneck. To optimize the Autoencoder, different widths of the encoder and decoder and the bottleneck were tested. Based on this a decision was made to set the width of the encoder and decoder to 64 and the bottleneck to four. The CNN model was trained five times.

5.2.4 Test and performance metrics

In general, evaluation of unsupervised or semi-supervised anomaly detection methods are challenged by missing knowledge about ground truth anomalies.

Autoencoders are used to reconstruct a signal, and anomalies are typically defined if the difference between reconstruction and input signal exceeds a certain threshold. Therefore it needs to be able to reconstruct the correct signal both if a correct signal is used as input, and if a disrupted signal is used as input. A disrupted signal can entail any types of faults added to the signal. In this case, the decision was made to use an average of the daily values.

Two metrics were used for evaluation of the performance: Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). The challenge with MAPE is that if the signal has a low amplitude, a small deviation in the signal will give a high error.

[1]I would like to thank Mark Philip Philipsen for his assistance in setting up the CNN.

On the other hand, the MAE favors the pumps with a low energy usage, as the model does not need to be good at identifying variations in the signal if they are on a limited scale. For this reason, both MAPE, MAE and visual inspection were used to evaluate the models.

5.3 Results

The results are presented in Figure 5.3 and Table 5.2.

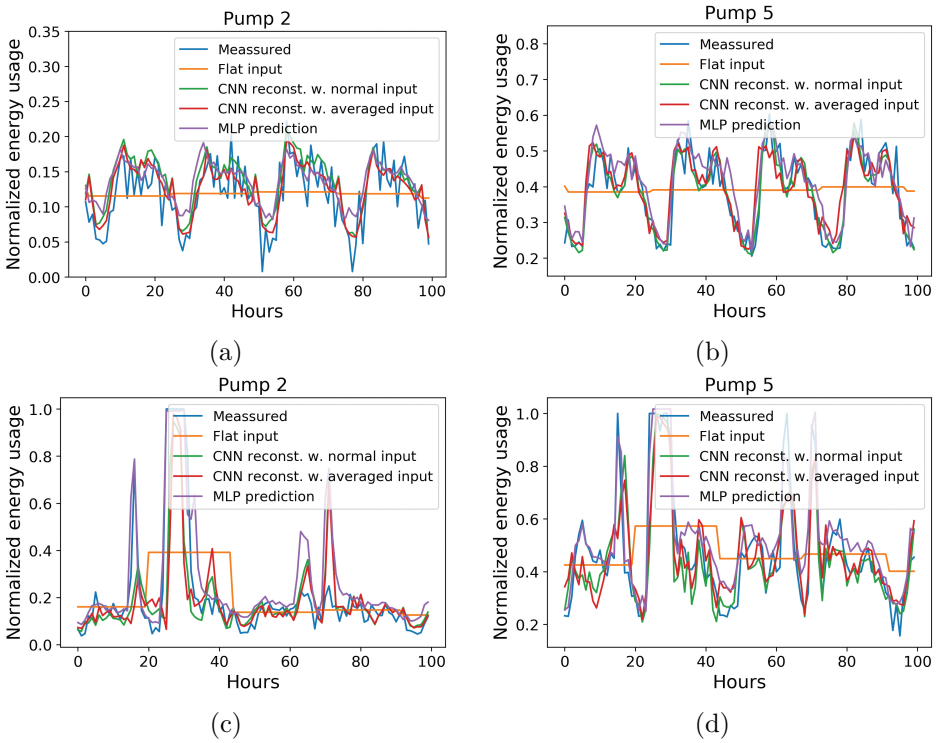


Fig. 5.3: Examples of the measured energy usage, flat input signal, CNN reconstructions, and MLP predictions for pump 2 and pump 5 on dry days ((a) and (b)) and rainy days ((c) and (d)). Notice that the second axis varies between the plots to show more details.

On average, the MLP model performs better when considering the MAE. Furthermore, on average it obtains a better MAPE than the CNNs model when the CNN is fed with the disrupted signal. However, when the correct signal was fed to the CNN the CNN obtained a better MAPE than the MLP model.

Table 5.2: Performance of the three approaches when using a CNN Autoencoder with the buttleneck width set to four and the width of the encoder and decoder set to 64. The results have been obtained by training the models five times, and the std indicates the standard diviation for the five runs.

Pump nr	**CNN normal signal**		**CNN flat signal**		**MLP prediction**	
	MAPE	*MAE*	*MAPE*	*MAE*	*MAPE*	*MAE*
1	22.20 ± 0.29	0.0391 ± 0.0003	23.95 ± 0.17	0.0422 ± 0.0001	22.78 ± 2.60	0.0368 ± 0.0012
2	40.95 ± 0.64	0.0454 ± 0.0003	41.62 ± 0.60	0.0480 ± 0.0004	49.88 ± 4.69	0.0485 ± 0.0010
3	18.68 ± 0.69	0.0438 ± 0.0002	20.45 ± 0.21	0.0520 ± 0.0004	15.33 ± 0.46	0.0370 ± 0.0005
4	28.62 ± 0.28	0.0647 ± 0.0005	31.73 ± 0.29	0.0740 ± 0.0003	28.78 ± 0.98	0.0620 ± 0.0013
5	12.68 ± 0.09	0.0553 ± 0.0006	16.35 ± 0.21	0.0657 ± 0.0005	13.21 ± 0.40	0.0515 ± 0.0008
Total	24.63	0.0497	26.82	0.0564	26.00	0.0472

Generally, the CNNs have a lower standard deviation than the MLP models, indicating that the method is more stable.

Visual inspection of the results shows that all the models were able to predict or reconstruct general tendencies in the signal, though they could not follow the hourly variations, as seen in Figure 5.1. Furthermore, they sometimes over- and under-shoot during high load in the system. This is problematic for the purpose of fault detection as this would entail that high load caused by rain would be detected as faults. Therefore, developers should be careful when replacing original signals with reconstructed or predicted signals if the replacement is caused by large variations between the measured and predicted values.

A general challenge when using hourly signals is that a large part of the information in the original signals is lost. Furthermore, rainfalls often last less than an hour, and thereby the benefit of using CNNs, which can encounter timely variations, are smaller than if there was a higher temporal resolution in the data.

5.4 Contributions

- It was shown that MLP and Autoencoder CNNs can predict and reconstruct most of the hourly energy signals. However, if this should be used for anomaly detection more research is needed to avoid rainfalls being detected as anomalies.
- It was shown that CNNs obtained a more stable performance than MLP, whereas MLP showed a slightly better performance in average.
- We argued that better temporal resolution of the energy usage could entail a better model performance for the CNN model.

References

[1] A. Azadeh, M. Saberi, A. Kazem, V. Ebrahimipour, A. Nourmohammadzadeh, and Z. Saberi, "A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization," vol. 13, no. 3, pp. 1478–1485. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494612003006

[2] M. Bertolini, D. Mezzogori, M. Neroni, and F. Zammori, "Machine learning for industrial applications: A comprehensive literature review," vol. 175, p. 114820. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S095741742100261X

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," vol. 41, no. 3, pp. 1–58. [Online]. Available: https://dl.acm.org/doi/10.1145/1541880.1541882

[4] J. Dai, J. Tang, S. Huang, and Y. Wang, "Signal-based intelligent hydraulic fault diagnosis methods: Review and prospects," *Chinese Journal of Mechanical Engineering*, vol. 32, no. 1, sep 2019. [Online]. Available: https://doi.org/10.1186%2Fs10033-019-0388-9

[5] J. Ellerbæk Nielsen, M. R. Rasmussen, S. Højmark Rasmussen, D. Getreuer Jensen, and A. Hertz Kristensen, "PUFDO - PUMPE FLOW TIL DRIFTSSTATUS OG -OVERBLIK DANVA VUDP PROJEKTRAPPORT." [Online]. Available: https://www.danva.dk/media/7223/pufdo-slutrapport-samlet.pdf

[6] C. Fan, F. Xiao, Y. Zhao, and J. Wang, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data," vol. 211, pp. 1123–1135. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261917317166

[7] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," vol. 287, p. 116601. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261921001409

[8] C. S. Kallesoe and T. Knudsen, "Self calibrating flow estimation in waste water pumping stations," in *2016 European Control Conference (ECC)*. IEEE, pp. 55–60. [Online]. Available: http://ieeexplore.ieee.org/document/7810263/

[9] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, apr 2020. [Online]. Available: https://doi.org/10.1016%2Fj.ymssp.2019.106587

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," vol. 12, pp. 2825–2830.

[11] W. Zhou, X. Li, J. Yi, and H. He, "A novel UKF-RBF method based on adaptive noise factor for fault diagnosis in pumping unit," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1415–1424, mar 2019. [Online]. Available: https://doi.org/10.1109%2Ftii.2018.2839062

[12] A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," vol. 8, no. 6. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1280

Chapter 6

Drift detection in sensors at Wastewater Treatment Plants

WWTPs and Wastewater Recovery Facilities, prospectively referred to as WWTPs, are responsible for cleaning sewage to a degree that complies with national legislation. In Denmark the requirements for total Nitrogen (N) removal depends on the recipient. Furthermore, in some cases it can be beneficial to discharge below the requirements as there is a fee per kg. discharged total N. The main process for nitrogen removal is as follows: the nitrogen arrives to the WWTP as NH_4; at the plant, bacteria nitrify it to NO_3 under aerobic conditions, which are obtained by aerating the process tanks; and finally, the NO_3 is denitrified to CO2 and free N_2 through anaerobe conditions. Historically, aerobic and anaerobic conditions have been obtained by aerating the Process Tank (PCT) when the NH_4 concentration reaches a certain threshold and stopping the aeration when the NH_4 concentration is below a certain concentration. An illustration of this is shown in Figure 6.1

Aeration of process tanks is the main energy consumer at the plants, and several studies on optimization of WWTPs exist [21].

The global need for reaching the SDGs and increasing energy prices have led to increased focus on optimizing the operation of WWTPs.

For process optimization, several plants have replaced the alternating operation approach with more modern techniques such as Proportional–Integral–Derivative (PID) control. In PID controlled process tanks, a set point for the NH_4 concentration is set and the goal for the PID controller is to reach this set point. To this end, the PID adjusts the aeration based on the proportional error (P), which is the error between the measured NH4 concentration and the set point, the Integral error (I), which is the total error over time, and the derivative error (D), which is the expected error in the next timestamp. How fast the controller changes according the P, I, and D is defined by

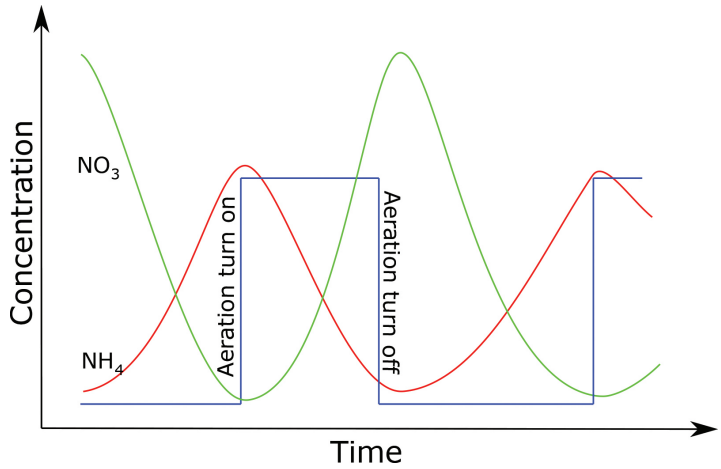


Fig. 6.1: Illustration of how NH_4 concentration (red) increases overtime until it reaches a certain level and the air pumps are activated (blue). Thereafter, the NH_4 nitrifies to NO_3 (green) during aeration. When the air pumps are turned off again the NH_4 concentration starts to increase and the NO_3 denitrifies to N_2 and CO_2

three constants. If the constants are high, the controller adjusts quickly but typically overshoots the adjustment. In PID controlled process tanks, the aeration is typically performed in the inlet to the PCT while the outlet is kept anaerobic. In these cases the approach is often to keep the concentration of the different substances in the outlet of the PCT constant or slightly changing according to the load.

Today, several commercial software systems for optimizing the control of WWTPs are on the market, but few focus on the data quality [7]. Especially, sensor drift is often present in the data. This is problematic as it can entail reduced total N removal and over-aeration with large increase in energy consumption as a consequence. The cost savings archived by implementation of advanced automatic control systems can easily be counteracted by bias in sensors [17].

Sensor drift is a commonly known problem, and the magnitude of the drifts can easily reach one mg/l for Dissolved Oxygen (DO) sensors [16], ammonia sensors, and potassium sensors [8]. Furthermore, some utilities accept that NO_3 sensors have offsets of up to $\pm$ 1 mg/l and that the NH_4 sensors have offsets of up to $\pm$ 0.5 mg/l. The size of the drifts are very large when considering the concentrations of the substances. For instance, a drift of ± 1 mg in the NH_4 sensor represents a large part of the signal as the NH_4 levels rarely exceed 3 mg/l. This is problematic as most WWTPs use the data from NH_4 sensors for aeration control [8].

6.1 Existing approaches

Faulty sensor data are commonly known in industrial settings [18]. Several different fault and anomaly types can be found in sensor data, including outliers, missing data, bias, drift, and constant values.

Corominas et al. [7] found that the most popular data-driven methods for fault detection in WWTPs were Principal Component Analysis (PCA) followed by Independent Component Analysis (ICA) and Clustering. However, despite several methods having been developed, only 16 % of the data-driven methods developed for WWTPs were commercialized [7].

Several papers on anomaly and drift detection in WWTPs have been published. In 2002 Thomann et al. [19] showed that control charts could be used for drift, shift and outlier detection in real WWTPs, and in a review, Newhart et al. [15] confirmed that Control Charts are well-suited for monitoring of single, low-noise variables.

Later, different PCA based methods have been popular for anomaly detection in WWTPs [1, 3, 6], but also methods such as univariate statistics [4] and Q-statistics have been used [3].

Recently, several papers on drift detection have been published. An overview of papers available at Scopus in the period 2020-2021 can be seen in Table 6.1.

Table 6.1: Papers on drift detection published in 2020-2021.

Author	Year	Data	Method
Ba-Alawi et al. [2]	2021	Simulated dry weather	Stacked Denoising Autoencoders
Cecconi and Rosso [5]	2021	>1 year of in-control data from a real plant; faults were subsequently introduced.	ANN and Shewhart Control Charts
Kazemi et al. [10]	2020	Simulated data	Incremental PCA
Kazemi et al. [9]	2021	Simulated data	Support Vector Machine Ensemble Neural Network Extreme Learning
Klanderman et al. [11]	2020	Trained on in-control data, tested on simulated data and real data containing one fault	Autocorrelation and Fused Lasso
Luca et al. [12]	2021	Simulated data	PCA and statistic
Mamandipoor and Majd [14]	2020	Real data (11 months)	Long Short-term Memory Network
Xu et al. [20]	2021	Simulated and real data	Independent Component Analysis

From Table 6.1 it can be seen that in four of the eight papers the methods are only tested on simulated data. Xu et al. [20] tested the method in both a simulated and a real case. In the real case, 213 samples were available; of these, 45 represented normal behaviour and were used for training the model. However, these samples were also used for testing together with the rest of the dataset. This is problematic as it allows the algorithm to have a zero tolerance for samples not totally similar to one of the 45

training samples. Thereby the test can show an incorrect high performance which would not be present if the test was carried out on a fully unknown dataset.

Klanderman et al. [11] tested the algorithm on a real dataset containing one fault, which they were able to detect. Mamandipoor et al. [14] had almost a year of data from a real WWTP. Faulty behaviour in the dataset was annotated by an expert and the remaining data were considered correct. An example of correct and faulty behaviour was included in the paper, and it could be seen that the faulty data were very distinctive from the normal behavior data.

Cecconi et al. [5] installed six sensors in a WWTP and trained an Artificial Neural Network (ANN) to predict the value of each of the sensors based on the remaining five sensors. The difference of the ANN model and the measured values was used as input to a Shewhart Control Chart. The sensors were cleaned every week and three weeks of data were used for testing, but there were no faults in the test data, for which reason faults were added subsequently and successfully found. Furthermore a PCA based solution was used as input to a Shewhart Control Chart. However, this method was better for point anomalies than longer and more problematic anomalies such as drift and calibration biases. A challenge with the developed ANN was that it should be retrained every time sensors are replaced, after adjustment in the treatment process and to encounter seasonal changes such as dry and wet weather periods. Furthermore, retraining should be done if an alarm is given after successful calibration of a sensor [5].

In addition to the work presented in Table 6.1, Mali and Laskar et al. [13] used Monte Carlo and ANN to find bias in simulated data.

As seen above, several papers on drift detection have been published, but the methods are typically developed based on simulated data, data from highly controlled setups, or data where anomalies have been applied after the data were collected.

In real sensors the same fault can develop differently in different sensor types and does not occur systematically. For instance, optical Disolved Oxygen (DO) sensors typically drift more and have more complicated drift patterns due to fouling than membrane DO sensors [16]. Another challenge is that real world data is often subject to co-occurrence of faults and anomalies. Furthermore, anomalies such as outliers in the flow to the plant or the composition of the wastewater are often consequences of rain or discharges and should not be considered as faults.

As seen above, several methods for drift detection have been presented in the literature, but they are often based on simulated data, and for the studies using real data, the data do not reflect the circumstances at most operating WWTPs. Thereby there is a gab between the scenarios used for research and development in the literature and most operating WWTPs.

6.2 Drift detection in operating WWTPs

The insights from Section 6.6.1 motivated us to investigate if the methods presented in the literature could be applied to operating WWTPs where precautions to ensure a high data quality for fault and drift detection have not been made, as this better reflects the actual conditions at most operating WWTPs. In this section our findings are documented.

6.2.1 Data

In total, five datasets from three different WWTPs were available. Each dataset represented a Process Tank (PCT) and contained 2 min. measures of the NH_4, NO_3, and DO measured in the PCTs, as well as the flow to the WWTP. One of the PCTs used alternating aeration, while the remaining PCTs primarily used PID for aeration control.

Due to a small number of lab measurements of the exact concentrations of the substances, attempts were made to make a ground truth dataset by manually labeling the data for the PCT with alternating operation, as this was considered the easiest to label. However, it was difficult to label the data without introducing faults because the operator had dealt with drift in the NH_4 sensor by changing set points and because an offset of up to 0.5 mg/l and 1 mg/l was accepted for the NH_4 and NO_3 sensors, respectively.

6.2.2 Supervised learning

The lack of a ground truth dataset entailed that traditional supervised learning was not possible. Therefore, a consideration was made to use one-class learning to predict the different concentrations and the flow. As sensor drift appears over a longer period, the prediction should not be based on the immediate previous prediction, to prevent the algorithm predicting the sensor measurement instead of the real concentrations and flow. The concentrations and flow during a day depend on rush hour, rain fall, and when the aeration pump is active. These variations are smaller when considering the daily average. Therefore, a Random Forest model was trained to predict the concentrations and flow on the first 80 % of the data and tested on the remaining 20 % of the data. However, change in the control settings entailed that the predictions were highly imprecise. For this reason, the decision was made to use unsupervised learning.

6.2.3 Unsupervised learning

As PCA is a popular approach in the literature, the data was normalized and plots for all combinations of PCs on a daily basis were made and visually inspected. It was observed that the patterns in the plots changed over time, especially with change

in control settings being distinguishable. For PCTs with alternating operation over-aeration was distinguishable as well. However, this was even more clear when plotting all possible combinations of parameters without performing PCA. An example of how a change in set points changes the patterns in the data can be seen in Figure 6.2. Figure 6.3 shows data from a PID-controlled plant. Here it is worth noticing that patterns in the PID controlled PCT are different from the patterns in the plant with alternating operation. This also underlines the challenges when parameters in the PID controller are changed. For instance, settings in PID controllers can entail a pattern similar to alternating operation.

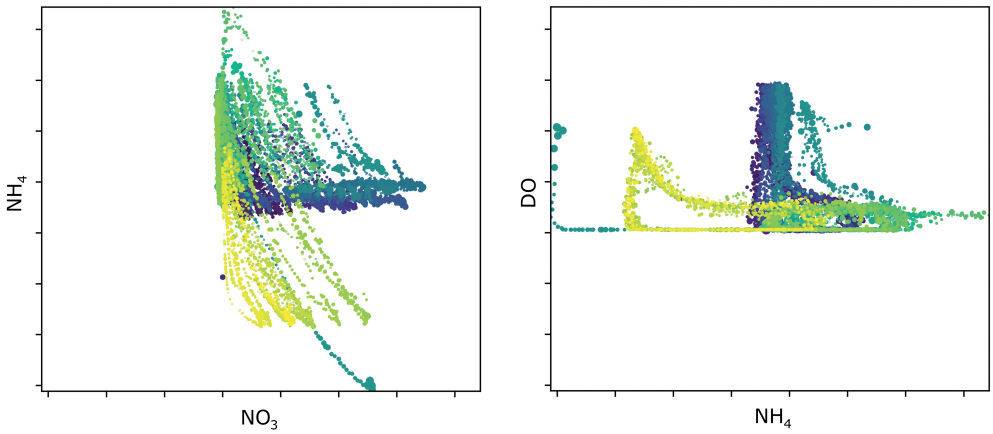


Fig. 6.2: The figure shows eight days of data from a PCT with alternating operation. The colors indicate the time, where the dark blue are the first data points and the yellow are the last data points. In the shown period the set points have been changed. The dark blue and green colors indicate measures collected before the change in set point and the yellow and light-green colors indicate measurements collected after the change in set point.

For general tendencies the data were averaged on a daily level and PCA was performed. Attempts were made to remove the least contributing PC; however, as the faults were present in all PCs, this did not entail identification of faulty data.

Another unsupervised approach considered was to use an ANN Autoencoder. However, as the one-class learning approach did not work due to change in control settings, it was not expected that this approach would work either, for which reason it was not tested.

Based on the challenges described above, a decision was made to use an algorithm which could detect if a data point was behaving abnormal instead of predicting the exact data point. Furthermore, initial tests of the algorithm Local Outlier Factor (LOF) showed promising results, for which reason LOF was chosen.

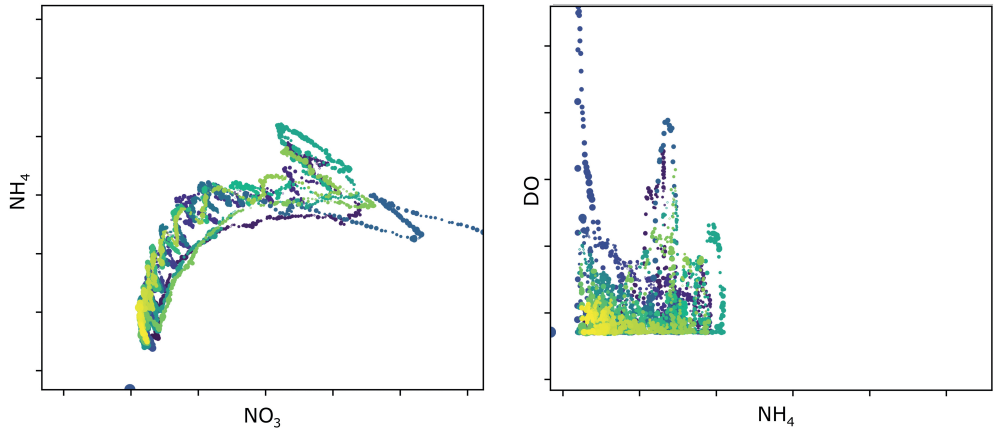


Fig. 6.3: The figure shows five days of data from a PID-controlled PCT. The colors indicate the time, where the dark blue are the first data points and the yellow are the last data points. Measurements of the sensors during the five days showed that the sensors had not drifted over the accepted amount.

6.2.4 Local Outlier Factor

The LOF for a data point is found by measuring the distance to the closest data points. Thereby data points which are far from other data points will obtain a higher LOF than data points positioned close to other data points. Figure 6.4 is an illustration of LOF for two-dimensional data points. However, the concept is similar for the four-dimensional data points used in this work.

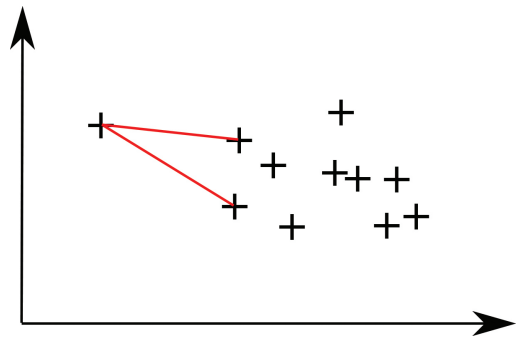


Fig. 6.4: Illustration of the LOF for two-dimensional data points. The red lines shows the distances used to calculate the LOF

To ensure that the method could be used in operating plants a decision was made to use a Moving LOF, where the LOF of a given data point was based on the previous

99 data points. In this work the distances to the 20 closest data points were used to calculate the LOF. Subsequently, a threshold was applied, and all data points with a higher LOF were detected as anomalies.

6.2.5 Assessment of anomalies

The detected anomalies were manually analyzed and categorized into one of the following categories: *missing data*, *increased presence*, *change in control settings*, *Drift or over-aeration*, and *Other*. In cases where an anomaly could fit into more than one of the categories, the anomaly was put into the category which triggered the LOF to exceeded the threshold. For instance, in some cases it was observed that the NH_4 sensor had drifted without the LOF exceeded the threshold and then an increase in the flow entailed that the LOF exceeded the threshold. This anomaly would be categorized as increased presence.

6.2.6 Results

The data, Moving LOF, and the detected anomalies for the PCT using alternating operation are presented in Figure 6.5.

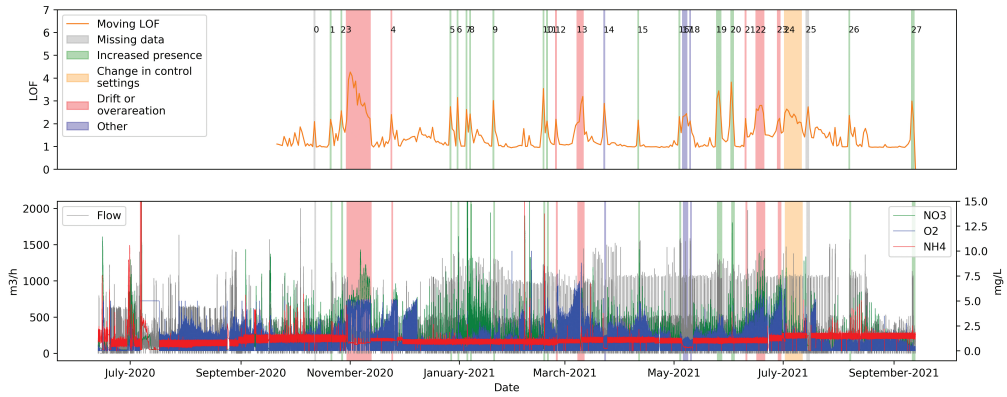


Fig. 6.5: The upper graph shows the Moving LOF (orange). The different color markings indicate that an anomaly is detected and the type of anomaly. The anomalies are numbered for identification. The lower graph shows the full dataset for the PCT with alternating operation. When observing the NH_4 concentrations (red line) it can be seen that it alternates between an upper and a lower set point, which is characteristic for alternating operation. It can also be seen that these set points are changed during the period of data collection. These changes in set points occur after a drift has been detected, indicated by the red patches. The figure is adapted from [8].

From Figure 6.5 it can be seen that several anomalies were detected, and a majority of them were caused by sensor drift. It can also be seen that in most cases the drift

was handled by changing the set points. In Figure 6.5, drift can be spotted as it entails an increase in DO and NO_3 over several days. The increase in DO is caused by the NH_4 sensor measuring to high values. When the sensor measures to high values, the lower set point is not reached before the nitrification process starts to slow down due to low NH_4 concentration. As NO_3 needs anaerobic conditions for denitrification, the NO_3 level also increases along with the DO levels. When the size of the drift increases, so does the amount of over-aeration and NO_3. A detailed plot of anomaly four, which is caused by drift in the NH_4 sensor, is shown in Figure 6.6. More examples of the detected anomalies can be found in Paper E.

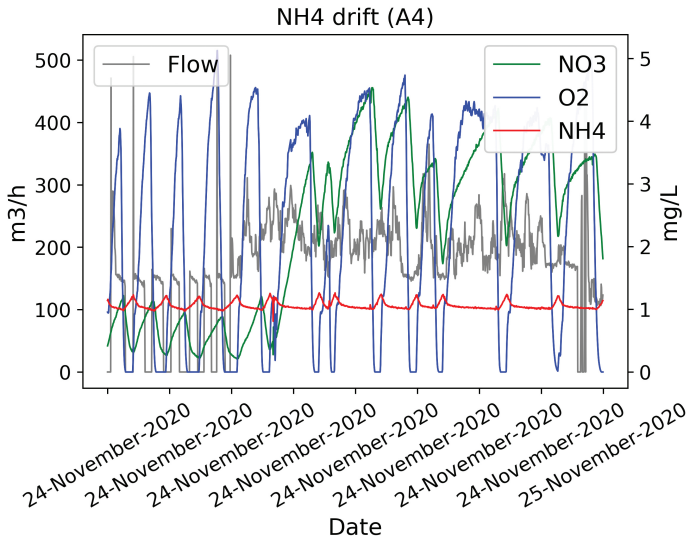


Fig. 6.6: The figure shows anomaly four from the PCT using alternating operation. When the NH_4 concentrations reach the upper threshold the aeration starts and induces the concentration of DO to increase. The NH_4 concentration then starts to decrease. However, it flattens before it reaches the lower set point. This entails a long period of aeration and a high level of NO_3 as the denitrification cannot happen during aerobic conditions. The figure is adapted from [8].

Figure 6.7 shows one of the PCTs which were primarily controlled with PID controllers. Similar plots for the remaining three PID-controlled PCTs and detailed plots of different anomaly types can be found in Paper E. An overview of all the detected anomalies with descriptions is provided in Table 6.2.

From Figure 6.7 it can be seen that despite the plant being primarily PID-controlled, the patterns of the NH_4 sensor indicate alternating operation in the first part of the data collection. Furthermore, short periods with patterns indicating alternating operation are present in the middle of the period for data collection. It can be seen that most of

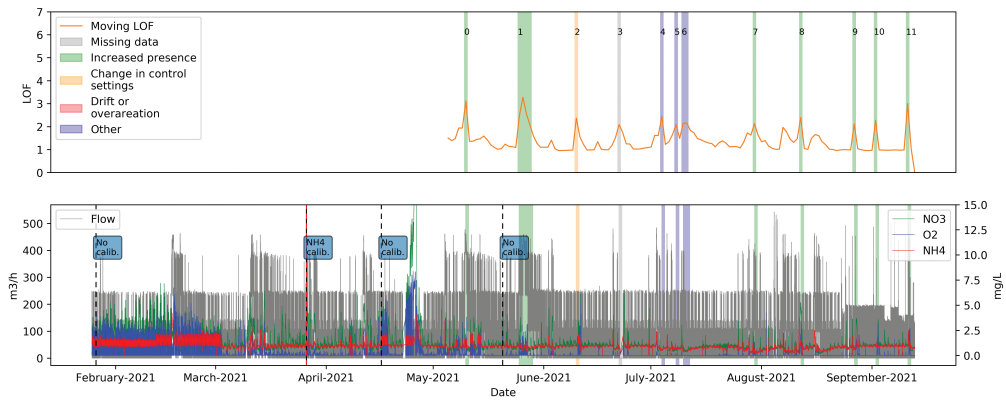


Fig. 6.7: The upper graph shows the Moving LOF (orange). The different color markings indicate that an anomaly is detected and the type of anomaly. The anomalies are numbered for identification. The lower graph shows the full dataset and information on calibration. When observing the NH_4 concentrations (red line) it can be seen that in the beginning of the period the pattern indicates alternating operation, and in the beginning of Marts, the more flat NH_4 concentrations indicate that the plant is controlled by a PID controller interrupted by short periods of alternating operation. The figure is adapted from [8].

the detected anomalies in this PCT were caused by increased presence of a parameter, e.g., increased flow due to rain. Similar patterns and tendencies were observed in the other PID-controlled PCTs. An overview of all the detected anomalies are presented in Table 6.2.

Summarizing the results, several different types of anomalies were detected. The most common cause of the anomalies were increased presence of flow or substances, which are not faults. For the PCT with alternating operation the second most common cause of anomalies was drift. For the PID-controlled PCTs the second most common cause of anomalies were change in control settings.

No cases of sensor drift towards lower values were detected, but when visually inspecting the data several cases of longer periods with low DO and NO_3 concentrations could be observed indicating drift in these sensors. However, as an offset of $\pm$ 1 mg/l for the NO_3 sensor was accepted from the utility it was hard to distinguish between acceptable and unacceptable low concentrations. From an algorithmic view, it is more difficult to detect drift in negative direction than in positive direction, as the size of a negative drift is limited by the sensor not measuring concentrations lower than 0 mg/l. Both positive and negative drift in the NH_4 sensor can entail a too high discharge of total N.

Using a Moving LOF on historical data entails that the sensors have not been calibrated when a drift is detected. Thereby, the drift is sometimes present in several of

Table 6.2: Description of all the detected anomalies. The table is adapted from [8].

WWTP 1 PCT 1	WWTP 2 PCT 1	WWTP 2 PCT 2	WWTP 3 PCT 1	WWTP 3 PCT 2
0. Increased flow	0. Missing data	0. Increased flow combined	0. Change in PID. There is an	0. Increased flow
1. Increased flow	1. Increased flow + NH4	with changed control	increased flow starting the	1. Change in
2. Change in PID	drift (up)	settings the previous day	day before and continuing	control settings
3. Missing data	2. Increased flow + NH4	1. Increased flow	two days after the anomaly	2. Increased flow
4. Other	drift (up)	2. Increased flow	was detected. In the period	3. Increased flow
5. Other	3. NH4 drift (up)	3. Increased flow	of the anomaly, the pattern	4. Increased NO3,
6. Other	4. NH4 drift (up)	4. Increased flow and	of the sensors changed,	low DO
7. Increased flow	5. Increased flow	increased NO3 unrelated	indicating change in	5. Increased flow,
8. Increased NH4,	6. Increased flow	to flow	control settings.	increased NO3,
NO3 and DO	7. High concentrations of	5. Increased NO3	1. Increased NH4	low DO
9. Increased NO3	NH4 and NO3	6. Increased NO3	concentration due to	
and DO	8. High concentrations of NO3	7. Increased flow	missing aeration	
10. Increased NO3	present or NO3 sensor	8. Data shows low flow, very	2. Increased flow, inducing	
and DO	drifted (up)	large amounts of DO and	high NH4, NO3 and	
11. Increased flow	9. Increased flow	increasing NH4.	DO concentrations	
	10. Increased flow	9. Increased flow	3. Increased flow, inducing	
	11. Increased flow	10. NO3 and NH4 the first day,	high NH4 and NO3	
	12. NH4 drift (up)	increased flow the	concentrations	
	13. This anomaly starts as NH4	second day	4. Increased NH4	
	drift (up). The second day	11. Increased concentrations of	concentrations inducing	
	data is missing for almost	NH4, NO3 and DO.	high NO3 concentrations	
	13 h. Hereafter, the lower	Possible because the other	5. Increased NH4	
	setpoint seems to be slightly	PCT at the WWTP was out	concentrations inducing	
	increased with 0.1, which	of operation, see anomaly	high NO3 concentrations	
	handles the problems with	14 for WWTP2 PCT1	6. Increased flow, inducing	
	over-aeration. The last day	12. Increased flow	high NH4 and	
	of the anomaly is due to an	13. Increased concentrations of	NO3 concentrations	
	increased flow.	NH4, NO3 and DO. One		
	14. All parameters are low	day with increased flow.		
	except for the flow. Maybe	Possible because the other		
	this PCT has been out of	PCT at the WWTP is out of		
	operation or experiments	operation, see anomaly		
	had been performed.	17–18 for WWTP2 PCT1.		
	15. Increased flow	14. Increased concentrations of		
	16. Increased flow	NH4, NO3 and DO. One		
	17. Low parameters, see 14	day with increased flow,		
	18. Low parameters, see 14	see 13.		
	19. Increased flow	15. Increased flow		
	20. High levels of NH4 present	16. Increased flow		
	day the first day, increased	17. Increased flow		
	flow the second day	18. Missing data		
	21. NH4 drift (up)	19. NH4 sensor drifted (up)		
	22. NH4 drift (up)	20. Increased flow		
	23. NH4 drift (up)	21. Increased flow		
	24. Change in setpoint. In the	22. Increased flow		
	period up to the detection	23. Increased flow		
	of this anomaly the			
	setpoints were increased			
	multiple times. This also			
	happened two days before			
	this anomaly was detected.			
	The day before this			
	anomaly was detected, the			
	setpoints were decreased			
	inducing over aeration. The			
	day after the setpoint was			
	increased again, which was			
	the case for the remainder			
	of the anomaly. The LOF			
	decreased over time as it			
	learnt the new behaviour			
	25. Missing data			
	26. Increased flow and NH4			
	concentration			
	27. Increased flow and NH4			
	concentration			

the data points used to calculate the LOF, which entails that the drifts prospectively are less distinguishable.

6.2.7 The gap between academia and practice

Despite a large number of data-driven drift detection algorithms being presented in the literature, it is difficult to obtain a sufficient performance in operational WWTPs such as those which provided data for this study. The primary reason for this was that the data used for research and development in the literature and the data available from most of the operating WWTPs are based on different circumstances. The data used in the literature is generally characterized by a large degree of control. This can be either simulated data or data from a plant with a large focus on sensor cleaning. In these cases, drifts can systematically be introduced for each of the different sensors individually. Furthermore, for simulated WWTPs, there are full control with external factors such as flow and composition of the wastewater. Contrary to sensors in simulated WWTPs, sensors in real WWTPs often have an offset within an acceptable range. However, this entails that faults cannot be fully isolated. Generally, there is no control or a low degree of control with the flow to the plants and the composition of the wastewater. Furthermore, drift and external circumstances, such as rain or large discharges of wastewater from industries, can motivate operators to change the control settings at the plants. However, these changes in control settings are rarely documented, and thereby difficult to encompass in the models.

To achieve a sufficient dataset for drift detection, the sensors at the WWTP need to be maintained, measured, and calibrated on a frequent basis, preferably once a week. Furthermore, the dataset needs to include several months of data, preferably covering over a year or more. In this period all changes in control settings should be limited to as few as possible, and if a change is needed, for instance due to increased flow, this should be documented, and the the original settings should be used as soon as possible. As lab measurements are not expensive, the main hindrance is the culture among the operators of the WWTPs. Therefore, if a utility is able to achieve a sufficient dataset for drift detection, the problem with sensor drift is most likely already solved, as this would entail a change in the culture among the operators at the WWTPs. Furthermore, if any condition is changed, such as the control settings or catchment area, a new dataset is needed, entailing that the period of having a functioning algorithm is short compared to the period for data collection.

Another solution for handling of drift is by implementation of self-calibrating sensors.

Despite the data quality not being sufficient for data-driven drift detection, the data quality is sufficiently high for other purposes, and parameters such as the energy usage could be well suited for bench-marking of the plant.

For further details on the gap between solutions for drift detection in academia and practice and for recommendations on bridging this gap, please refer to Paper E.

6.3 Contributions

- It was demonstrated that the data-driven solutions for drift detection developed in academia often are fitted to in-control dataset, which does not reflect the conditions in operating WWTPs.
- We provided insight into the factors entailing low data quality.
- It was argued that if a utility can achive a dataset with sufficiently high data quality for drift detection, the challenges with drift might already be solved. However, despite the datasets available for operating WWTPs having low data quality for drift detection, the data quality might be sufficient for bench-marking of the plants.

References

[1] J. Alferes, S. Tik, J. Copp, and P. A. Vanrolleghem, "Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection," vol. 68, no. 5, pp. 1022–1030. [Online]. Available: https://iwaponline.com/wst/article/68/5/1022/17654/Advanced-monitoring-of-water-systems-using-in-situ

[2] A. H. Ba-Alawi, P. Vilela, J. Loy-Benitez, S. Heo, and C. Yoo, "Intelligent sensor validation for sustainable influent quality monitoring in wastewater treatment plants using stacked denoising autoencoders," vol. 43, p. 102206. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214714421002932

[3] F. Baggiani and S. Marsili-Libelli, "Real-time fault detection and isolation in biological wastewater treatment plants," vol. 60, no. 11, pp. 2949–2961. [Online]. Available: https://iwaponline.com/wst/article/60/11/2949/16099/Realtime-fault-detection-and-isolation-in

[4] I. Baklouti, M. Mansouri, A. B. Hamida, H. Nounou, and M. Nounou, "Monitoring of wastewater treatment plants using improved univariate statistical technique," vol. 116, pp. 287–300. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957582018300387

[5] F. Cecconi and D. Rosso, "Soft sensing for on-line fault detection of ammonium sensors in water resource recovery facilities," vol. 55, no. 14, pp. 10 067–10 076. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.est.0c06111

[6] T. Cheng, A. Dairi, F. Harrou, Y. Sun, and T. Leiknes, "Monitoring influent conditions of wastewater treatment plants by nonlinear data-based techniques," vol. 7, pp. 108 827–108 837. [Online]. Available: https://ieeexplore.ieee.org/document/8789409/

[7] L. Corominas, M. Garrido-Baserba, K. Villez, G. Olsson, U. Cortés, and M. Poch, "Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques," vol. 106, pp. 89–103. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364815217302359

[8] B. D. Hansen, T. B. Hansen, T. B. Moeslund, and D. G. Jensen, "Data-driven drift detection in real process tanks: Bridging the gap between academia and practice," *Water*, vol. 14, no. 6, p. 926, mar 2022. [Online]. Available: https://doi.org/10.3390%2Fw14060926

[9] P. Kazemi, C. Bengoa, J.-P. Steyer, and J. Giralt, "Data-driven techniques for fault detection in anaerobic digestion process," vol. 146, pp. 905–915. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957582020319431

[10] P. Kazemi, J. Giralt, C. Bengoa, A. Masoumian, and J.-P. Steyer, "Fault detection and diagnosis in water resource recovery facilities using incremental PCA," vol. 82, no. 12, pp. 2711–2724. [Online]. Available: https://iwaponline.com/wst/article/82/12/2711/75837/Fault-detection-and-diagnosis-in-water-resource

[11] M. C. Klanderman, K. B. Newhart, T. Y. Cath, and A. S. Hering, "Fault isolation for a complex decentralized waste water treatment facility," vol. 69, no. 4, pp. 931–951. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/rssc.12429

[12] A.-V. Luca, M. Simon-Várhelyi, N.-B. Mihály, and V.-M. Cristea, "Data driven detection of different dissolved oxygen sensor faults for improving operation of the WWTP control system," vol. 9, no. 9, p. 1633. [Online]. Available: https://www.mdpi.com/2227-9717/9/9/1633

[13] B. Mali and S. H. Laskar, "Incipient fault detection of sensors used in wastewater treatment plants based on deep dropout neural network," vol. 2, no. 12, p. 2121. [Online]. Available: http://link.springer.com/10.1007/s42452-020-03910-9

[14] B. Mamandipoor, M. Majd, S. Sheikhalishahi, C. Modena, and V. Osmani, "Monitoring and detecting faults in wastewater treatment plants using deep learning," vol. 192, no. 2, p. 148. [Online]. Available: http://link.springer.com/10.1007/s10661-020-8064-1

[15] K. B. Newhart, R. W. Holloway, A. S. Hering, and T. Y. Cath, "Data-driven performance analyses of wastewater treatment plants: A review," vol. 157, pp. 498–513. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0043135419302490

[16] O. Samuelsson, A. Björk, J. Zambrano, and B. Carlsson, "Fault signatures and bias progression in dissolved oxygen sensors," vol. 78, no. 5, pp. 1034–1044. [Online]. Available: https://iwaponline.com/wst/article/78/5/1034/63680/Fault-signatures-and-bias-progression-in-dissolved

[17] O. Samuelsson, G. Olsson, E. Lindblom, A. Björk, and B. Carlsson, "Sensor bias impact on efficient aeration control during diurnal load variations," vol. 83, no. 6, pp. 1335–1346. [Online]. Available: https://iwaponline.com/wst/article/83/6/1335/79992/Sensor-bias-impact-on-efficient-aeration-control

[18] H. Y. Teh, A. W. Kempa-Liehr, and K. I.-K. Wang, "Sensor data quality: a systematic review," vol. 7, no. 1, p. 11. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-0285-1

[19] M. Thomann, L. Rieger, S. Frommhold, H. Siegrist, and W. Gujer, "An efficient monitoring concept with control charts for on-line sensors," vol. 46, no. 4, pp. 107–116.

[20] C. Xu, D. Huang, D. Li, and Y. Liu, "Novel process monitoring approach enhanced by a complex independent component analysis algorithm with applications

for wastewater treatment," vol. 60, no. 38, pp. 13 914–13 926. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.iecr.1c01990

[21] W. Zhang, N. B. Tooker, and A. V. Mueller, "Enabling wastewater treatment process automation: leveraging innovations in real-time sensing, data analysis, and online controls," vol. 6, no. 11, pp. 2973–2992. [Online]. Available: http://xlink.rsc.org/?DOI=D0EW00394H

Chapter 7

Value Creation

In the previous chapters, potential use cases for machine learning were identified and assessed, and research was performed within four use cases with potentials for value creation. Through this process experience was gained within the cross-fields of the water sector and machine learning. In this chapter, the most important takeaways have been collected, and machine learning for value creation in the water sector has been assessed from three perspectives, for the benefit of future decision making.

First, an analysis of the method used for assessing use cases for machine learning has been made. This analysis is based on experience from the use cases, which were subject to research and development in this thesis. The result of the analysis is an overview over which parameters should be considered when evaluating the investment and risk in relation to machine learning use cases in the future.

One key hindrance for machine learning solutions in the water sector is related to the data quality. Therefore, the second perspective considered in this chapter is related to data quality and how to create value from the data available today.

The third perspective considered is related to the potential for value creation in the water sector using machine learning.

7.1 Analysis of the method for assessment of use cases

Often, decision makers in the water sector lack experience with machine learning, which is problematic as it makes it difficult for them to make qualified decisions on how to optimally invest in machine learning. In Chapter 2 several use cases for machine learning in the water sector were identified and assessed according to their economic potential, the required investment considered as development time, and the risk related to development of the use cases. For these assessments several parameters were defined. However, based on experience within the four use cases researched in this work and experience

with development of use cases generally in EnviDan, there is basis for reconsidering the parameters and the importance of the parameters affecting the development time and the risk related to development. Therefore, the parameters considered in Chapter 2 have been reconsidered to provide a better basis for future decision making in the interdisciplinary area of the water sector and machine learning. Some of the parameters, such as the number of input and output parameters, have been included in other parameters, while new parameters, such as differences between the development environment and the production environment have been included in this evaluation. Each of the considered parameters are discussed, and recommendations for reducing cost and risk are given. An overview of the updated evaluation on parameter importance is shown in Table 7.1.

Table 7.1: Overview of how important selected key factors are for the time duration for a use case and the risk related to it.

Parameter	Importance for time	Importance for risk
Data accessibility	High	Low
Correlation between input and output	Medium	High
Readiness of data	High	Medium
Time for customization	Low-medium	Low
Machine learning method	Low-medium	Low-medium
Difference between development environment and production environment	Medium	Medium

7.1.1 Data accessibility

The amount of work required to access the data depends on whether the data have been collected. If the data need to be collected, this would be much more time consuming than if it is available. Further, if all the data are available from one database, the work related to feature extraction is lower than if the data are spread across multiple databases or locally stored data sheets. From a business perspective, especially locally stored data sets can be a challenge for expanding a solution to different customers.

As shown in Chapter 3 a variable which impacts conditions in the real world does not always contribute to the performance of a machine learning model.

If multiple data sources need to be accessed, it might be beneficial to use fewer but more relevant features for model development, as not necessarily all features which influence the condition in the real world have an effect on the model performance. This was also the case for the features extracted for our work on sewer deterioration modeling

in Chapter 3.

7.1.2 Correlation between input and output

If the correlation between input and output is not sufficient for obtaining the Minimal Viable Product (MVP), the product manager must be prepared to allocate more time for accessing the data, reconsider the MVP, or the form of the product. Otherwise, much time can easily be spent without creating any value. If the MVP cannot be obtained due to lack of correlation between input and output other data sources can be considered. However, there should be a valid argument for the notion that the considered data can improve the performance as accessing new parameters can be time consuming with little or no impact on the model performance. As documented in Chapter 3, a large number of features does not necessarily improve the model performance. In many cases the best solution would be to change the product to a format which requires a lower degree of timeliness, consistency, validity and completeness in the data. This could, for instance, be by using statistical solutions instead of machine learning solutions as shown in the work on drift detection in WWTPs, Chapter 6. If none of the above mentioned solutions can give a MVP, the project manager should consider either dropping the project or reevaluating the cost for development, as it would require new data collection, which in some cases cannot be done within the budget.

Lack of correlation between input and output entails a large risk for 1) increased time spent on data access and method development, and 2) that much time is spent on trying to develop the product without succeeding. Therefore, in order to minimize the risk, the project manager should be very clear on the need in order to obtain a MVP and be quick to drop the product if a prototype does not show promising. In our experience, it is important to avoid spending too much time on improving the machine learning model to obtain the MVP, as improving the machine learning model often only contributes marginally to the performance.

7.1.3 Readiness of data

Considerations should be given to how much work needs to be put into preparing the data. If the data only require standard data cleaning, this does not contribute much to the time expanses whereas if the data needs labeling the time duration will increase. The total time duration for labeling depends on the amount of data which requires labeling and how detailed the labels should be. If much data need labeling an annotation tool can be used. If a usable annotation tool exists, it is preferable to use this to avoid spending time on programming a tool. If labeling is required it is worth considering if the patterns, which should be labeled, are clear for the person who has to label the data. If the patterns are unclear, it can entail a large uncertainty in the labels [6], or it can entail that the data cannot be annotated, as was the case for our work on anomaly detection in WWTPs, Chapter 6.

7.1.4 Time for customization

Compared to the development time the time for customization is relatively low. However, if the product should be distributed to a large number of customers who have specific wishes for adjustment to their specific cases, this can entail a large number of similar scripts which can be be unmanageable and difficult to maintain. The research on sewer deterioration modeling, presented in Chapter3, provides knowledge to a large software asset management product in EnviDan. In this product several adjustments for the different utilities are required. An example of how the deterioration models can be adjusted for the customer is in the definition of the condition states.

In cases which require a large amount of customization, it is particularly important to allocate time to write systematic and well-documented code. However, allocating time for ensuring a high quality in the code is a decision which needs to be taken by the product manager, as the developers usually are rewarded more for fast implementation compared to code quality. Further benefits from high quality code and documentation are that it is easier to later return to the code for adjustment and to get acquainted with for newcomers.

For software products Non-Functional Requirements (NFR) such as scalability, reusability, and execution time need to be encountered to ensure an optimal code infrastructure, reducing the time spent on maintanance and the number and the severity of faults in the code.

Since 2019, the need for allocating time for NFR and documentation has been increasingly recognized by the management in EnviDan.

7.1.5 Machine learning method

The importance of the required machine learning approach depends on the researcher or developer experience levels and access to sparring. For an organization with little to no experience with machine learning or the required machine learning approach (supervised, unsupervised and reinforcement learning) this parameter is more important than for an organization which has a broader range of experience. Since the beginning of 2019, EnviDan has moved from being a company with little experience within machine learning, which only had a small number of products or services using supervised learning, to a company which has experience within all of the three main categories of machine learning.

7.1.6 Differences between the development environment and the production environment

A general challenge in data science is that most solutions are based on historical data, which have some degree of bias according to how they are used. This is the case for both a) sewer deterioration modeling, which is highly dependent on the used inspection

strategy and the year of inspection, b) prediction of the methane production in industrial biogas plants where the available feedstocks change over time, and for c) drift detection in WWTPs, where several external factors such as catchment area and climate change over time.

If a model is tested on data similar to the data used for training and subsequently applied to data with a lower similarity to the training data, this can entail a lower performance of the models when applied in real cases as described in Chapter 3. If this is not managed it can result in a lack of performance which is not discovered before it is reported by the users. This can damage company reputation and entail a need for further development. Furthermore, if the model is trained and tested on simulated data, the model performance might not be sufficient when applied in real cases as shown in Chapter 6.

7.2 Machine learning in the water sector

Data quality refers to the data being "fit for use" and entails a certain level of completeness, timeliness, consistency, and validity. The degree to which these factors need to be met depends on the specific purpose of the data [4, 7]. A general challenge met several times in this work was a lack of data quality compared to the expected model performance.

Sewer deterioration modeling was challenged by a lack of data quality which entailed a lower model performance than hoped for in the start of the project. However, the performance was still sufficient for value creation, as the deterioration model was combined with a consequence model, whereby the risk related to the pipes was found. However, for future usage, it could be considered to include statistical models in the product as this would make it more explainable.

For drift detection in WWTPs, lack of data quality entailed that a machine learning based solution could not be made from the existing data. In the future, the purpose of the product could be changed to be a statistical benchmark of the energy usage. For instance, it is suggested that statistical comparison of the energy usage can both give input on whether the plant is running optimally and be used for evaluation of the effect of a new control method. However, for a fair evaluation it would still be necessary to keep a human in the loop, as the external factors such as climate, season, and catchment area changes over time.

The main challenge for anomaly detection in pumping stations was the low data resolution, as the information in the signal was obscured. In this project the approach for detecting the anomalies was to reconstruct the signals, but the reconstructions were not precise enough for anomaly detection. Despite the data quality not being sufficient for anomaly detection, the data quality was sufficient for estimating the amount of intruding water in sections of sewer systems, which is one of EnviDan's services.

Contrary to the above mentioned challenges, several machine learning solutions have been developed and commercialized in EnviDan while this project has been running. An example of this is PUFDO, a method for calculating the flow through a pumping station based on the energy usage. Another example is localisation of drainage systems connected to sewers. To undertake this, radar images of rainfall were connected to events at the pumping stations with a flow meter, and machine learning was used to identify in which pixels, and thereby locations, rainfall entailed an event at the pumping stations.

The Data, Information, Knowledge and Wisdom (DIKW) pyramid is often used within information science [3]. Figure 7.1 shows the DIKW pyramid and places the methods visualisation, statistics and machine learning according to whether they can contribute with information, knowledge or wisdom. Furthermore, the figure shows that the higher up the hierachy, the stronger decision support systems can be made. However, this also requires a high degree of completeness, timeliness, consistency, and validity in the data, whereas solutions such as visualisation and statistics can provide less strong decision support but also entail lower requirements for the data.

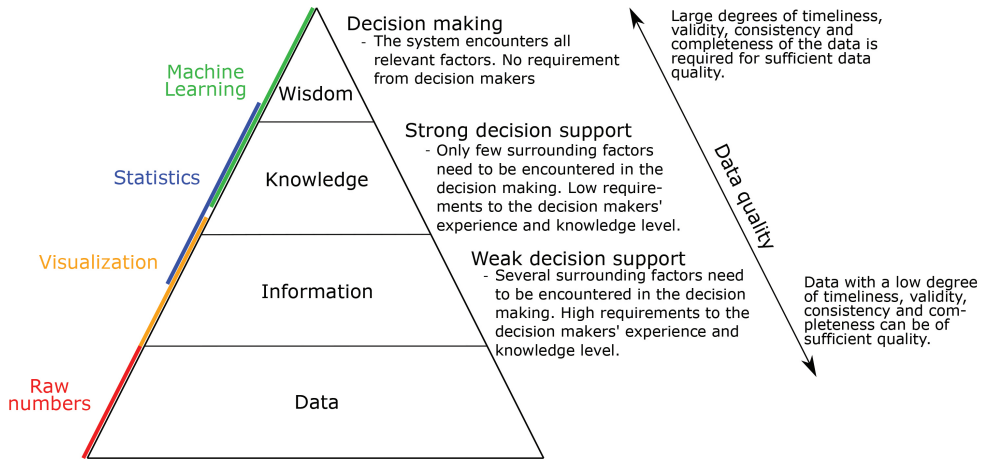


Fig. 7.1: The Data, Information, Knowledge, and Wisdom (DIKW) pyramid is adapted from Frické [3] and put in relation to different digital methods, strengths of decision support tools and requirements for data quality.

Another barrier for machine learning in the water sector is that the sector has primarily been using deterministic models, which are based on a high degree of expert knowledge. In many cases, machine learning methods entail a movement away from the expert based models, which can be a challenge for water professionals who have used these models through their whole carrier. Furthermore, machine learning models are often associated with black boxes, which can make it hard to trust the models. In

this connection it is important to increase the knowledge about machine learning. A key challenge in machine learning is that often a solution can provide good test results, but if the developer of the machine learning solutions does not have a comprehensive understanding of the machine learning method and the data, the test results can often be misleading.

7.3 Drivers for digitalization, AI and machine learning in the water sector

The focus of this work has been machine learning for value creation in the water sector. However, machine learning is a branch of AI which is based on learning algorithms. Likewise, Artificial Intelligence (AI) is a group of methods which can be used in digital setups. If strictly focusing on machine learning or AI, several solutions with potential for value creation will be overlooked. Furthermore, often a solution does not require AI or machine learning but can be solved by utilization of, for instance, simple signal processing.

Recently several drivers for further implementation of digitization AI and machine learning in the water sector has occurred. One of the drivers is higher awareness of the environment among citizens. For example, in the summer 2020, there was a large debate about HOFOR discharging large amounts of mechanically treated wastewater to the sea, resulting in a group of citizen creating an organization for clean water in Øresund.

Another driver is the fact that the Danish government together with other parties agreed that the water sector should be climate neutral in 2030 [5], which is also in line with the SDGs. As described in Chapter 1 AI can be useful for obtaining the SDGs. Furthermore, European programs such as European Horizon, national funds such as the Innovation Fund Denmark, and private funds such as Novo Nordisk Foundation, Carlsberg Foundation, The Velux Foundations, and The Lundbeck Foundation allow for more research in AI, both at the universities and in the private sector, which would not be possible otherwise.

In the period from 2019 to 2021 there has been an increase in companies using AI. A survey showed that, generally, AI contributed with more savings in 2020 than in 2019, whereas the revenue obtained from AI remained at a steady state [2].

Algorithmia [1] reports that only 55 % of companies actively working with machine learning have deployed a model. Furthermore, Algorithmia has observed an increase in companies which have used machine learning but not deployed a machine learning model. The reason for this increase is most likely related to the fact that it takes time to mature AI in a company to a sufficient degree for deploying the AI models [1], a complex process, which has also been experienced by EnviDan.

When considering companies obtaining at least 20 % of their Earnings Before Inter-

ests and Taxes (EBIT) they more often used best AI practices such as machine learning operations (MLOps), cloud technologies and risk-mitigation than companies with a lower EBIT from AI. However, companies with lower EBIT from AI are increasing their engagement in core best practices [2].

The trend toward using best AI practices can also be observed at EnviDan. Challenges with data quality, models degrading over time, and customisation have motivated an increased focus on MLOps. Furthermore, the long-term benefits from using MLOps have been increasingly recognized among the decision makers, despite a required initial investment.

Generally, there is a movement towards more digitization in the water sector, which also entails visualization. It is expected that visualization of data and statistics will entail a larger awareness of how data is collected and what it can be used for. For instance, it might be easier to observe sensor drift and thereby calibrate the sensors earlier if the data is visualized. Furthermore, the impact of, e.g, change of set points will be more obvious. It is considered that this might be an indirect driver for higher data quality.

7.4 Contributions

- Recommendations for assisting future decision making in the interdisciplinary area of the water sector and machine learning were provided to assist optimal decision making.
- Challenges related to data quality and how to overcome these challenges were discussed. Furthermore, it was recommended to consider other ways of creating value from data, rather than strictly focusing on machine learning, for instance, by including visualisation and statistic.
- Finally, examples of drivers for digitization, AI, and machine learning in the water sector were identified.

References

[1] Algorithmia, "2020 state of enterprise machine learning," Algorithmia, Tech. Rep.

[2] M. Analytics, "The state of ai in 2021," McKinsey Analytics, Tech. Rep., 2021.

[3] M. Frické, "The knowledge pyramid: the DIKW hierarchy," *KNOWLEDGE ORGANIZATION*, vol. 46, no. 1, pp. 33–46, 2019. [Online]. Available: https://doi.org/10.5771%2F0943-7444-2019-1-33

[4] R. Mahanti, *Data quality: dimensions, measurement, strategy, management, and governance*. ASQ Quality Press.

[5] Miljøstyrelsen, “Rapportering af ”parismodel” for vandsektoren i danmark,” Miljøstyrelsen, Tech. Rep., 2021.

[6] C. B. Rasmussen, K. Kirk, and T. B. Moeslund, “The challenge of data annotation in deep learning—a case study on whole plant corn silage,” *Sensors*, vol. 22, no. 4, p. 1596, feb 2022. [Online]. Available: https://doi.org/10.3390%2Fs22041596

[7] A. Scarisbrick-Hauser and C. Rouse, “The whole truth and nothing but the truth? the role of data quality today,” vol. 1, no. 3, pp. 161–171. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/17505930710779333/full/html

Chapter 8

Conclusion

Machine learning is a branch of AI which uses learning algorithms. Several industries have benefited from machine learning technologies. Compared to other industries the water sector has been behind in the implementation of machine learning and AI. However, in recent years the usage of machine learning and AI technologies has accelerated, and these are important technologies for reaching the SDGs.

Potential use cases for machine learning in the water sector were systematically identified, assessed, and visualized in a way that one on hand provided new knowledge into the scientific community, and on the other hand gave the decision makers in EnviDan an overview of potential use cases for machine learning and an insight into the related investments from a data science perspective. The approach for visualizing the use cases has subsequently been used for other purposes as well in EnviDan.

Four of the assessed use cases were subject to research in this work. These use cases were *sewer deterioration modeling*, *prediction of methane yield from biogas plants*, *fault detection in pumps*, and *drift detection in WWTPs*.

The work on sewer deterioration modeling included development of a sewer deterioration model, research in optimizing the model and a comprehensive feature analysis of which features contributed to the performance of the deterioration model and how these varies across different utilities. Furthermore, research in ageing curves was performed. This research filled a gap in the understanding of deterioration models and how the models are affected by the available data, which is important for ensuring a high quality of the deterioration models. Furthermore, the research on sewer deterioration modeling has contributed to improvements in EnviDan's software product for asset management.

The second topic, which was forecasting of methane yield from biogas plants, included a proof of concept study focused on whether a machine learning model combined with a deterministic model could improve the forecasts of the methane yield compared to only using one of the models. The results revealed that for forecasting one day ahead

the hybrid model could indeed improve the forecasts.

For fault detection in pumping stations, it was investigated whether CNN-autoencoders could improve anomaly detection when compared to MLP. This was done using data available at EnergiNet's DataHub, which contains hourly measurements of all energy billing meters in Denmark, including the meters at pumping stations. The results revealed that the resolution of the data was too low for the purpose, and that both methods would entail detecting anomalies which were not faults.

For the last topic, drift detection in WWTPs, research was applied regarding how to apply machine learning methods for drift detection in real operating WWTPs. The results showed that there was a large gap between solutions developed in academia and their applicability in most operating plants. The primary reason for this was low data quality, though despite the data not being sufficient for drift detection it might still be sufficient for other purposes such as benchmarking of the WWTPs.

Based on the experience gained from the four use cases the models for assessing the investment and risk were reconsidered, and an updated overview of which parameters should be considered for investment and risk was provided.

Synthesizing all these perspectives, today the biggest potential for value creation in the water sector is not necessarily within machine learning but within applying visualization, statistics, and simple AI. However, in longer terms, it is expected that an increase in data quality, quantity, and general knowledge about machine learning will support a movement towards using more complicated AI and machine learning in the water sector.

This work has provided insight into the interdisciplinary areas of the water sector and machine learning. Research was performed in close collaboration between water professionals and academia which ensured a high relevance of the research. Furthermore, this work has paved the way for more machine learning in the water sector which will benefit the environment and contribute to reaching the SDGs.

The state of machine learning in the water sector is constantly evolving, and there is a large research potential. For sewer deterioration modeling there is a tendency towards either using machine learning or statistics. However, it is suggested that a combination of the two methods could obtain better performance than the models individually. In this connection it is suggested to investigate further how the different defect types are distributed in the sewer networks. For forecasting the methane yield from biogas plants, it is suggested to focus on predictions from one day to two weeks ahead. For fault detection in pumping stations a dataset with high resolution could be used to improve the performance, though the potential in this use case was based on the data being easily accessible and uniform for several different utilities. Within drift detection in WWTPs, it is suggested to put more research focus to statistical bench-marking tools, providing decision support for operators and managers at the plants.

Part II

Papers

Paper A

General Sewer Deterioration Model Using Random Forest

Bolette Dybkjær Hansen, David Getreuer Jensen, Søren Højmark Rasmussen, Jamshid Tamouk, Mads Uggerby, and Thomas Baltzer Moeslund

The paper has been published in the
Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019 pp. 834–841, 2019.

© 2019 IEEE
The layout has been revised.

Abstract

Collapse of sewers can induce significant damage to roads and buildings, resulting in large economical costs. Therefore, utilities wish to repair or replace the sewers before they collapse. In order to investigate if a sewer needs maintenance or replacement it can be inspected with Closed Circuit Television (CCTV), but as CCTV inspection is very expensive, and hence only a small percentage of the sewers are inspected. This underlines the importance of choosing the correct sewers for inspection and have resulted in development of several deterioration models. However, the best performing existing models are tailored to individual cities and need to be calibrated in order to be generalized to new areas. As the cost for collecting a data set for calibration is high, the utilities could benefit from a sewer deterioration model which generalizes across location.

This paper presents a deterioration model based on Random Forest, which is trained on data from 35 utilities spread across the country of Denmark. The model was able to predict the sewer condition with a specificity at 0.80 and a sensitivity at 0.76, which is comparable to the best existing models. This shows that it is possible to make a deterioration model which generalizes across data from different regions, sewers and utilities. This is a significant improvement compared to the current situation where models need to be learned for each new set of data.

A.1 Introduction

The sewer system is an out of sight but very expensive infrastructure to build and maintain. Sewer collapse can induce significant damage in roads and buildings placed on top of them, but due to the underground location of sewers, they are hard to monitor.

Today the sewers are inspected through Closed Circuit Television (CCTV) inspection, which is done by sending a vehicle with a camera into the sewers, and manually inspecting all the videos. This is a very time consuming and expensive process, for which reason only a small fraction of all sewers are inspected during a year.

Still, sewer inspections are important in order to prevent accidental incidents such as collapse of sewers by helping utilities and municipalities to choose the correct sewers for either rehabilitation or replacement. Most utilities have a limited budget for inspections, and for that reason it is necessary to prioritize which sewers to inspect.

Several models for choosing the correct sewers for inspection have been developed. These models prioritize sewers based on the risk related to failure of a sewer and are based on the likelihood for failure combined with the consequence of failure. Hence, the quality of the underlying deterioration model is fundamental for the quality of the management tools, and despite some asset management tools are based on evaluated deterioration models [1, 2] it is not the case for all the models [3, 4]

Deterioration models estimate the deterioration level of the sewers. They can either

be deterministic, statistically, or Artificial Intelligence based. In a review over deterioration models performed by Ana and Bauwens [5] in 2010, they chose to focus on statistical models. The reason for not focusing on deterministic models was due to deterministic models often being too simplistic and the data availability for this type of modelling being limited. The reason for not focusing on AI based models was, among other things, issues with data availability, need for computationally power and explanatory issues with identifying causality. The statistical modelling types covered by the review are: Cohort survival models, Markov chain models, semi-Markov models, logistic regression models, and multiple discriminant analysis [5]. Since then, models based on bayesian statistics [6, 7], monte-carlo and fuzzy logic [8], and the machine learning algorithm Random Forrest (RF) [9–11] have been suggested. When applying statistical models they are typically calibrated to data whereas machine learning models are trained on data.

The type of predictions performed by the models can be split into two categories, which relate to estimating the probability of the sewer to be in each of the condition states included in the model [4, 12, 13], and to predict the exact condition state of the sewer [6–11, 14]. This paper will focus on the models which predict the exact state of the sewers as it is easier to evaluate and compare the performance of these models.

A common feature for several inspection notations is that the sewers have been categorized into a number of classes, often five [6, 7, 9–11] or six [14]. It is hard to predict the specific state of the sewers, especially to distinguish between the two worst classes. This is i.e. shown by Harvey and McBean [9], who first predicted the sewer condition in five states, but chose to combine state 1-3 and state 4-5 in order to detect if the sewer is in an acceptable state or not. Combining states into binary categories has also been done by Fuchs-Hanusch et al. [14], Kabir et al. [6], and Laakso et al. [11], whereas Rokstad and Ugarelli [10] did not combine the states but evaluated the sensitivity and specificity according to finding sewers predicted to be in the worst condition.

The models are typically based on data from single cities [2, 4, 6, 8–10, 13, 15], which in some cases have been supplied by data from other places [7]. In contrast to this [11] used data from an area of 230 km^2 in the southern part of Finland. On the other hand, [16] focused on making a general approach for finding the best deterioration model for different utilities or municipalities.

Generally, deterioration models take in a number of factors affecting the sewers and predict the state of the sewer based on this. Hawari et al. [8] investigated which factors influence the deterioration by studying the literature and consulting experts in order to rate the importance of the different factors. The outcome was three groups of factors: Physical, operational and environmental factors. Physical factors relate to parameters such as age, pipeline dimension, material buried depth etc. Operational factors relate to i.e flow, content of the water and roots. The environmental factors relate to the surroundings such as soil type, location including traffic load, and ground disturbance such as construction work. [8] The number of features which significantly can be shown

to influence the deterioration process depends on the sample size [12] which means that the number of features which should be taken into consideration in a model depends on the amount of data available for calibration or training of the model.

In order to calibrate a deterioration model sufficiently a data set containing the variance of the sewer system is required. However, for municipalities or utilities which have not earlier used CCTV inspection on the sewers, it is very expensive to produce a sufficient data set [12, 16, 17]. Additionally, due to lack of data, the models are prone to over- and under-fitting: Some models does not include sufficient parameters to describe the underlying context and some include too many parameters according to the data available for calibration [12].

Despite the large number of models, there is still a great research potential within the area of deterioration models. As earlier mentioned most models are developed for a small area such as a city. This is problematic as the accuracy and bias of the model depends on the amount of data used to calibrate or train the model [10] and it is very expensive to generate a sufficiently large data set.

If a model could be applied to several areas, with no need for calibration in order for it to work, a large number of utilities could benefit from the model.

The contribution of this paper is to present a deterioration model which does exactly that, i.e. a general model. The model is based on data from 35 different utilities in Denmark.

A.2 Method

A data set containing the sewer condition and corresponding physical and environmental parameters from 35 different utilities spread across the the country of Denmark was used in order to develop a deterioration model for predicting the condition state of individual pipes. The metrics used to evaluate the condition of the sewers is called damage percentage.

A.2.1 Damage percentage and comparison to other metrics

The damage percentage is a standardized metric for evaluating the condition of the sewers, and it is followed by all utilities in Denmark. When a CCTV-inspection is performed in a sewer, all observations are noted down. The observation categories are: Water level, physical condition, operational condition and special constructions. Water level relates to the amount of water in the sewer. The physical condition includes cracks, surface damage, production error, deformation, staggered assembly of pipes and hanging assembly material. The operational condition includes roots, infiltration, deposits, coatings and obstacles. The special conditions relate to smaller pipes being connected to the main pipe including the condition and orientation of these, the quality of the assembly and changeover during construction change. Dependant on the type

and severity of the observations each observation contributes to the damage percentage with a certain amount divided by the length of the pipe. [18]. Due to the calculation method, it is possible to have a damage percentage above 100.

The damage percentage is a metric which is very hard to compare to the metrics used in other deterioration models, as these are often based on standards in which the condition is split into certain condition states. For this reason, the damage percentage was transformed to a logarithmic scale in which Damage percentage at 0-99 will be ranging from 0-10 by utilization of equation A.1. Damage percentages above 99 will result in a slightly higher value on this scale.

$$TS = 5 \cdot log_{10}(DP + 1) \tag{A.1}$$

In Equation A.1 TS is the transformed scale and DP is the damage percentage.

The benefit of transforming the damage percentage to a logarithmic scale ranging from 0-10 for damage percentages up to 99 is, that it can be further split into condition states matching condition states in other deterioration models. For instance, for comparison to a model with five condition states (CS), the scale is split into five CS as follow: If TS < 2 the pipe is in CS1. If $2 \leq$ TS < 4 the pipe is in CS2 etc. Likewise it can be split into a binary problem where TS < 6 is considered to be an acceptable state and TS ≥ 6 is considered to be an unacceptable state. Both scenarios are illustrated in Figure A.1.

A.2.2 Data set

In total the data set contains 146,856 CCTV inspection reports over sewer pipes and pipe sections. A pipe section here refers to pipes which contain obstacles preventing a CCTV inspection to continue from the current position and thereby split the inspection of the pipe into sections. For simplicity pipe sections are from here referred to as pipes. All the pipes in the data set have been CCTV-inspected and based on this inspection the damage percentage have been calculated. In addition to the damage percentage, the provided data set contains information on material, construction year, depth, size, type of sewer, elevation, soil type, position according to city, industry, trees, buildings, roads etc. In total there are 47 variables in the data set of which 35 are binary and 12 are continuous. An overview of the binary variables can be found in Table A.1, and an overview of 10 of the continuous variables can be seen in Table A.2. In addition to the continuous variables presented in Table A.2 the data set contained the positions of the sewers according to a UTM based coordinate system covering Denmark.

When transforming the TS into five CSs, as described in Section A.2.1 the distribution is as follows: CS1 = 28,993 pipes, CS2 = 23,571 pipes, CS3 = 38,771 pipes, CS4 = 37,953 pipes, and CS5 = 17,568 pipes.

This large data set enables utilization of machine learning approaches when modelling deterioration models. Furthermore, the large data set enables inclusion of a pro-

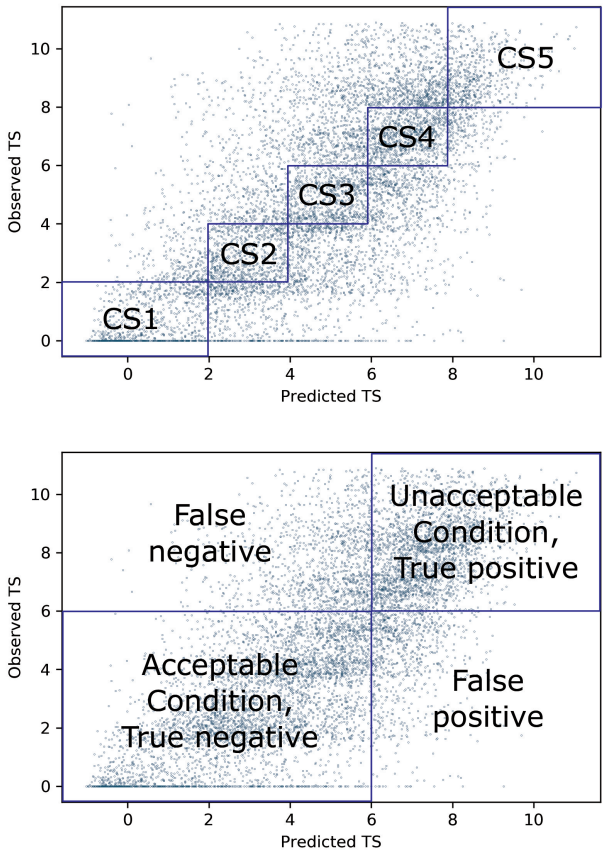


Fig. A.1: Illustration of how the TS can be split into a five CS problem (top) and a binary problem (bottom)

portionally high number of input variables in order to avoid over- and under-fitting of the model.

A.2.3 Choice of algorithm

As earlier described the data set contains both binary and continuous variables. As the machine learning algorithm Random Forest (RF) is tolerant to parameters having different scales, and has proven efficient when compared to the statistical model called

Table A.1: Overview over binary parameters in the data set

Parameter	Percentage of the data set containing the parameter
Material	
Concrete	68.0 %
Plastic	21.0 %
Clay	1.8 %
Fully rehabilitated pipes with "Non-dig" methods	1.8 %
Other	7.4 %
Content	
Sewage	35.0 %
Rain water	29.4 %
Combined	35.5 %
Rehabilitation parameters	
Has any sewer in this position been rehabilitated (incl replacement)	3.7 %
Puncturate rehabilitation	0.0 %
Rehabilitation, but details unknown	0.0 %
Area type	
City zone	73.0 %
City center	1.8 %
Industrial area	8.6 %
Surrounding buildings	
High buildings	1.9 %
Low buildings	74.4 %
Surrounding Trees	
Distance to trees < 12 m	19.7 %
Distance to trees < 4 m	7.9 %
Soil type*	
DS	21.4 %
ES	3.1 %
FS	7.3 %
TS	12.6 %
HV	0.3 %
ML	42.9 %
SC	0.1 %
MS	3.6 %
DL	1.1 %
YS	3.1 %
HS	4.2 %
HG	0.6 %
Road classes	
local (tertiary)	31.5 %
local (secondary)	13.6 %
local (primary)	8.9 %
Transprotation road	4.3 %
Other road	0.9 %

*The parameters presented in this category refer to different earth types found in the underground.

Table A.2: Overview over continuous parameters in the dataset

Parameter	Mean	Std
Age	32.9 years	17.2 years
Dimension	311.8 mm	195.8 mm
Length	43.8 m	23.0 m
Depth	2.4 m	0.9 m
Year of construction	1975.2 year	16.9 years
Terain	31.0 m	24.7 m
Year of rehabilitation (Year of construction if not rehabilitated)	1976.5 year	17.9 year
Number of buildings	4.0 buildings	14.8 buildings
Number of grids	1.9 grids	5.7 grids
Distance to road center	86.8 m	31.1 m

GompitZ [10], developed by Gat [19], and linear regression models [11], it was chosen to use Random Forest Regressor (RFR). A RFR algorithm consists of a number of decision trees which each makes a prediction of the output. The outcome of the RFR algorithm is the average of all the decision trees. For more information on RF algorithms see Breiman [20].

As it might be relevant to look into other algorithms in addition to RFR 14 other methods were initially tested in order to decide if other algorithms should be used as parallel to RFR or in combination with RFR. These were Bagging [21] with Decision Tree, Bagging with Random Forest, K-Nearest Neighbors with uniform weights, K-Nearest Neighbors with distance weights, Linear Ridge, Lasso Regression, Multi-layer Perception Regressor, AdaBoost [22] with Decision Tree, Recursive Feature Elimination by using Gradient Boosting [23], Principal Components Regression, Support Vector Regression [24], Naive Bayes [25], Quadratic Discriminant Analysis, and Extra Trees [26]. In all cases the hyper parameters were manually adjusted in order to obtain the best result for each method. As Random Forest, Bagging with Decision Tree and Bagging with Random Forest showed the best results a combination of these methods were tested by using the mean prediction of the three methods as the result. All tests were carried out using the Python library Scikit-learn [27].

RFR gave the best results for the individual algorithm, and similar results to the combined method: When training the models several times, the combined method would in some cases give a slightly better result than the RFR, however as the combined method was more complex than the RFR, it was chosen only to use RFR, but to train the RFR 5 times and use the better model.

The RFR used for predicting the TS consisted of 177 decision trees with a maximum depth of 26. Furthermore, the maximum number of features to consider for a split was

set to 71 % of the total number of features. Mean square error was used as criterion for measuring the quality of the split. The parameters were chosen based on testing different combinations of variables.

When using RF to predict sewer conditions Rokstad and Ugarelli [10] observed a tendency to predict towards the mean. In order to reduce this tendency we found the straight line best describing the relation between the true value of the training set and the predicted value of the training set. The displacement between this line and the optimal line, which intersects (0,0) and (10,10) represents a bias for each point as illustrated in Figure A.2. Therefore the displacement was added to every prediction performed on the test set.

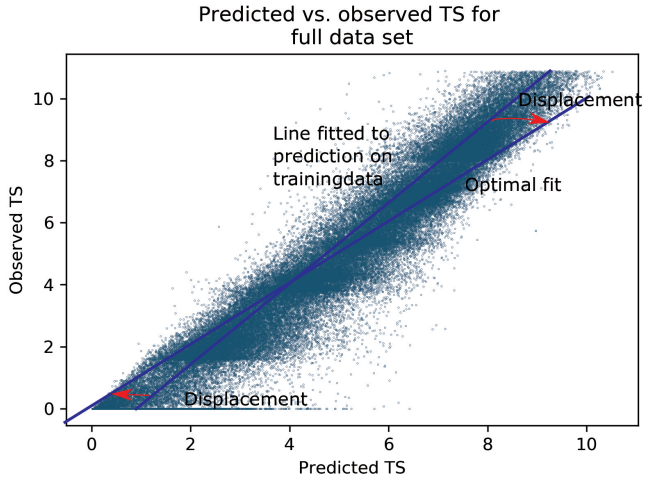


Fig. A.2: Example of how the displacement of the prediction is found

A.2.4 Training of models

Three variants of the RFR were trained. In the first variant the full data set was used. In the second variant only a part of the data set was used in order to account for an uneven distribution of the sewers' condition. In order to generate this data set the TS was split into five condition states as described in Section A.2.1. As fewest pipes were in CS5 (17,568 pipes), an equally number of sewers from the other CS's were randomly selected resulting in a data set containing 87,840 pipes. We call this data set an Approximately Equally Distributed data set, from here on refereed to as the AED-data set. The third variation of the algorithm was trained on the AED-data set, but

the geographical position of the pipes was removed from the data set, in order to test how dependent this methodology is on the geographical position of the sewers.

For training the models the three data sets were randomly split in a training set containing 90 % of the pipes and a test set containing 10 % of the pipes. Hereafter, for all the variants the RFR was trained five times and the best of these models were selected. The best models were in this case defined as the ones which predicted the TS with the lowest Mean Absolute Error.

A.2.5 Evaluation of the models

In order to compare our models to excising models it was investigated how other models have been evaluated. A commonly used method for evaluating deterioration models is evaluating how many sewers are categorized in each group, which is i.e. done by Caradot et al. [15]. This type of evaluation is not considered for comparison in this article, as it does not tell how well the model performs on single pipes. Other authors have made confusion matrices for a five class problem [9, 10] and several authors who originally had data for a five or six class problem have stated that pipes which are in one of the two highest states are in an unacceptable condition whereas the rest of the pipes are in an acceptable condition in order to make a binary model [6, 9, 11, 14]. For this reason our model is evaluated as a five class problem and as a binary model.

When working with binary models the balance between sensitivity and specificity can be adjusted. Laakso et al. [11] chose to fix the sensitivity to 0.80 which was a compromise between the utility wishing a high sensitivity and avoiding a too low specificity.

With inspiration from this we also test the specificity for a two class problem when the sensitivity is fixed on 0.80. As our model is based on regression, the balance between sensitivity and specificity can be adjusted by considering predictions of the TS in a certain range below 6 as being in an unacceptable state. This coresponds to moving the vertical line shown in Figure A.3 towards the left as illustrated by the red arrows.

A.3 Results

The predicted TS compared to the observed TS for all three models is illustrated in Figure A.4.

The confusion matrices for the three data sets when considering a five class problem can be seen in Table A.3.

The confusion matrices for the three data sets when considering a binary problem can be seen in Table A.4 for both the original models and when the sensitivity is fixed at 0.80. The corresponding sensitivities and specificities can be seen in Table A.5.

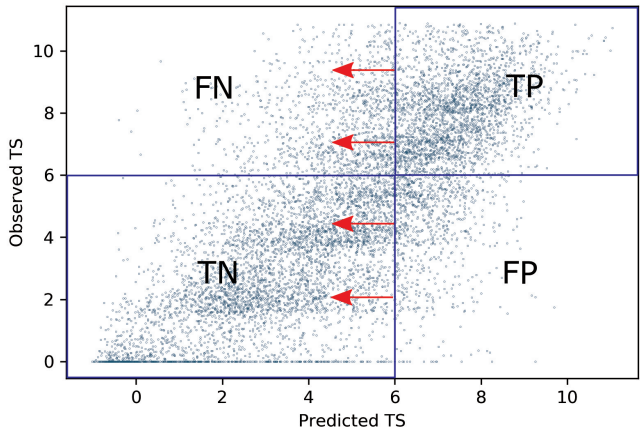


Fig. A.3: Illustration of how the balance between sensitivity and specificity can be adjusted when considering a binary problem. FN = false negative, TP = true positive, TN = true negative and FP = false positive

A.4 Discussion

As described in section A.1 CCTV inspections are important in order to monitor the condition of sewers. However, as CCTV inspections are expensive, it is important to chose the correct sewers for inspection. Due to this several deterioration models have been developed. Several different approaches have been used in order to evaluate these models. An example of this is shown by Caradot et al. [15] who considered a four class problem and evaluated how many sewers were categorized in each class. This type of evaluation is not considered for comparison in this article, as it does not tell how well the model performs on single pipes. Another example is Elmasry et al. [7] who evaluated the Mean Absolute Error and Root Mean Square Error. A more comparable evaluation was performed by Harvey and McBean [9] and Rokstad and Ugarelli [10] who made models for predicting a five class problem. Both authors showed confusion matrices for their predictions on a test set. In order to compare our model to these two models the number of true predictions for each class was calculated and can be seen in Table A.6.

As seen in Table A.6 Harvey and McBean [9] are very good at predicting the sewers in SC1, but only predict up to 28.1 % of the sewers in the other states. In contrast to this Rokstad and Ugarelli [10] had a more even performance for the five CSs, and are actually quite good at finding sewers in SC5 compared to Harvey and McBean [9] and our models. However, it is worth noticing that Rokstad and Ugarelli [10] in all cases had

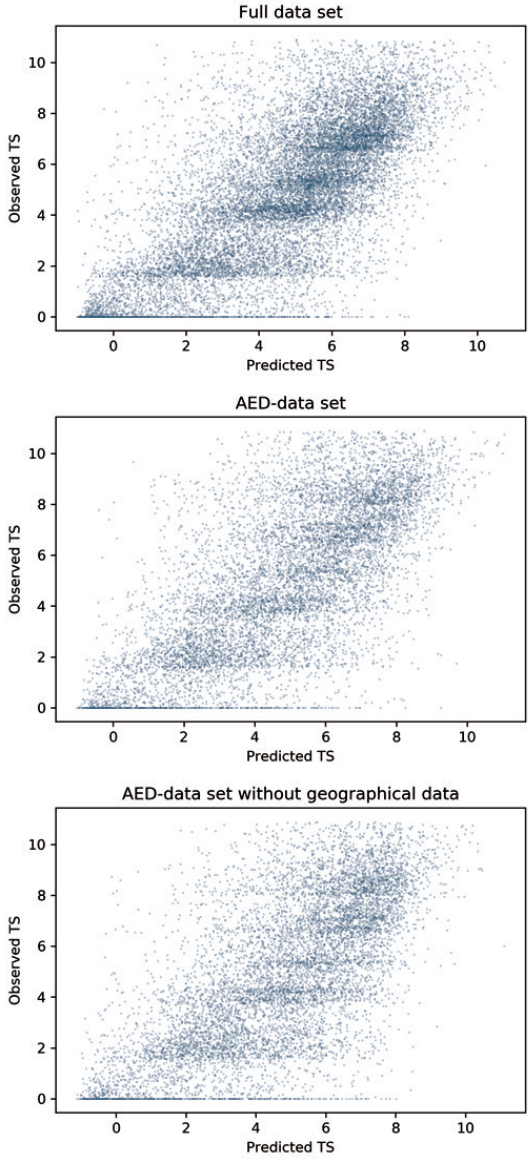


Fig. A.4: Observed vs predicted TS for training the RFR on the three data sets

Table A.3: Confusion matrices for the predictions performed by each of the three models when considering a five class problem

		Predicted - Full data set					
		CS1	CS2	CS3	CS4	CS5	Total
6*Observed	CS1	1641	818	388	61	1	2909
	CS2	352	1019	814	168	6	2359
	CS3	123	742	2136	935	29	3965
	CS4	39	230	1172	2079	166	3686
	CS5	19	99	394	926	329	1767
	Total	2174	2908	4904	4169	531	14686

		Predicted - AED-data set					
		CS1	CS2	CS3	CS4	CS5	Total
6*Observed	CS1	939	535	228	68	7	1777
	CS2	263	764	537	171	12	1747
	CS3	63	357	789	485	34	1728
	CS4	22	138	524	939	168	1791
	CS5	19	84	370	802	466	1741
	Total	1306	1878	2448	2465	687	8784

		Predicted - AED-data set without geographical information					
		CS1	CS2	CS3	CS4	CS5	Total
6*Observed	CS1	929	502	279	80	3	1793
	CS2	263	741	566	160	8	1738
	CS3	65	351	833	487	23	1759
	CS4	23	147	526	948	98	1742
	CS5	15	99	389	951	298	1752
	Total	1295	1840	2593	2626	430	8784

a notable higher percentage of sewers being predicted with an error of more than 2 CSs from the observed CS than our model. An example can be seen when calculating the number of sewers predicted to be in CS1 but actually being in CS5. For Rokstad 2015 6 % of the sewers in CS5 were predicted to be in CS1, which is only the case for 1 % for our model. Despite the fact that it is difficult to compare this kind of observations, it is important to consider as the deterioration models often are integrated as a part of an asset management tool where the risk combined with the predicted CS is used to decide which sewers to inspect. If a sewer in CS5 is predicted to be in CS4 or CS3 this might be considered for inspection if it is a high-risk sewer, whereas this might not be the case if it is predicted to be in CS1.

As it is hard to estimate the exact CS Harvey and McBean [9], Fuchs-Hanusch et al. [14], Kabir et al. [6], and Laakso et al. [11] have combined CSs in order to make

Table A.4: Confusion matrices for the three data sets when considering a binary problem

	Observed condition	Predicted	
		Acceptable	**Unacceptable**
Original models			
Full data set	Acceptable	8033	1200
	Unacceptable	1953	3500
AED-data set	Acceptable	4475	777
	Unacceptable	1157	2375
AED-data set without	Acceptable	4529	761
geographical info	Unacceptable	1199	2295
Sensitivity fixed at 0.80			
Full data set	Acceptable	6895	2338
	Unacceptable	1091	4362
AED-data set	Acceptable	3986	1266
	Unacceptable	706	2826
AED-data set without	Acceptable	3916	1374
geographical info	Unacceptable	700	2794

Table A.5: The sensitivity and specificity for the three data sets

	Sensitivity	**Specificity**
Original models		
Full data set	0.64	0.87
AED-data set	0.67	0.85
AED-dataset without geographical info	0.66	0.86
Sensitivity fixed at 0.80		
Full data set	0.80	0.75
AED-data set	0.80	0.76
AED-dataset without geographical info	0.80	0.74

a binary problem. Harvey and McBean [9], Kabir et al. [6], and Laakso et al. [11] all used inspection data which had five CSs and chose to split it such that the first three CSs were considered as acceptable states and the two worst states were considered as unacceptable states. Fuchs-Hanusch et al. [14] used inspection data which had six CS's and considered the best four states as acceptable and the worst two as unacceptable. A comparison between results from the binary deterioration models and our model can

Table A.6: Comparison of models developed for a five class problem

	Harvey and McBean 2014	**Rokstad 2015 (GompitZ*)**	**Our model on AED-data set**
True class 1	90.1 %	55.9 %	52.8 %
True class 2	10.2 %	24.8 %	43.7 %
True class 3	17.3 %	55.1 %	45.7 %
True class 4	28.1 %	51.1 %	52.4 %
True class 5	0 %**	40.4 %	26.8 %
Average	29.1 %	45.5 %	44.3 %

*Rokstad 2015 presented an average for the GompitZ model when calibrating several times. **Only one sewer was in CS5.

be seen in Table A.7.

As seen in Table A.7 our model and Harvey and McBean [9] show very similar results. Harvey and McBean [9] show slightly better sensitivity than our models. On the other hand our models show a slightly better specificity, indicating that the difference between the two models can be explained by the choice of balance between specificity and sensitivity. For cementitous pipes Kabir et al. [6] show results similar to our original models, indicating that the difference between the performance of the models can be explained by the choice of balance between sensitivity and specificity. For clay pipes Kabir 2018 outperforms all the other models whereas for metalic and plastic pipes the number of bad pipes is too low for making a fair comparison. The performance of our model distinctly exceeds the performance of the models presented by Fuchs-Hanusch et al. [14] and Laakso et al. [11].

It is very important to point out that all the models used for comparison except for Laakso et al. [11] utilized data from a single city. When using this approach the models need to be calibrated or trained for all the different areas they are used in. This means that utilities with limited amounts of data either most settle for inferior models or invest in gathering a sufficiently large data set for calibration or training of a better model. In contrast to this our model is based on data from an entire country and hence does not need any recalibration etc. In other words, our model is a general model.

As shown in Table A.5 the model we trained on the AED-data set where the geographical position of the sewers was left out performed almost as good as the models for which the geographical position was included. This shows that deterioration models does not necessary need to be fitted to a single city and underlines the benefit of utilities collaborating on collecting and storing inspection data despite having different geographical positions and thereby also different environmental conditions.

When comparing the model trained on the full data set and the model trained on the AED-data set the model trained on the AED-data set is considered to have a slightly

Table A.7: Comparison between our model and existing models

	Sensitivity	Specificity	Size of test set (accept/unaccept)
Our model			
full data set	0.80	0.75	(9233/5453) pipes
AED-data set	0.80	0.76	(5252/3532) pipes
AED-data set without geo. info	0.80	0.74	(5290/3494) pipes
Harvey and McBean 2014	0.82	0.73	(318/38) pipes
Fuchs-Hanusch 2015	0.55*	0.73*	
Kabir 2018			
Cementitous	0.64	0.87	(1698/69) pipes
Clay	0.75	0.86	(414/68) pipes
Metallic	0.66	0.97	(157/3) pipes
Plastic	0.50	0.98	(129/4) pipes
Average	0.63	0.90	
Laakso2018			
RF model Original	0.50	0.80	∼112 km pipes**
RF model FNR fixed on 0.20	0.80	0.47	∼112 km pipes**

*Read from graph **This number has been calculated from the information that 30 % of 1241 km pipes have been inspected, and 30 % of these were used as test data.

better performance than the model trained on the full data set. However, due to the fact that RFR models are trained randomly, each new training sessions will give slightly different results, which means that no generalizable conclusions can be made based on this difference. The reason for the model trained on the AED-data giving similar results to the model trained on the full data set can be due to the amount of training data for both models being sufficient to show the tendencies in the data. However, it could also be that having an equally distributed data set compensates for having less data.

Another point worth noticing is that despite the tendency to develop binary models, it is still beneficial to use models predicting the condition in a continuous scale or in a certain number of condition states when considering selection of sewers for inspection. This is due to the fact that inspections are expensive, and if a high number of sewers are recommended for inspection the utility will have to choose between them. This is underlined by the fact that our model might not be able to find all the sewers in CS5 when considering a five class problem, but if a sewer is found to be in CS5 there is a

high probability of it actually being in this state.

A.5 Conclusion

This paper presents a Random Forest based deterioration model for sewer condition which shows results comparable to the best deterioration models available for comparison. The data used to develop our models is, in contrast to the other models with a high performance, collected from 35 utilities providing sewer service to different cities with different geography and different environmental conditions. This introduces the possibility of utilities collaborating on collecting a data set instead of producing a new data set for each utility, which can lead to savings for the individual utility.

In conclusion, we have showed that it is possible to make a deterioration model that generalizes across data from different regions, sewers and utilities. This is a significant improvement compared to the current situation where a new model needs to be learned for each new set of data.

References

[1] M. Elmasry, T. Zayed, and A. Hawari, "Defect-based ArcGIS tool for prioritizing inspection of sewer pipelines," *Journal of Pipeline Systems Engineering and Practice*, vol. 9, p. 04018021, nov 2018.

[2] K. Baah, B. Dubey, R. Harvey, and E. McBean, "A risk-based approach to sanitary sewer pipe asset management," *Science of The Total Environment*, vol. 505, pp. 1011–1017, feb 2015.

[3] M. Marzouk and A. Osama, "Fuzzy-based methodology for integrated infrastructure asset management," *International Journal of Computational Intelligence Systems*, vol. 10, no. 1, p. 745, 2017.

[4] A. G. Altarabsheh, *Managing Urban Wastewater System Using Complex Adaptive System Approach*. PhD thesis, PURDUE UNIVERSITY GRADUATE SCHOOL, 2015.

[5] E. V. Ana and W. Bauwens, "Modeling the structural deterioration of urban drainage pipes: the state-of-the-art in statistical methods," *Urban Water Journal*, vol. 7, pp. 47–59, feb 2010.

[6] G. Kabir, N. B. C. Balek, and S. Tesfamariam, "Sewer structural condition prediction integrating bayesian model averaging with logistic regression," *Journal of Performance of Constructed Facilities*, vol. 32, p. 04018019, jun 2018.

[7] M. Elmasry, A. Hawari, and T. Zayed, "Defect based deterioration model for sewer pipelines using bayesian belief networks," *Canadian Journal of Civil Engineering*, vol. 44, pp. 675–690, sep 2017.

[8] A. Hawari, F. Alkadour, M. Elmasry, and T. Zayed, "Simulation-based condition assessment model for sewer pipelines," *Journal of Performance of Constructed Facilities*, vol. 31, p. 04016066, feb 2017.

[9] R. R. Harvey and E. A. McBean, "Predicting the structural condition of individual sanitary sewer pipes with random forests," *Canadian Journal of Civil Engineering*, vol. 41, pp. 294–303, apr 2014.

[10] M. M. Rokstad and R. M. Ugarelli, "Evaluating the role of deterioration models for condition assessment of sewers," *Journal of Hydroinformatics*, vol. 17, pp. 789–804, sep 2015.

[11] T. Laakso, T. Kokkonen, I. Mellin, and R. Vahala, "Sewer condition prediction and analysis of explanatory factors," *Water*, vol. 10, p. 1239, sep 2018.

[12] M. Ahmadi, F. Cherqui, J.-B. Aubin, and P. L. Gauffre, "Sewer asset management: impact of sample size and its characteristics on the calibration outcomes of a decision-making multivariate model," *Urban Water Journal*, vol. 13, pp. 41–56, feb 2015.

[13] F. Tscheikner-Gratl, R. Sitzenfrei, W. Rauch, and M. Kleidorfer, "Integrated rehabilitation planning of urban infrastructure systems using a street section priority model," *Urban Water Journal*, vol. 13, pp. 28–40, jul 2015.

[14] D. Fuchs-Hanusch, M. Günther, M. Möderl, and D. Muschalla, "Cause and effect oriented sewer degradation evaluation to support scheduled inspection planning," *Water Science and Technology*, vol. 72, pp. 1176–1183, jun 2015.

[15] N. Caradot, H. Sonnenberg, I. Kropp, A. Ringe, S. Denhez, A. Hartmann, and P. Rouault, "The relevance of sewer deterioration modelling to support asset management strategies," *Urban Water Journal*, vol. 14, pp. 1007–1015, may 2017.

[16] H. Park, S. H. Ting, and H. D. Jeong, "Procedural framework for modeling the likelihood of failure of underground pipeline assets," *Journal of Pipeline Systems Engineering and Practice*, vol. 7, p. 04015023, may 2016.

[17] X.-X. Yuan, "Principles and guidelines of deterioration modelling for water and waste water assets," *Infrastructure Asset Management*, 2016.

[18] DANVA, *Fotomanualen Beregning af fysisk indeks ved TV-Inspektion.* 1 ed., 2005.

[19] Y. L. Gat, "Modelling the deterioration process of drainage pipelines," *Urban Water Journal*, vol. 5, pp. 97–106, jun 2008.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[21] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123 – 140, 1996.

[22] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *journal of computer and system sciences*, vol. 55, pp. 119–139, 1997.

[23] J. H. Friedman, "Stochastic gradient boosting," 1999.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[25] H. Zhang, "The Optimality of Naive Bayes," *American Association for Artificial Intelligence*, 2004.

[26] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, pp. 3–42, mar 2006.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, 2011.

Paper B

Sewer Deterioration Modeling: The Effect of Training a Random Forest Model on Logically Selected Data-groups

Bolette D. Hansen, Søren H. Rasmussen, Thomas B. Moeslund, Mads Uggerby, and David G. Jensen

The paper has been published in the
Procedia Comput. Sci. Vol. 176 pp. 291–299, 2020.

© 2020 The Authors
The layout has been revised.

Abstract

Breakdown of sewers can induce significantly damage to roads and buildings placed upon it. For this reason, timely maintenance of the sewer system is essential. However, due to the under-ground position of the sewers they are very expensive to monitor, as this is done by CCTV inspection. Therefore, it is important to choose the right sewers for inspection and several decision-support tools have been developed to help the operators to select which sewers to inspect. These decision support tools all contain a model which predicts the condition of the sewers, and recently several models have been proposed in order to increase the performance. The scope of this paper is to investigate the effect of training a Random Forest model on logically selected groups of data, as opposed to training of a joined model on the full data set. The selected data groups were based on expert knowledge: The first data groups were based on the sewer material (concrete, plastic, clay, reinforced with lining and other material). The concrete data set was then further sub-divided into wastewater types (sewage, rain and combined) whereas the plastic data set was sub-divided into road classes. The results showed that the model trained on the full data set performed better than the models trained on logically selected data-groups as it encounters the heterogeneity of the data set. Furthermore, this answers an important question raised by end users of the deterioration models.

B.1 Introduction

The sewer system is an essential infrastructure, but it is very expensive to build and maintain. However, maintenance of the system is important in order to account for the ageing sewer system and maintain functionality. If a sewer is not maintained in time it can break down and induce damage to roads and buildings placed upon it. Replacing the sewers after break-down have shown more expensive than to perform intime rehabilitation or replacement of the sewer [1]. However, due to the under-ground position of the sewers they are hard to monitor

Monitoring of the sewer system is often performed by Closed Circuit Television inspection (CCTV inspection). CCTV inspection is performed by sending a robot with camera through the sewer system and manually annotating all the events present in the sewers. This is very time consuming and thereby expensive for which reason effort is put into automate the inspection technics [2]. However, there is still a long way before these techniques can fully replace manual inspection. Therefore, the utilities need to prioritize, which sewers to inspect. In order to help the utilities identifying, which sewers to prioritize several decision-support tools have been developed [3–6]. These tools typically consist of a deterioration model, predicting the condition state of the sewer or the probability of the sewer to be in a certain condition and a consequence model for determining the criticality in case of sewer failure. By combining them the

users have a risk based prioritizing tool. One decision support tools also includes an economic model for prioritizing the sewers [5].

Where the consequence models are based on easily accessible data such as distance to buildings, industry areas, hospitals, schools, beaches and roads, the deterioration of the sewers is harder to determine. It depends on several factors including both environmental, operational, hydraulic and physicality of the sewers. In order to optimize the prediction of the condition several deterioration models have been developed or evaluated since 2018 [3, 5–16]. The approaches for modeling the deterioration of the sewers have traditionally been split in to deterministic, statistically and artificial intelligence or machine learning approaches [17]. The models developed today are typically either based on statistical or machine learning methodologies. Since 2018 the following statistical approaches have been utilized: Bayesian Network [6], Gompertz models [8, 9], Logistic Regression [10, 14], and Monte Carlo Makov Chain [7, 12]. Likewise, since 2018, the following machine learning approaches have been utilized: Random Forest [8, 10, 14, 16], other forest based methods [13], Ant Colony Optimization [3], and Support Vector Machine [11].

Some of the methods have been tested in a way that makes them comparable with others by e.g. the sensitivity and specificity of the results [14–16, 18] or the number of true positive, true negative, false positives and false negatives [8, 10]. However, it is still hard to compare the different models as one might prioritize a high sensitivity over the specificity and opposite. Comparison of the models is further complicated by the privacy of the data sets. This means that the models are developed based on data sets with different ratio between good and bad pipes. An example of how the different data sets can give different results is presented by [10] who tested different models on both a German data set and at Columbian data set. Likewise, the deterioration of the pipes varies as for instance concrete pipes are very sensitive to corrosion due to sewerage [19]. In many cases one model is trained or calibrated to all the sewer pipes at the same time [6–8, 10, 11, 14, 16]. However [15] made Bayesian Logistic Regression models for concrete, clay, metallic and plastic pipes respectively, while others focused on specific material groups such as plastic [12] or the material groups with the most pipes represented [13]. This leaves the question of whether it is beneficial to develop a deterioration model based on specific data groups such as material, wastewater type and road type, or if it is sufficient with one model for all data groups. The answer to this question leads to another question, which is, whether or not a model should be evaluated according to the performance of separate data groups or the overall performance of the model. In the cross-field between data science and material science, these questions are sometimes raised by the end users of the deterioration models, such as decision makers at utilities. Therefore, investigation of this will help aligning the expectations between the end users of the deterioration models and the developers.

The scope of this paper is to investigate the effect of developing several Random Forest models based on logical grouped data sets compared to developing one general

model for the whole data set. The use of machine learning methods in this field is still in its infancy and Random Forest is at present state of the art.

B.2 Method

B.2.1 Data and preprocessing

The data has been withdrawn from a database where 35 Danish utilities have entered the results of CCTV inspections performed on their sewer networks. Only pipes which allowed for extraction of all predicter and target variables were withdrawn from the database. Likewise, in order to avoid pipes with incorrect registrations only pipes with realistic registrations were withdrawn. Examples of suspicious data could be if a pipe did not fit into following spans: 0 years pipe age 169 years, 63 mm $<$ pipe dimension $<$ 3000 mm and 0.6 m $<$ pipe depth $<$ 10 m. furthermore, a datapoint would be considered suspicious if several pipes had the exact same damage percentage. The final data set consisted of 119.919 pipes.

The annotation of the occurrences in the CCTV inspections follow the Danish standard for CCTV inspection of sewers [20]. Hereafter, the damage percentage of the sewers was calculated. The damage percentage is a weighted evaluation of each occurrences' contribution to the total damage percentage [21]. It is worth noticing that, due to the way of calculating the damage percentage, it is possible to obtain a damage percentage above 100. Finally, the damage percentage is transformed to a continuous scale ranging from 0-10. The formula for this is inspired by the Danish standard for calculating the physical index [21] and can be seen in equation B.1.

$$TS = 5 * log(DP + 1) \tag{B.1}$$

In equation B.1 TS refers to the transferred scale and DP refers to the damage percentage. The predictor variables available is the same as those presented by [16] which in general terms relate to: Dimension and length of the pipes, material type, wastewater type, if and how the sewer have been rehabilitated, surrounding areas (industry, city etc.), surrounding buildings and trees, soil types, road classes and position of the pipes. In addition to these parameters the slope and the ground water level have been found. As the slope is not accessible for all pipes, this data set contains a little less data than the data set used for [16]. The data set consists of both binary and continuous variables.

B.2.2 Logical data groups

The logical data groups in this study was found by consultation with construction engineers with several years of experience within sewer management. It was concluded that the main separation parameter should be the material. Due to a high number of

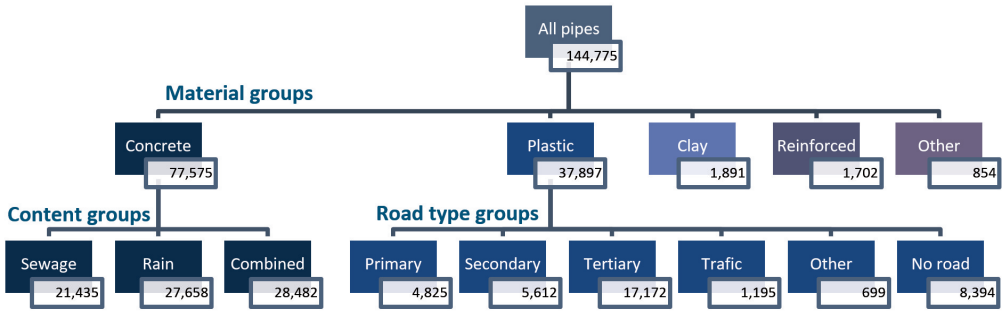


Fig. B.1: Overview of the investigated data groups and the corresponding number of pipes.

concrete and plastic sewers, these data groups were further split into sub-groups. For concrete pipes the sub-groups were based on the wastewater types running in the sewers (sewerage, rain and combined) and for the plastic pipe the sub-groups were based on the road type placed upon them (traffic road, primary road, secondary road, tertiary road and other road). An overview of all the logical data groups investigated in this study and corresponding number of pipes within the data group can be seen in Fig. B.1

B.2.3 Models

The model used in this study is a Random Forest model with the same settings as described in [16]. A benefit of using a Random Forest model is that it can handle data sets with both binary and continuous variables. Before training the model the data set for each data group is randomly split with 90 % for training and 10 % for test. The model is trained to predict the TS for each of the data groups. The model was set up using scikit-learn in Python [22].

B.2.4 Test

The models are trained to predict the TS, however, in order to compare the results with other methods the sensitivity and specificity of the models should be calculated. Furthermore, the precision of the models was calculated as this is important when the decision maker must prioritize which sewers to inspect. Another benefit of calculating the precision is that it can be used to account for the mixed distribution of good and bad pipes in different data groups when evaluating the models. In order to calculate the sensitivity, specificity and precision of the models all pipes with a TS 6 was considered to be in bad condition, and all pipes with a TS $<$ 6 was considered to be in good condition. As earlier described the performance of different models can be very hard to compare due to the prioritization of sensitivity versus specificity. For this reason,

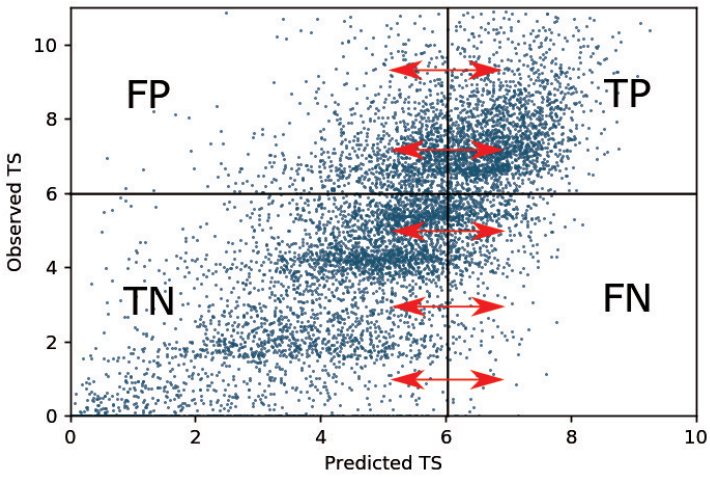


Fig. B.2: Illustration of how the balance between sensitivity and specificity can be adjusted by changing the splitting point for the predictions. The figure is inspired from [16]

the sensitivity of the models was fixed to 0.80 for testing purposes. 0.80 is based on a compromise presented by [14]. In order to adjust the sensitivity of a regression model the split between which pipes are predicted as being in a bad condition and which are predicted to be in a good condition can be changed till the sensitivity hits 0.80 as illustrated in Fig. 2. Finding the sensitivity of 0.80 was done using the Brents algorithm or Golden search for scalar optimization in the Python library SciPy [23]. If the two search methods came up with different sensitivity the one closest to 0.80 was chosen. If the sensitivity was too far away from 0.80, indicating that a local minimum was found for both methods, boundaries for the TS were set in order to find the correct minima. All the models were trained 10 times, and the mean and standard deviation for each data group was calculated.

In order to compare the total performance of the models trained on the logically grouped data sets with the model trained on the full data set the weighted sensitivity, specificity and precision were found. This was done for each level of data groups presented in Fig. B.2.

The formulas for calculating the weighted sensitivity can be seen in equation B.2.

$$WSens = \frac{TP_{DG1} + TP_{DG2} + ... + TP_{DGn}}{TP_{Tot} + FN_{Tot}} \tag{B.2}$$

In equation B.2 the $WSens$ is the weighted sensitivity, TP_{DG} is the number of true

positives in each data group, TP_{Tot} is the total number of true positive predictions and FN_{Total} is the total number of false negative. The formula for calculating the weighted specificity can be seen in equation B.3.

$$WSpec = \frac{TN_{DG1} + TN_{DG2} + ... + TN_{DGn}}{TN_{Tot} + FP_{Tot}} \tag{B.3}$$

In equation B.3 $WSpec$ is the weighted specificity, TN_{DG} is the number of true negatives, TN_{Tot} is the total number of true negative and FP_{Total} is the total number of false positive. The formula for calculating the weighted precision can be seen in equation B.4.

$$WPrec = \frac{TP_{DG1} + TP_{DG2} + ... + TP_{DGn}}{TP_{Tot} + FP_{Tot}} \tag{B.4}$$

In equation B.4 $WPrec$ is the weighted precision.

B.3 Results

The predictions for respectively the general model and the models trained on pipes within a specific material group can be seen in Fig. B.3. From the figure it can be seen where the splitting point should be placed in order to obtain a sensitivity at 0.80 for each of the models.

An overview of the specificity and precision for each model can be seen in Table B.1. This table also contains information on the performance of the general model for each data group when the splitting point is adjusted to the specific data groups. The model trained on the full data set obtained a sensitivity at 0.78 ±0.01 when the specificity is fixed at 0.80. The weighted specificity for the models trained on the logically grouped data set was calculated to be 0.73 ± 0.00. Furthermore, the general models obtain a precision of 0.62 ± 0.01 and thereby outperform the models trained on the material specific data sets as these obtain a precision of 0.57 ± 0.00.

Table B.1: Overview of the performance of the general model, the general model's performance on each material group and the performance of the models trained on material specific data when the sensitivity is fixed at 0.80. The mean and standard divination have been found by training each model 10 times.

	General model				Models trained on material specific data			
	Sensitivity	↑Specificity	↑Precision	Splitting point	Sensitivity	↑Specificity	↑Precision	Splitting point
All pipes	0.80 ± 0.00	0.78 ± 0.01	0.62 ± 0.01	5.18 ± 0.03	0.80 ± 0.00	0.73 ±0.00	0.57 ± 0.00	-
Concrete	0.80 ± 0.00	0.69 ± 0.02	0.65 ± 0.01	5.39 ± 0.02	0.80 ± 0.00	0.69 ± 0.01	0.65 ± 0.01	5.38 ± 0.02
Plastic	0.80 ± 0.00	0.82 ± 0.00	0.24 ± 0.02	2.86 ± 0.21	0.80 ± 0.01	0.83 ± 0.02	0.25 ± 0.02	2.93 ± 0.18
Clay	0.80 ± 0.01	0.60 ± 0.01	0.66 ± 0.07	5.44 ± 0.23	0.81 ± 0.02	0.59 ± 0.11	0.65 ± 0.07	5.63 ± 0.28
Reinforced w. lining	0.80 ± 0.01	0.52 ± 0.04	0.44 ± 0.05	4.10 ± 0.20	0.82 ± 0.03	0.55 ± 0.03	0.45 ± 0.04	4.08 ± 0.12
Other material	0.82 ± 0.05	0.76 ± 0.06	0.44 ± 0.08	3.68 ± 0.51	0.80 ± 0.05	0.80 ± 0.09	0.36 ± 0.10	3.33 ± 0.42

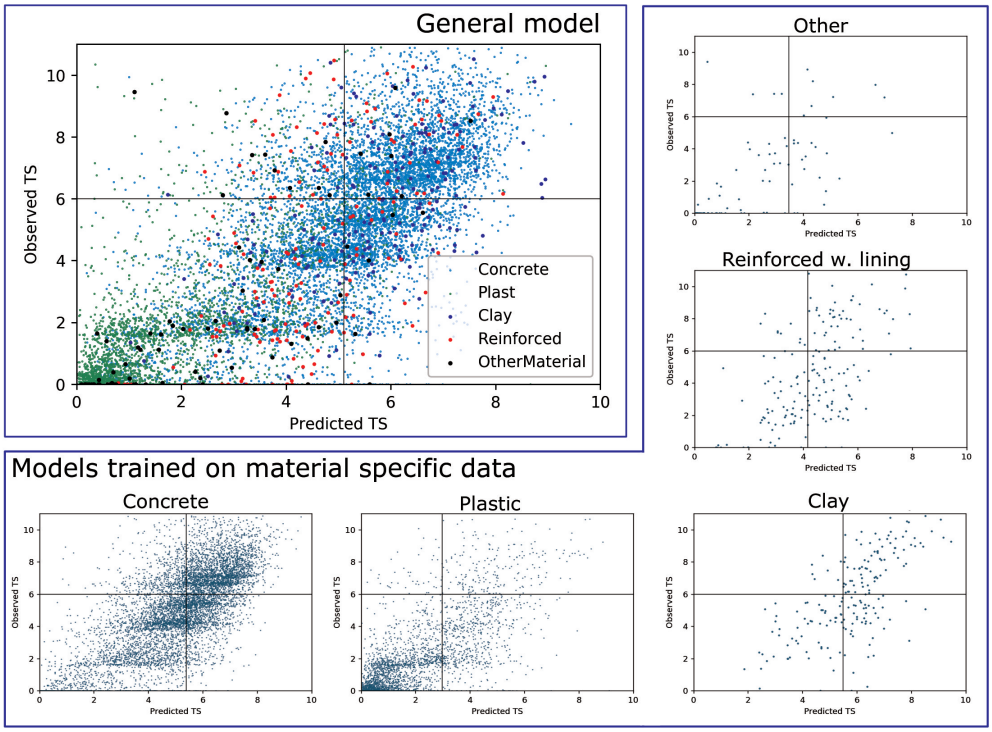


Fig. B.3: Results for the general model trained on all the pipes and for the models trained on material specific data. The vertical lines in the graphs show the position of the splitting point in order to obtain a sensitivity of 0.80.

B.3.1 Sub-division of the concrete group

An overview of the performance of the model trained on all the concrete pipes and the models trained on respectively the concrete pipes containing sewage, rain and combined water can be seen in Table B.2. From the table it can be seen that the model trained on the concrete data set and models trained on subgroups of the concrete data set performs similar as they obtain a specificity of respectively 0.69 ± 0.00 and 0.69 ± 0.01 and a precision of respectively 0.65 ± 0.01 and 0.65 ± 0.00.

Table B.2: Overview of the performance of the concrete model, the concrete model's performance on each wastewater group, and the performance of the models trained on concrete pipes containing sewage, rainwater and combined water respectively when the sensitivity is fixed at 0.80. The mean and standard divination have been found by training each model 10 times.

	Concrete model				Models trained on concrete and wastewater data			
	Sensitivity	↑Specificity	↑Precision	Splitting point	Sensitivity	↑Specificity	↑Precision	Splitting point
All concrete pipes	0.80 ± 0.00	0.69 ± 0.01	0.65 ± 0.01	5.38 ± 0.02	0.80 ± 0.00	0.69 ± 0.00	0.65 ± 0.00	-
Sewage	0.80 ± 0.00	0.69 ± 0.02	0.66 ± 0.02	5.49 ± 0.05	0.80 ± 0.00	0.69 ± 0.02	0.67 ± 0.01	5.50 ± 0.05
Rain	0.80 ± 0.00	0.72 ± 0.01	0.60 ± 0.01	5.12 ± 0.04	0.80 ± 0.00	0.71 ± 0.01	0.59 ± 0.01	5.06 ± 0.04
Combined	0.80 ± 0.00	0.65 ± 0.02	0.67 ± 0.01	5.51 ± 0.04	0.80 ± 0.00	0.65 ± 0.01	0.67 ± 0.01	5.51 ± 0.03

B.3.2 Sub-division of the plastic group

It has turned out that it is not possible to make a fair evaluation of whether it is beneficial to split the plastic pipes into road groups with the amount of data available for this study. This is because that only very few bad pipes are available for some of the groups. Especially the category 'other road' is sensitive to this and in one case only one bad pipe was present in the test set. In 'other cases', it was discovered that when the sensitivity in one test was fixed at 0.80 the specificity could easily vary with 0.10 due to the small amount of bad pipes in the data group.

B.4 Discussion

Sewer management is a very hot topic with several new models being published since 2018. In this paper we investigated if it would be beneficial to develop one general model trained on a data set containing all the available types of pipes or if it would be more beneficial to split the data set into sub-data sets based on logically selected groups of data.

As seen in Table B.1. there is no significant difference in the performances of the general model and the models trained on material specific data sets, when considering different material types and corresponding splitting points, as the variation in specificity and precision is within one standard deviation. The similar performance of the two approaches is underlined by the facts that 1) the splitting points for the two approaches

are within one standard deviation when considering the different material groups and 2) the distribution of the predictions in Fig. B.3 looks similar.

When considering the overall performance of the models, the general model, with a specificity at 0.78 ± 0.01 and a precision at 0.62 ± 0.01, significantly outperforms the models trained on the material specific data sets which obtains a specificity at 0.73 ± 0.00 and a precision at 0.57 ± 0.00. The reason for this is that the general model allows for a sensitivity > 0.80 on e.g. the concrete pipes and < 0.80 on e.g. the plastic pipes, and thereby encounters the fact that the plastic pipes generally are in a better condition than the concrete pipes. This is underlined by the fact that when using the general models' threshold, the sensitivity is 0.84 ± 0.00 for the concrete pipes and 0.30 ± 0.02 for the plastic pipes. Likewise, the specificity and precision for the concrete pipes, when using the general threshold, decreases to respectively 0.63 ± 0.01 and 0.62 ± 0.01 for the concrete pipes and increases to respectively 0.98 ± 0.00 and 0.54 ± 0.04 for the plastic pipes. Thereby the general model encounters the heterogeneity in the data, which is not accounted for when treating each data group individually.

The results for whether it makes sense to split the concrete data set into three sub-data sets for the wastewater types showed that it did not make significant difference. Likewise, it was investigated if it would be beneficial to split the plastic data set according to road type, however, the plastic pipes were in too good condition in order to make a fair comparison.

The good performance of the general model according to the models trained on the logically grouped data sets is most likely due to the modelling method used in this paper: Random forest is an ensemble of different decision trees. The decision trees are randomly grown, but weight is put on the most informative splits, so that the most influential parameters are likely to contribute to the trees [24].

A comparison of our results to the results presented in the literature for different data groups can be seen in Table B.3.

Table B.3: Comparison of our results to the results obtained in the literature. *The numbers were found by calculating the weighted sensitivity and specificity using equation 2 and 3 respectively. ** The precisions for the different data groups were calculated based on confusion matrices presented by the authors

	Our models (best versions)			Best results from Kabir 2018 [15]		
	Sensitivity	↑Specificity	↑Precision	↑Sensitivity	↑Specificity	↑Precision**
All pipes	0.80 ± 0.00	0.78 ± 0.01	0.62 ± 0.01	0.69*	0.88*	0.28
Concrete	0.80 ± 0.00	0.69 ± 0.01	0.65 ± 0.01	0.64	0.89	0.18
Plastic	0.80 ± 0.00	0.83 ± 0.02	0.24 ± 0.02	0.50	0.98	0.40
Clay	0.80 ± 0.01	0.60 ± 0.01	0.66 ± 0.07	0.75	0.86	0.44
Reinforced w. lining	0.80 ± 0.01	0.55 ± 0.03	0.44 ± 0.05	-	-	-
Metallic	-	-	-	0.67	0.97	0.33
Other material	0.80 ± 0.05	0.80 ± 0.09	0.44 ± 0.08	-	-	-

As seen in Table B.3 it can be hard to compare our results to the results obtained

by [15], as we have specified the sensitivity to be 0.80 whereas [15] has chosen the balance between sensitivity and specificity based on the form of the ROC curve. For clay pipes [15] performs better than our clay models do when considering specificity whereas our models have a better precision. Regarding the performance on the plastic pipes it is hard to compare our model to [15] as they only have four plastic pipes in bad condition. In all cases except for plastic, our models perform best regarding precision whereas Kabir has a higher sensitivity. The fact that we both obtain a higher sensitivity and precision than [15], while obtaining a lower specificity, indicates that our data set is less heterogeneous than the one used by [15]. The difference in heterogeneity in the data sets can be due to 1) the pipes in the data set used by [15] generally being in a better condition than the pipes in our data set or 2) that we use a broader definition of when a pipe is said to be in a bad condition than [15].

When comparing the results from the model trained on the full data set, without considering any sub-groups, our model performs slightly better than the best model presented by [16]. [16] used the same model setup as used in this article to obtain a specificity at 0.76 for their best model when the sensitivity was fixed at 0.80. This is most likely due to inclusion of the slope and groundwater level which were not included in [16]. Our model also outperforms [14] whom obtained a specificity of 0.47 when the sensitivity was set to 0.80.

B.5 Conclusion

This paper contributes to the general knowledge about development of deterioration models by investigating how the performance of the predictions is influenced by the training data set. It was investigated how the performance was affected when training on a full data set and when training on logically grouped sub-data sets. The results showed that there is no significant difference between the two approaches when considering their performance on specific data groups. Moreover, it was shown, that the general model performed better when considering the overall performance as it encounters the heterogeneity of the data.

References

[1] H. Korving, F. H. L. R. Clemens, and J. M. van Noortwijk, “Statistical modeling of the serviceability of sewage pumps,” vol. 132, no. 10, pp. 1076–1085.

[2] J. B. Haurum and T. B. Moeslund, “A survey on image-based automation of CCTV and SSET sewer inspections,” vol. 111, p. 103061.

[3] A. Altarabsheh, M. Ventresca, and A. Kandil, “New approach for critical pipe prioritization in wastewater asset management planning,” vol. 32, no. 5, p. 04018044.

[4] M. Elmasry, A. Hawari, and T. Zayed, "An economic loss model for failure of sewer pipelines," vol. 14, no. 10, pp. 1312–1323.

[5] M. Elmasry, T. Zayed, and A. Hawari, "Multi-objective optimization model for inspection scheduling of sewer pipelines," vol. 145, no. 2, p. 04018129.

[6] S. M. Ghavami, Z. Borzooei, and J. Maleki, "An effective approach for assessing risk of failure in urban sewer pipelines using a combination of GIS and AHP-DEA," vol. 133, pp. 275–285.

[7] N. Balekelayi and S. Tesfamariam, "Statistical inference of sewer pipe deterioration using bayesian geoadditive regression model," vol. 25, no. 3, p. 04019021.

[8] N. Caradot, M. Riechel, M. Fesneau, N. Hernandez, A. Torres, H. Sonnenberg, E. Eckert, N. Lengemann, J. Waschnewski, and P. Rouault, "Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in berlin, germany," vol. 20, no. 5, pp. 1131–1147.

[9] N. Caradot, M. Riechel, P. Rouault, A. Caradot, N. Lengemann, E. Eckert, A. Ringe, F. Clemens, and F. Cherqui, "The influence of condition assessment uncertainties on sewer deterioration modelling," vol. 16, no. 2, pp. 287–296.

[10] N. Hernández, N. Caradot, H. Sonnenberg, P. Rouault, and A. Torres, "Support tools to predict the critical structural condition of uninspected pipes for case studies of germany and colombia," vol. 13, no. 4, pp. 794–802.

[11] N. Hernández, N. Caradot, H. Sonnenberg, P. Rouault, and A. Torres, "Optimizing SVM model as predicting model for sewer pipes in the two main cities in colombia," in *New Trends in Urban Drainage Modelling*, pp. 926–931, Springer International Publishing, Sept. 2018.

[12] P. Lin, X. Yuan, and E. Tovilla, "Integrative modeling of performance deterioration and maintenance effectiveness for infrastructure assets with missing condition data," vol. 34, no. 8, pp. 677–695.

[13] T. Laakso, T. Kokkonen, I. Mellin, and R. Vahala, "Sewer life span prediction: Comparison of methods and assessment of the sample impact on the results," vol. 11, no. 12, p. 2657.

[14] T. Laakso, T. Kokkonen, I. Mellin, and R. Vahala, "Sewer condition prediction and analysis of explanatory factors," vol. 10, no. 9, p. 1239.

[15] G. Kabir, N. B. C. Balek, and S. Tesfamariam, "Sewer structural condition prediction integrating bayesian model averaging with logistic regression," vol. 32, no. 3, p. 04018019.

[16] B. D. Hansen, D. Getreuer Jensen, S. H. Rasmussen, J. Tamouk, M. Uggerby, and T. B. Moeslund, "General sewer deterioration model using random forest," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 834–841, IEEE.

[17] E. V. Ana and W. Bauwens, "Modeling the structural deterioration of urban drainage pipes: the state-of-the-art in statistical methods," vol. 7, no. 1, pp. 47–59.

[18] M. M. Rokstad and R. M. Ugarelli, "Evaluating the role of deterioration models for condition assessment of sewers," vol. 17, no. 5, pp. 789–804.

[19] X. Li, F. Khademi, Y. Liu, M. Akbari, C. Wang, P. L. Bond, J. Keller, and G. Jiang, "Evaluation of data-driven models for predicting the service life of concrete sewer pipes subjected to corrosion," vol. 234, pp. 431–439.

[20] B. Laden, DANVA, and Fotomanualgruppen, *Fotomanualen: TV-inspektion af afløbsledninger*. DANVA. OCLC: 769296284.

[21] Dansk Vand- og Spildevandsforening and DANVA, *Fotomanualen: beregning af fysisk indeks ved TV-inspektion.* Dansk Vand- og Spildevandsforening. OCLC: 488644072.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, "Scikit-learn: Machine learning in python," vol. 2011, no. 12, p. 28252830.

[23] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in python," vol. 17, pp. 261–272.

[24] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," vol. 40, no. 2, pp. 139–157.

Paper C

Comprehensive Feature Analysis for Sewer Deterioration Modeling

Bolette D. Hansen, Søren H. Rasmussen, Mads Uggerby, Thomas B. Moeslund, and David G. Jensen

The paper has been published in
Water Vol. 13, 819, 2021.
The original paper is accompanied by unpublished supplementary material, which contains in-depth information on the correlation of defects in topographically connected pipes. This supplementary material can be found in this thesis Part III Appendix G.

© 2021 by the authors.
The layout has been revised.

Abstract

Timely maintenance of sewers is essential to preventing reduced functionality and breakdown of the systems. Due to the high costs associated with inspecting a sewer system, substantial research has focused on sewer deterioration modeling and identification of the most useful features. However, there is a lack of consensus in the findings. This study investigates how the feature importance depends on the definition of bad pipes and how the feature importance changes between utilities with similar data bases. A dataset containing 318,457 pipes from 35 utilities with a condition state (CS) ranging from one to four was used. The dataset was cleaned, and a backward step analysis (BSA) was applied to two ways of binarizing the CS. Additionally, a BSA was applied for each utility with 100 pipes in CS four. The results showed that a selective definition of bad pipes reduced the performance and changed the order of which features contributed the most. In each case, either year of construction, age, groundwater, year of rehabilitation, or dimension was the most important feature. On average 6.5 features contributed to the utility-specific models. The feature analysis was sensitive to the inspection strategy, the size of the dataset, and interdependency between the features.

C.1 Introduction

A sewer system is a hidden but very expensive type of infrastructure to maintain [1]. Breakdown of a sewer can result in significant damage to roads and buildings. Furthermore, reduced functionality of the sewers can lead to flooding and exfiltration, for example, which can affect a number of externalities, such as property, traffic disruption, public health and the environment [1, 2]. For these reasons, the sewers' operators need to replace them in a timely manner, especially if the sewers are critical. However, sewers' underground location makes them difficult to monitor. Today, monitoring of the sewers is typically done by Closed Circuit Television inspection (CCTV inspection) [1]. CCTV inspection is done by manually sending a TV-inspection robot into the sewer and annotating all observations. As this is very time consuming, expensive, and imprecise due to a number of subjective factors [3, 4], much research has been put into automating these processes [1, 5]. However, full automatization of sewer inspection is not imminent. The high costs associated with CCTV-inspection forces utilities to prioritize which sewers to inspect. In Denmark, the paradigm for risk-based rehabilitation has been based on area. The areas which should be subject to CCTV-inspection were prioritized based on age of the pipes and the experience of the operators. Based on the findings in the CCTV-inspections, it was chosen whether an area should be rehabilitated or not. This has resulted in rehabilitation of pipes which could have been operational for several years as the inspection showed that the pipe might not be operational for the whole period until the next time the area would be chosen for rehabilitation. Today, for

economical optimization and better use of the pipes lift time, there is a trend toward risk based CCTV-inspection planning and rehabilitation on a pipe level.

Maintenance of sewer systems on pipe level entails new requirements for computer systems to keep track of the individual pipes, as the utilities now need to keep track of several tens of thousands of pipes instead of a limited number of areas. To assist the utilities in choosing which sewers to inspect, several decision support systems have been developed [6–9]. Usually these systems are risk models, consisting of a deterioration model and a consequence model. The deterioration models predict the condition of the sewers or the likelihood of a sewer's condition. The consequence models describe the severity of a potential sewer failure and can include economic, environmental, and social consequences [1]. Generally, the deterioration models suffer from low accuracy.

Development of sewer deterioration models is complicated by a high uncertainty in the data. This uncertainty is influenced, among other things, by subjectivity in the annotation of CCTV inspections, lack of data, and subjective selection of which pipes to inspect [3]. Dirksen et al. [4] found that defects with distinct features like roots were easy to find, while the probability of getting a false negative for other defect types varied around 0.25. The probability of a false positive was found to be around 0.04 [4]. Another issue often affecting the deterioration datasets is a lack of information [1], which results in low quality data. Furthermore, the datasets are affected by the fact that they have typically been collected for a specific purpose, such as quality assurance before asset handover or road renovation, diagnosis of malfunctioning and random inspections. This introduces a selective survival bias in the data [1]. Other factors that complicate deterioration modelling are that the datasets in general are highly skewed, both according to the number of pipes in the different classes and according to the predictor variables [10–12]. Furthermore, the size of the natural variability between sewers is unknown.

A large number of deterioration models have been developed; however, a lack of publicly available datasets due to privacy issues makes it difficult to compare the models [5]. Furthermore, the condition state (CS) is typically based on the local standard for CCTV inspection, which can be based on, for example, the European standard [13], Pipeline Assessment Certification Program [14], or a country specific standard [12, 15–17]. Moreover, in order to evaluate the deterioration models many authors tend to classify the multiclass or regression problem as a binary problem [12, 13, 18–20]. However, the model performance is very sensitive to how picky the evaluation is designed to be. For example, the performance of the precision and recall will increase if considering both pipes in the worst CS and the second worst CS as bad pipes, compared to considering only pipes in the worst CS to be in bad condition.

In addition, when deciding how to define the target variable the developer of the deterioration model needs to decide which predictor variables to use. Several methods have previously been used for parameter selection and the feature importance test. O'Reilly [21] in 1989 investigated the correlation between defects and individual param-

eters such as age, material, diameter, location, depth, wastewater type, soil type etc. in 180 km of sewers. Hansen et al. [11] investigated the potential benefits of developing deterioration models based on data groups defined by experts but found no improvement in model performance. Yin et al. [22] used a backward variable elimination process, through which they removed a parameter at a time and examined how the performance changed. Davies et al. [23] used a backward selection method and Laakso et al. [13] used the Boruta algorithm and found eight features to be influential.

Carvalho et al. [10] used eight different methods to investigate the feature importance and found that the different methods showed very different results. For example, if analyzing the features by removing the most significant features step by step, the importance of the other features will change, as there is often redundancy in the signal from the different predictor variables. This is not encountered when using the build-in feature analysis in Random Forest [10], however. Due to the uncertainty in the data, Roghani et al. [3] found that using the two or three most informative predictor variables was sufficient to build the deterioration model. However, using a deterioration model was better than just basing it on the inspection age.

Mohammadi et al. [24] reviewed 24 statistical and AI based papers on sewer deterioration. Nineteen of the reviewed papers provided information on whether a parameter was relevant. Nineteen features were considered, and none of the features were used in all the papers. Furthermore, none of the features considered relevant in more than three of the papers were considered relevant in all the papers. This illustrates a high variability in feature importance. Likewise, none of the features whose significance level was specified in more than one case were irrelevant in all the studies they were used in [24]. Finding the most significant features is important as accessing, extracting, and preprocessing each feature is very time demanding. In a review of deterioration models Hawari et al. [25] concluded that more work needs to be done to identify which data municipalities should collect in order to develop reliable deterioration models [25].

As described above, the performance of the deterioration models is affected by many conditions and a number of choices needs to be made for each model. This makes it possible to develop well performing models within academia. However, to create value, the models must meet the utilities' needs. For example, Guzmán-Fierro et al. [26] worked with a target variable ranging from 1 to 5 but developed a model that encountered only the pipes in CS 1 and CS 5. In reality, it is not possible to leave out the pipes in between, at least during the preliminary inspection.

In summary, sewer deterioration modeling has been a hot topic for the last two decades and myriad factors influence the performance of the models. Finding the optimal model cannot necessarily be done by selecting the model with the highest performance according to the literature. Likewise, there is a great deal of disagreement about which predictor variables are significant. The existing sewer deterioration models presented in the literature are characterized by large deviations in data, methodology, etc. Today researchers tend to perform feature analyses on single datasets. However,

a rarely touched perspective is the statistical variation in the features influencing the results when using similar datasets.

The contributions of this study are investigations of:

- The overall feature importance in a dataset containing information from several different utilities, including identification of potential drawbacks
- How the performance and feature importance of the models are affected by how the model developer has distinguished between good and bad pipes
- How the feature importance varies between utilities when the parameters in the datasets have been found in the same way for all utilities.

To the best of the authors' knowledge, this study provides the most comprehensive analysis of feature importance in sewer deterioration modeling and the first investigation of feature importance across several utilities with similar data bases. This information adds value to the process of developing deterioration models for utilities, which have a limited budget.

The following section of the paper, Section 2, provides a description of the data available, preprocessing, model selection, and the method used for feature importance. Section 3 contains three subsections, one for each of the contributions, while Section 4 contains a discussion of the key findings and comparisons to the literature. Section 5 contains a summary of the most important conclusions covered by the paper.

C.2 Materials and Methods

A dataset containing pipes from 35 utilities across Denmark was extracted from a common database for CCTV inspections. Pipes with suspicious values were not extracted. Examples of suspicious data points included those in which the following criteria were not met: 63 mm $<$ dimension $<$ 3000 mm, 0 years $<$ age $<$ 169 years and 0.6 m $<$ depth $<$ 10 m. Most of the inspections were performed from the start of the 1990s until today. The full dataset contains CCTV inspection from 318,457 pipes. For each pipe access to 24 different predictor variables was attempted; however, all predictor variables were only available for 196,174 pipes. An overview of the predictor variables can be seen in Table C.1.

All CCTV-inspections followed the Danish standard for CCTV inspections [27]. The inspections contained information on several observation types and corresponding severity of each. Based on the type of defect and its severity, the observations were categorized as CS 1–4. The way in which each observation should contribute to the CS was based on input from a Danish utility. The CS of a given pipe was then set to the worst of the observations. An overview of how the different defect types and severities contribute to the CS can be seen in Table C.2.

Table C.1: Overview of predictor variables and the corresponding data types as well as distribution or units. For continuous and numeric data types, the mean and std are presented. For categorical data types, the percentage of pipes in each category is presented and for binary data types, the percentage of true values is presented.

Predictor variable (abbreviation)	Data type	Distribution and units
Length	Continuous	43.59 ± 26.54 m
Age	Numeric	25.7 ± 21.1 Years
Material	Categorical	Concrete (61.22 %), plastic (33,77 %), clay (1,50 %), full reline (2.23 %), other (1.22 %)
Dimension	Continuous	306.6 ± 198.9 mm
Wastewater type (Wastewater)	Categorical	Sewage (38.18%), rain (31.66 %), combined (29.54 %)
Slope	Continuous	12.22 ± 11.61 mm/m
Year of construction (YoConst)	Numeric	Year 1982.2 ± 21.1
Year of rehabilitation (YoRehab)	Numeric	Year 1983.7 ± 21.6. This is set to YoConst if not rehabilitated
Type of rehabilitation (Rehab)	Categorical	Total replacement (5.11 %), Full reline, also included as material (2.23 %), Punctuate (0.04 %), unknown (0.00 %)
X coordinate (X)	Continuous	Adjusted UTM (m)
Y coordinate (Y)	Continuous	Adjusted UTM (m)
Utility ID	Numeric	
Ground level	Continues	32.07 ± 24.57 m
Depth	Continuous	2.43 ± 0.89 m
Groundwater level according to pipe (Groundwater)	Continuous	-4.48 ± 3.88 m
Soil type	Categorical	ML[1] (44.15 %), MS[2] (19.38 %), OPS[3] (9.79 %), FDS[4] (6.53 %), MaS[5] (4.28 %), MoS[6] (4.15 %), OMS[7] (3.45 %), FS[8] (2.77 %), MC[9] (1.00 %), MG[10] (0.46 %), Marsk (0.25 %), Lake (0.004 %)
Road type	Categorical	Tertiary (38.99 %), secondary (13.54 %), primary (13.00 %), traffic (3.99 %), other (1.88 %), no road (28.60 %)
Distance to road center (DistRoad)	Continuous	1.91 ± 1.93 m for pipes less than 10 m from road center. The remaining pipes have been assigned the value 99 m
Distance to nearest trees (Trees)	Categorical	<4 (6.51 %), <12 (25.50%), >12 (74.50 %)
Number of road grate (NoGrates)	Numeric	3.46 ± 3.56 grates
City type	Categorical	City zone (79 % incl. city center and industrial area) city center (18.43 %), industrial area (15.22 %)
Number of buildings (NoBuildings)	Numeric	6.09 ± 5.22 buildings
Area with tall buildings (BuildingHigh)	Binary	9.44 % True
Area with low buildings (BuildingLow)	Binary	77.06 % True

[1]Morain clay, [2]Meltwater sand, [3]Outwash plain sand, [4]Freshwater deposit of sand, [5]Marine sand, [6]Morain sand, [7]Old marine sand, [8] Fly sand, [9] Meltwater clay, [10]Marine gravel.

C.2.1 Preprocessing of Data

Thirty-five datasets were included in this study: one containing data from all the pipes and one for each of the utilities that had more than 100 bad pipes.

The preprocessing of the datasets was done by first removing features represented in less than 20% of the cases and then removing data points containing NaNs. An overview of the number of pipes available before and after data cleaning, the number of features removed from the dataset, and the number of pipes in bad condition can be seen in Table C.3.

All datasets were randomly split with 90% for training and 10% for testing. Due to the high imbalance between good and bad pipes, the training sets were randomly downsampled to contain an equal number of good and bad pipes.

Table C.2: Overview of how a CCTV-observation of a defect with severity zero to four entails the condition state to be in state one to four.

	Settled deposits	Attached deposits	Deformation	Obstacle	Displaced joint	Connection	Infiltration	Intruding sealing material	Surface damage	Transitional component	CNTP[1]	CNCU[2]	CNDR[3]	Manufacturing defect	CNCH[4]	Break and collapse	Roots	Water level
Severity 0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Severity 1	1	2	1	2	1	2	2	1	1	2	1	1	1	1	1	1	1	1
Severity 2	2	3	2	3	2	3	3	2	2	2	2	2	2	2	2	2	2	2
Severity 3	3	4	3	3	3	3	4	3	3	3	3	3	3	3	3	3	3	3
Severity 4	4	4	3	4	4	3	4	4	4	4	4	4	4	4	4	4	4	3

[1]Connection with lining defection or intruding connection, [2]Connection through cut hole in reline pipe, [3]Connection through drill hole in pipe, [4]Connection through chop hole in pipe. The color indicates the CS and goes from green to red.

C.2.2 Model Selection

As shown in Table C.1, the predictor variables available for this study have different data types, which is well handled by forest based models. Forest based models can be used to solve either regression or classification problems. They consist of several decision trees, which evaluate the data points according to a treelike structure. The construction of each decision tree is based on statistical variations in the datasets and an introduced randomness. Each decision tree votes for a specific outcome and based on these votes the forest makes a prediction. Two forest based model types were considered for classification: XGBoost [28] and Random Forest [29].

The Random Forest model was implemented using the Python library scikit-learn [30]. The number of decision trees was set to 177 and the max depth was set to 26 based on Hansen et al. [12]. The remaining hyperparameters were set to the default value. The number of estimators and max depth for the XGBoost model was first defined with inspiration from the settings of the random forest model. Hereafter different ways of setting these parameters were tested. For classification multiclass softmax was used. For the remaining parameters, the default values were used.

XGBoost benefitted from the ability to handle missing data; however, a XGBoost model takes much more time to train than a Random Forest model. XGBoost did not show better results than Random Forest. Furthermore, Random Forest is often used for deterioration modeling in sewer [12, 13, 19, 31] and in water pipes [32]. For this reason, Random Forest was used for this study.

Table C.3: Overview of the datasets containing more than 100 bad pipes before and after cleaning, as well as the number of features removed in the cleaning process and the number of pipes in condition state (CS) three and four after cleaning.

Dataset	Number of Pipes in Total	Number of Pipes after Cleaning	Number of Features Removed	No. Pipes in CS 3 after Cleaning	No. of Pipes in CS 4 after Cleaning
All pipes	318,457	196,174	0	64,969 (33 %)	20,542 (10 %)
Utility 1	20,379	17,730	0	5,992 (34 %)	1,758 (10 %)
Utility 2	10,062	8,270	0	3,138 (38 %)	872 (11 %)
Utility 3	9,904	8,469	1	3,666 (43 %)	1,103 (13 %)
Utility 4	10,116	7,913	0	2,166 (27 %)	615 (8 %)
Utility 5	18,745	15,830	0	4,477 (28 %)	1,162 (7 %)
Utility 6	17,109	13,867	0	5,290 (38 %)	1,141 (8 %)
Utility 7	4,669	3,108	6	687 (22 %)	280 (9 %)
Utility 8	4,522	3,315	0	892 (27 %)	790 (24 %)
Utility 9	12,163	9,451	1	2,759 (29 %)	1,708 (18 %)
Utility 10	734	640	1	135 (21 %)	103 (16 %)
Utility 11	6,355	5,686	1	1,899 (33 %)	469 (8 %)
Utility 12	20,945	16,453	0	5,128 (31 %)	987 (6 %)
Utility 13	1,779	1,268	0	312 (25 %)	120 (9 %)
Utility 14	18,154	14,587	0	6,214 (43 %)	2,063 (14 %)
Utility 15	2,757	2,027	0	580 (29 %)	231 (11 %)
Utility 16	18,025	15,812	4	4,700 (30 %)	1,560 (10 %)
Utility 17	3,855	3,252	0	1,145 (35 %)	311 (10 %)
Utility 18	9,655	7,128	0	2,516 (35 %)	880 (12%)
Utility 19	9,253	7,006	0	2,092 (30 %)	806 (12 %)
Utility 20	10,520	8,370	0	1,870 (22 %)	893 (11 %)
Utility 21	7,959	6,914	1	1,868 (27 %)	668 (10 %)
Utility 22	4,458	4,040	7	1,685 (42 %)	381 (9 %)
Utility 23	2,974	2,427	0	1,048 (43 %)	492 (20 %)
Utility 24	14,300	11,942	1	5,953 (50 %)	1,720 (14 %)
Utility 25	18,879	13,978	0	2,348 (17 %)	1,360 (10 %)
Utility 26	3,864	2,812	0	611 (22 %)	274 (10 %)
Utility 27	7,171	6,033	1	1,956 (32 %)	990 (16 %)
Utility 28	16,672	14,064	1	4,442 (32 %)	1,673 (12 %)
Utility 29	10,939	9,428	0	4,068 (43 %)	761 (8 %)
Utility 30	5,750	4,540	6	1,940 (43 %)	431 (9 %)
Utility 31	4,213	3,944	1	1,765 (45 %)	560 (14 %)
Utility 32	4,877	4,073	0	1,451 (36 %)	253 (6 %)
Utility 33	6,164	5,661	8	1,092 (19 %)	251 (4 %)

C.2.3 Feature Importance

In order to make the feature analysis, three methods were considered: (1) The Random Forest built-in feature importance measure [30], (2) Clustering the features in groups of features, training all combinations of the feature clusters, and investigating which feature clusters are most present in the best models and which feature clusters are most present in the bad models and (3) Making a backward step analysis by training a model on all but one feature for all features represented and removing the least contributing feature. This should then be repeated until only one feature is left.

Before selecting which method to use, it is worth considering the redundancy of the features. This has been handled previously by ensuring high heterogeneity between the features [15]. This approach entails removing a large number of features, which might be similar in most cases but could vary in essential cases. An example of this is year of construction and year of rehabilitation. If the pipe has not been rehabilitated, the year of rehabilitation is equal to year of construction, inducing a high redundancy between the two features. However, as rehabilitation is directly related to the condition of the pipe, the feature should be included in the analysis. Moreover, by including all the predictor variables in the analysis it is possible to account for the variations between utilities and obtain knowledge about features otherwise removed from the dataset.

As the built-in Random Forest method calculates the feature importance by number of splits for each feature, it is sensitive to redundancy between features. Clustering the features and training a model for all combinations of the feature clusters was tested initially, but it showed a high variance between the different utilities and did not contribute information on the individual features. The benefit of using the step analysis is that it encounters all the features; however, in the cases where many features are irrelevant it will be random if a feature is the 10th or the 20th least contributing feature. Like the built-in Random Forest method, this approach is sensitive to redundancy in the features, but the influence is of a more transparent character. Based on the above, the decision was made to conduct a backward step analysis.

Backward Step Analysis

To conduct the backward feature step analysis, the dataset was split randomly, and a model was trained for each of the features that was left out. This was repeated 10 times, and the predictor variable, which on average contributed the least to the performance, was removed. This was repeated until only one feature was left. Furthermore, the average performance of a model trained on all the features 10 times was found. For the feature analysis it was necessary to get a single performance measure. For this reason, the performance was calculated as the f1-score which is a balanced evaluation of the precision and recall. The f1-score was calculated using the Python library Scikit-lean [30] and the formula for calculating the f1-score can be seen in Equation (C.1).

$$f1_{score} = 2 \cdot (precicion \cdot recall)/(precision + recall) \tag{C.1}$$

A challenge using the f1-score is that it only encounters precision and recall. Thereby it does not encounter that the test set has a skewed distribution. For example, by randomly selecting 50% of the pipes a higher f1-score will be obtained than by selecting a number of pipes corresponding to the number of bad pipes in the dataset. Therefore, to evaluate how well the models performed according to a random selection strategy, the performance was calculated when randomly selecting 50% of the pipes and when randomly selecting a number of pipes corresponding to the number of bad pipes in the dataset. Due to variations in the distribution of bad pipes in the datasets, the F1-score cannot be used to give a fair evaluation of performance between utilities.

An overview of the method used for making the backward step analysis, calculating the performance when using all features were encountered, calculating the performance when randomly selecting 50% of the pipes, and calculating the performance when randomly selecting the same number of bad pipes as present in the dataset can be seen in Figure C.1.

C.2.4 Experiments

Three experiments were carried out. The purpose of the first experiment was to identify potential drawbacks of the approach used and take these into account in the remaining experiments. This experiment is referred to as the baseline. The purpose of the second experiment was to investigate how the performance and the feature importance changed when changing the definition of the target variable. The purpose of the last experiment was to investigate how the feature analysis changed between different utilities.

Baseline

This experiment was carried out on the full dataset. The condition of pipes in CS one and two was considered good while that of the pipes in CS three and four was considered bad. In this experiment, the backward step analysis was run for all features and relevant adjustments were incorporated.

Target Variable

In this experiment two backward step analyses were made: in the first analysis both pipes in CS three and four were considered bad pipes. In the second analysis only pipes in CS four were considered bad pipes. To ensure a fair comparison between the two analyses, the amount of training data in the first analysis was downsampled to the amount of training data available for the second analysis.

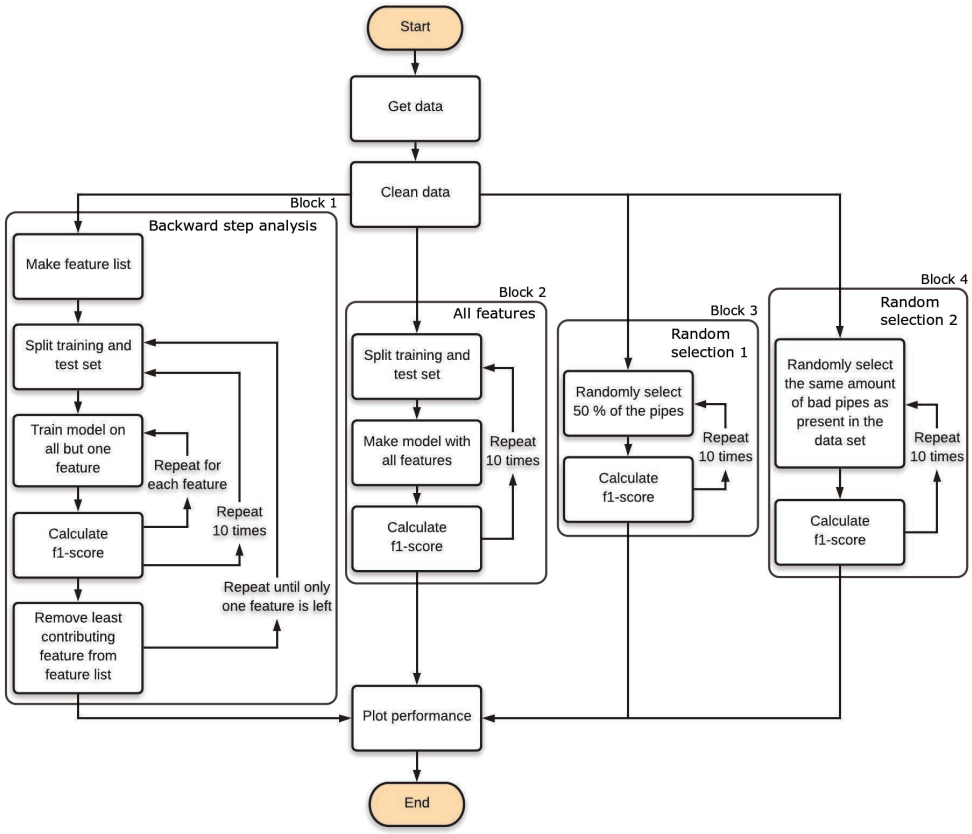


Fig. C.1: Feature step analysis when removing the features with the smallest contribution one by one. Block 1 shows the backward step analysis, block 2 shows the method for calculating the performance using all features, block 3 shows the method for calculating the performance when randomly selecting 50% of the pipes, and block 4 shows the method for calculating the performance when randomly selecting the same number of bad pipes as present in the dataset.

Difference between Utilities

A backward step analysis was performed for each of the utilities. This experiment was initially conducted solely considering pipes in CS 4 as being in bad condition; however, there was a relatively high variance in the features found relevant at the different utilities. This was particularly evident for utilities with few bad pipes entailing smaller datasets. For this reason, both pipes in CS three and four were considered bad pipes.

Each of the analyses was manually inspected to determine which parameters were

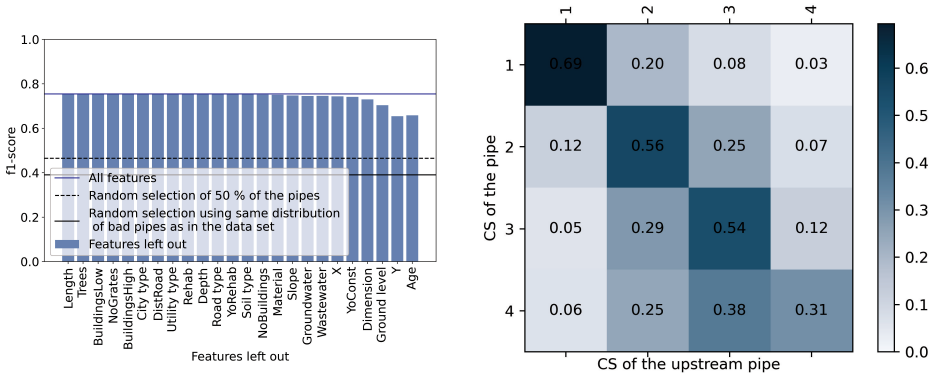
significant for each utility. This would preferably have been an automatic process, but as the results did not show a smoothly decreasing curve in all cases, an automatic approach would have required several assumptions.

An overview of the significant features for the different utilities was made, and the performance of the models was compared to the size of the dataset and the number of significant features.

C.3 Results

C.3.1 Baseline

The results of the baseline step analysis can be seen in Figure C.2a. The f1-score of the model using all the features is 0.75.



(a) (b)

Fig. C.2: (a) Backward feature step analysis for the baseline. Each chart shows the performance before the least contributing feature remaining has been removed, starting from the left to the right. The dashed and solid lines shows the performance if utilizing a random selection of respectively 50% of the pipes and a percentage corresponding to the distribution of bad pipes in the dataset. (b) Normalized confusion matrix for a pipe and the upstream connected pipe.

From Figure C.2a, it can be seen that both Y and X coordinates contribute to the predictions. This could indicate that when these parameters are included, the model learns the position of the pipes rather than that the actual parameters influence the condition state. In other words, this would correspond to using a nearest neighbor approach, which is problematic if applying the method to areas where training data is not available.

To clarify this suspicion, the probability of an upstream pipe present in a certain CS and given the pipe's condition was investigated. The normalized confusion matrix

for this can be seen in Figure C.2b, and it shows a clear correlation between the CSs of adjacent frames. Calculating the f1-score for pipes in CS three and four gives a f1-score of 0.69. As can be seen in the figure, the confusion matrix is not symmetric, which might be due to systematically occurring changes in the sewers. For example, it is common for an upstream pipe to be smaller than the downstream pipe but rarely the other way around. It should be noted that the figure is made from the same dataset as used for the feature analysis; however, not all pipes in the dataset have an inspected upstream pipe while others have more than one inspected upstream pipe.

Sewer inspections are not usually performed by taking representative samples from the whole sewer system but rather in subjectively selected areas. Therefore, the performance might be lower when applied to a part of the network that has not previously been inspected. To clarify this, another backward step analysis was applied but, instead of using a random split between training and test data, all pipes from four randomly selected utilities were used for testing and the remaining pipes for training. In so doing, the performance of the model based on all the predictor variable dropped by 10%. Furthermore, when performing the feature analysis, the utility ID and the X and Y coordinates were among the four worst predictor variables. For this reason, features related to location were not included in the remaining experiments. The new baseline can be seen in Figure C.3.

C.3.2 Target Variable

Figure C.4a shows the feature step analysis when considering pipes in both CS three and four to be in bad condition when using the same amount of training data as when considering only pipes in class four as being in bad condition. This model obtains a f1-score of 0.73 when using all features. Figure C.4b shows the feature step analysis when only considering pipes in CS 4 to be in bad condition. This model obtains a f1-score of 0.35.

Figure C.4a shows that the year of construction alone performs better than when combined with the relative groundwater level and ground level. This indicates that the relative groundwater level and ground level contributed positively to a group of features but introduced noise when included individually.

A smaller number of features are found to contribute when solely considering pipes in CS 4 as being in bad condition than when pipes in CS 3 also are considered as being in bad condition. All the parameters that contribute to the model performance in the first case mentioned, aside from wastewater type, also contribute in the second case.

C.3.3 Difference between Utilities

For each of the utilities in Table C.1 a backward step analysis was performed and manually inspected. Some of the utilities were observed to perform better when removing features up to a certain point. This was clearest for utility 10, which is the utility with

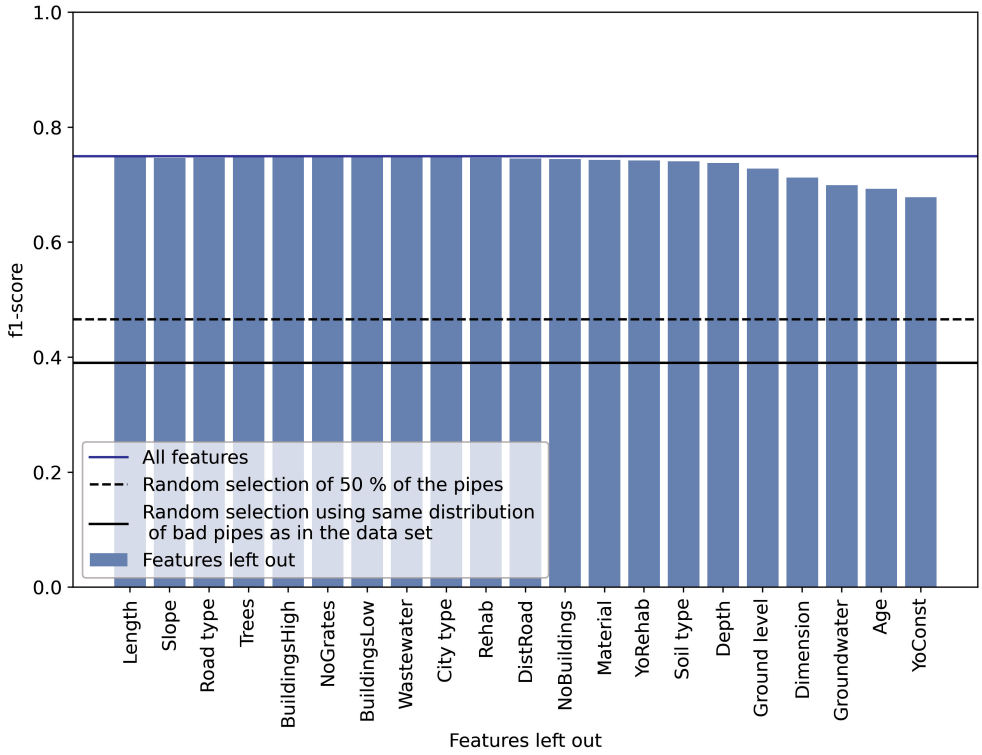
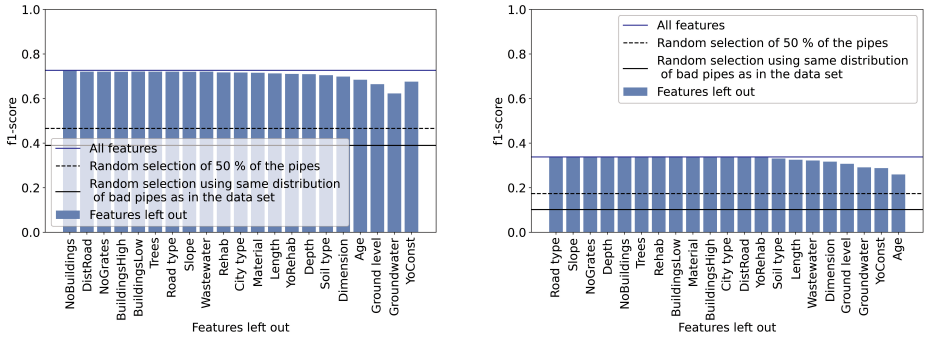


Fig. C.3: Backward feature step analysis for the new baseline where the geographical information has been removed. Each chart shows the performance when the least contributing feature remaining has been removed starting from the left and progressing to the right. The dashed black line shows the performance when randomly classifying 50% of the pipes as bad pipes and the solid black line shows the performance when randomly classifying a percentage of pipes, corresponding to the number of bad pipes in the dataset, as bad.

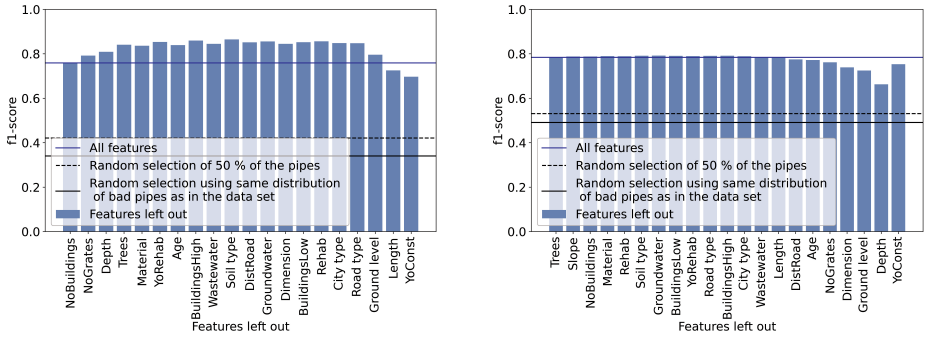
the smallest number of bad pipes, but the phenomenon could also be observed in some of the other utilities. The feature analysis for utility 10 can be seen in Figure C.5a. In most cases, the feature analysis shows a decrease in performance when features are removed. However, for some utilities the performance increases when the second-to-last feature is removed. This most often occurs if the second-to-last feature remaining is ground level or depth, but it has also been observed for groundwater to a smaller extent. An example of this can be seen in Figure C.5b, which shows the feature analysis for the utility with the largest number of bad pipes.

An overview of the predictor variables considered relevant for the different utilities



(a) (b)

Fig. C.4: (a) Same as baseline but with the same amount of training data as available when solely considering pipes in CS 4 as being in bad condition. (b) Feature step analysis when solely considering pipes in CS 4 as bad.



(a) (b)

Fig. C.5: Feature analysis for (a) the utility with the lowest number of bad pipes (utility 10), (b) the utility with the highest number of bad pipes (utility 14).

can be seen in Table C.4. In the table the performance is given as the f1-score when using the optimal number of features. The table also shows how many times a feature is found to be the most important feature.

On average 6.5 features were found to contribute to the performance. The table shows that year of rehabilitation and year of construction contain redundant information, and at least one of them is found to be significant in 24 of the utilities. For this reason, year of construction is considered more relevant than shown in the table if year of rehabilitation is not available and vice versa. Likewise, there might be some redundancy

Table C.4: Overview of which predictor variables contribute to the model performance for each utility. A "•" shows that the predictor variable contributes to the performance, a "o" shows that the predictor variable was included in the analysis but did not contribute to the model performance. The features are sorted in descending order, according to how often they contribute to the performance, and the utilities are sorted in descending order, according to number of pipes in CS 4. In addition, this table includes an overview of the most important feature and the best performance obtained for each utility.

Utility	Ground Level	Age	Groundwater	Wastewater	Length	Dimension	YoConst	YoRehab	Soil Type	Slope	Depth	NoBuildings	NoGrates	Material	DistRoad	Trees	Road Type	Rehab	City Type	Buildings Low	Buildings High	No. Features Relevant	Performance
Utility 14	•	•	o	o	o	•	•	o	o	o	•	o	•	o	•	o	o	o	o	o	o	7	0.78
Utility 1	•	•	•	o	•	•	o	o	•	o	o	o	•	•	o	o	o	o	o	o	o	8	0.71
Utility 24	•	•	•	o	o	•	o	o	•		•	o	•	•	o	o	o	o	o	o	o	8	0.76
Utility 6	•	•	•	•	o	o	•	o	o	o	•	o	o	o	o	•	o	o	o	o	o	7	0.71
Utility 16		•	•	•	•	o	•	•	o	o		•	o	o	•	o	o	•	o	o	o	9	0.71
Utility 12	•	o	•	•	•	•	o	•	•	o	o	•	o	o	o	o	o	o	•	o	o	9	0.69
Utility 28	•	•	•	o	•	•	•	o	o		o	o	o	o	o	o	o	o	o	o	o	6	0.77
Utility 5	•	•	•	•	o	•	•	o	•	•	o	•	o	o	o	o	•	o	o	o	o	10	0.78
Utility 29	•	o	•	•	•	•	•	•	•	•	o	o	•	•	o	•	o	o	o	o	o	12	0.76
Utility 3	•	o	•	o	o	•	•	o	o		o	o	o	o	o	o	o	o	o	o	o	4	0.77
Utility 9	•	•	•	•	•	o	•	o	o	o		o	o	o	o	o	o	o	o	o	o	6	0.77
Utility 2	o	•	•	•	o	o	o	•	o	o	o	o	o	o	o	o	o	o	o	o	o	4	0.82
Utility 25	•	•	•	o	o	•	o	•	o	•	o	o	o	o	o	o	o	o	o	o	o	6	0.77
Utility 18	•	•	o	•	•	•	o	o	•	o	•	o	o	•	o	o	•	o	o	o	o	9	0.79
Utility 27	o	o	•	•	•	o	o	•	o		o	o	o	o	o	o	•	o	o	o	o	5	0.66
Utility 19	o	o	•	o	•	o	•	•	o	o	o	•	•	o	o	•	o	o	o	o	o	7	0.71
Utility 4	•	•	o	•	o	•	o	o	•	•	o	o	•	o	o	•	o	o	o	o	o	8	0.65
Utility 20	•	•	o	•	•	•	•	•	o	•	o	o	o	o	o	o	o	o	o	o	o	8	0.76
Utility 21	•	•	o	•	•	•	•	o	o		o	o	o	o	o	o	o	o	o	o	o	6	0.73
Utility 30	•	•		o	•	•	o	o	o	o	o			o		•	o	o	o			5	0.70
Utility 11	o	•	•	•	•	o	•	o	o		•	o	o	o	o	o	•	o	o	o	o	7	0.69
Utility 31	•	•	•	o	•	o	o	o	•		o	o	o	•	•	o	o	o	o	o	o	7	0.77
Utility 22	•	•	•	o	•	•	o	o	o		•			o		o	o	o				6	0.84
Utility 32	o	o	•	•	o	o	o	•	o	o	o	o	o	o	o	o	o	o	o	o	o	3	0.71
Utility 8	•	o	o	•	o	•	o	•	•	o	o	•	o	o	•	o	o	o	o	o	o	7	0.74
Utility 23	•	•	•	•	•	o	o	o	•	•	•	o	o	o	o	o	o	•	o	o	o	9	0.80
Utility 17	o	o	o	o	o	o	•	o	o	o	o	o	o	o	o	o	o	o	o	o	o	1	0.81
Utility 33	•	•		•	o	o	o	•	o		•			o		o	o	o				5	0.56
Utility 7	o	•	•	•	•	o	o	•	•	•	o			•		o	o	o				8	0.71
Utility 26	o	•	o	o	o	o	o	o	•	o	o	o	o	o	o	o	o	•	o	o	o	3	0.75
Utility 15	o	o	o	•	o	•	•	o	o	o	o	o	o	o	•	o	o	o	•	o	o	5	0.80
Utility 13	o	o	o	•	o	o	•	•	o	o	•	•	o	o	o	o	o	o	o	o	o	5	0.86
Utility 10	•	o	o	o	•	o	•	o	o		o	o	o	o	o	o	•	o	o	o	o	4	0.85
Total	22	22	20	20	18	17	16	13	12	7	9	6	6	6	5	5	5	3	2	0	0		
Times best	0	8	6	0	0	1	13	5	0	0	0	0	0	0	0	0	0	0	0	0	0		

in number of buildings, buildings low, buildings high, and number of grids.

Table C.4 shows that year of construction is the most important feature in 13 of the utilities followed by age (8), groundwater (6), year of rehabilitation (5), and dimension (1). In general, there is a tendency for the continuous variables to be found relevant

more often than categorical and binary variables.

To identify general trends between the performance, the size of dataset and the number of relevant features, the relation between the number of bad pipes, the performance, and the number of features contributing to the performance is shown in Figure C.6.

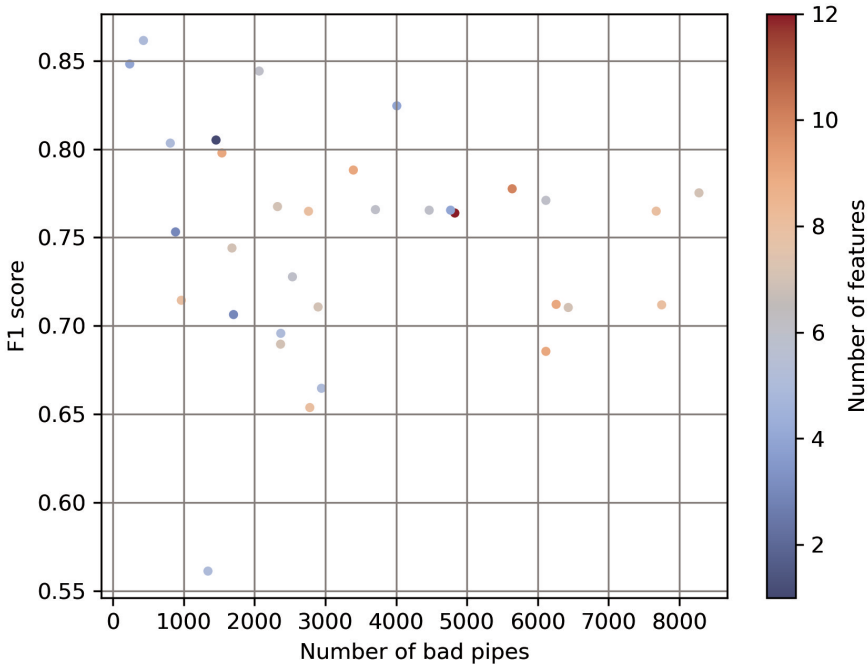


Fig. C.6: Plot of the relation between the number of bad pipes (first axis), the performance (second axis) and the number of relevant features.

In Figure C.6, the number of bad pipes is shown along the first axis, the f1 score is shown along the second axis and the number of relevant features is shown in a color scale ranging from blue to red. For utilities with more than 6000 bad pipes the number of features contributing to the performance staggered at six to nine.

As stated in Section 2.4.3, only pipes in CS four were initially considered as being in bad condition. In that analysis the performance and number of relevant features staggered for datasets with more than 1000–1500 bad pipes, which indicates that it is not solely the number of bad pipes that influences the results but also the total number of pipes inspected.

The smaller variation in performance and number of relevant features for utilities with a higher number of inspected pipes in bad condition, could indicate that these datasets contain a more representative segment of the pipes. Thereby they are less sensitive to a high or low occurrence of defects in an inspected area.

C.4 Discussion

Sewer deterioration modeling is complicated by several influencing factors. In this section the most prominent factors influencing the results are discussed, and the results are compared to previous findings in the literature.

C.4.1 Representativeness of Data

The results from the baseline experiment underlined the challenges of using historical data for sewer deterioration modeling, as the CCTV-inspections generally have been performed with a specific purpose, introducing a selective survival bias in the data [1]. However, as the datasets are comprehensive, most utilities do not have the finances to create a new dataset. Instead, the model developers must account for this by excluding the features in which the bias is most prominent, such as features related to geographical position. In the long term, utilities should include some spatial randomness in their strategy for CCTV-inspection.

C.4.2 Definition of Target Variable

Lack of publicly available data [5], numerous different standards for CCTV-inspections and different methods for evaluation of sewer deterioration models complicate the comparison of deterioration models. This also applies to the performance obtained in experiment two, where the f1-score drops from 0.73 to 0.35 when solely considering pipes in CS four as being in bad condition instead of considering pipes in both CS three and four. However, although the performance was affected, there was a high correlation in the predictor variables relevant for prediction of pipes in CS four and pipes in either CS three or four, which indicates that it is fair to make a binary evaluation of the feature importance.

Today CCTV inspections are performed by an operator who manually annotate the observations found in the sewers according to a given standard for tv inspections. These observations are often transformed into a general measure of the sewers condition. This condition measure can either be based on general standards or they can be utility specific. The benefit of utilizing the general standards are increased comparability between utilities whereas the benefit of utilizing a utility specific performance measure is that it can be adjusted to prioritize the types of defects relevant for the utility. For instance, a utility with limited capacity at the wastewater treatment plant might

increase weight on infiltration. Weighting some defects higher can cause the features related to these defects to become more important in a feature analysis. In the CS used in this study a higher weight has been put on attached deposit and infiltration according to other observation types as shown in Table C.2. This is consistent with the results showing a high importance of the relative groundwater level. As the groundwater maps available for this study were based on measurements every 500 m, the actual groundwater level can change significantly between the data points. It is likely that the ground level can compensate for these changes, which will induce a higher weight on this feature in the feature analysis.

C.4.3 Size of Datasets

When considering pipes in both CS three and CS four as being in bad condition, the performance and the number of features relevant staggered for datasets with more than 6000 pipes in bad condition. In the initial analysis only pipes in CS four were considered bad. In that analysis the performance and number of relevant features staggered for datasets with more than 1000–1500 pipes in bad condition. This indicates that the number of bad pipes required for optimal performance is correlated with how the target variable is defined and the total number of pipes inspected.

Furthermore, it is worth noticing that if solely considering the utilities with more than 1000 bad pipes, there is more consensus on which features contribute to the performance. For ground level the percentage of time it is found to be relevant increases from 69% to 78%. Similar tendencies are present for age (67% to 71%) and relative groundwater level (65% to 73%). A full overview is presented in Table C.5.

C.4.4 Irregularities in the Step Analysis

For some utilities, the performance improved when predictor variables were removed, indicating overfitting of the model. This was clearest for utility 10, which is also the utility with the smallest amount of training data. For datasets with more than 10,000 pipes, the tendency could still be observed in some cases after cleaning but removing features did not lead to an increase in performance of more than two to three percent.

In a few cases, the performance suddenly increased when removing one parameter. This could not be explained by stochasticity in the performance or overfitting. An example of this can be seen in Figure C.5b. This is most likely because some predictor variables perform well when combined but introduce noise when considered individually.

C.4.5 Comparison to the Literature

Mohammadi et al. [24] reviewed 24 papers, of which 19 had investigated which features were significant. In Table C.6 the results of this study are compared to the findings by Mohammadi et al.

Table C.5: Overview of how often a feature contributes to the performance when considering all the utilities and when considering only utilities with more than 1000 bad pipes.

Predictor Variables	All Utilities		Utilities with More Than 1000 Bad Pipes	
	Times Present	Percent of Times Found Relevant	Times Present	Percent of Times Found Relevant
Ground level	32	69	27	78
Age	33	67	28	71
Groundwater	31	65	26	73
Wastewater	33	61	28	61
Length	33	55	28	57
Dimension	33	52	28	57
Year of construction	33	48	28	46
Year of rehabilitation	33	39	28	39
Soil type	33	36	28	36
Slope	23	30	19	32
Depth	31	29	26	31
No. buildings	29	21	25	20
No. grates	29	21	25	24
Material	33	18	28	21
Dist. to road center	29	17	25	16
Dist. to trees	33	15	28	18
Road types	33	15	28	14
Rehabilitation type	33	9	28	7
City type	30	7	26	4
Building low	29	0	25	0
Buildings high	29	0	25	0

In the review by Mohammadi, there is a higher consensus about which predictor variables are significant. The most probable reason for this is that Mohammadi et al. reviewed studies whose authors selected a number of predictor variables. For example, four of the papers investigated between two and eight predictor variables and did not find any insignificant variables. In general, there is a consensus that length, age, dimension, ground water, and wastewater type are often important predictor variables. However, the model developer should consider the specific case when selecting predictor variables as there is no "gold standard".

C.4.6 CCTV-Inspection Planning

The still increasing access to pipe specific data and the increasing awareness of the benefits related to risk based pipe inspection and rehabilitation on pipe level are essentials when optimizing the management of sewer systems to save costs and resources. Sewer deterioration modeling is an essential element in this; however, the scientific literature

Table C.6: Comparison of how often the different features contribute to the performance in this study and in the review by Mohammadi et al.

Predictor Variables	Results		Mohammadi et al.	
	Times Present	Percent of Times Found Relevant	Times Present	Percent of Times Found Relevant
Ground level	27	78	-	-
Age	28	71	18	78
Groundwater	26	73	3	100
Wastewater	28	61	6	83
Length	28	57	11	91
Dimension	28	57	17	71
Year of construction	28	46	-	-
Year of rehabilitation	28	39	-	-
Soil type	28	36	5	20
Slope	19	32	12	42
Depth	26	31	16	44
No. buildings	25	20	-	-
No. grates	25	24	-	-
Material	28	21	15	67
Dist. to road center	25	16	-	-
Dist. to trees	28	18	-	-
Road types	28	14	5	40
Rehabilitation type	28	7	-	-
City type	26	4	-	-
Building low	25	0	-	-
Buildings high	25	0	-	-
Location	-	-	5	40
Up-invert	-	-	1	0
Down-invert	-	-	1	0
Bedding type	-	-	2	100
Corrosivity	-	-	2	50
Number of trees	-	-	5	60
Traffic	-	-	1	1
Flow	-	-	3	67
Hydrohalic	-	-	2	100
Location	-	-	5	40
Up-invert	-	-	1	0

dealing with the underlaying parameters influencing the deterioration models is sparse. The findings of this study enlighten some of these shortcomings, and the findings can be incorporated in future model development.

Generally, deterioration models can be used to give a snapshot of the sewer system and is used when no CCTV-inspection has been made or when the CCTV inspection is outdated. Typically, the deterioration models are based on datasets which have been

collected over several years. Therefore, users of deterioration models should be aware that the predictions of the CSs are evaluated on historical data and thereby cannot give a fair prediction of future condition states. For example, plastic pipes were rarely used 50 years ago, and plastic pipes older than 50 years have limited representation in the data. Furthermore, the surrounding environment, material quality etc. change over time. Future predictions of CSs are further complicated by variations in the degradation profile of different defect types. Some defect types occur stochastically and do not degrade over time such as defects related to pipe connections or installation of the pipes. Other defects degrade over time such as surface damage. Surface damage is often seen in concrete pipes due to the presence of hydrogen sulphide which erode the surface over time. Hydrogen sulphide is typically formed in pump pipes. Likewise, the degradation profile for defects related to roots in the pipes depends on the surrounding trees and their growth.

C.5 Conclusion

The primary contribution of this paper is a comprehensive analysis of the feature importance in sewer deterioration modeling. The paper addresses factors that influence sewer deterioration modeling and acknowledges weak or missing information in the literature, such as handling of biased datasets, the impact of how bad pipes are defined, and the variations in feature importance between utilities.

Deterioration models are usually based on CCTV-inspections performed over several years with a specific purpose in mind. This is problematic due to a selective survival bias in the data whereby the models do not perform as well on noninspected areas as they do on inspected areas. Ideally the datasets should be random in character, but due to economic constraints this is often infeasible. Instead, model developers should avoid utilization of geographically related parameters. Moreover, utilities should include randomness in their strategy for CCTV inspection.

Changing the definition of when a pipe is in bad condition produced large deviations in model performance. However, in the feature analysis it was the same features that contributed to the performance, although more features contributed when both pipes in CS three and four were considered bad than when only pipes in CS four were considered bad. This indicates that it is fair to use an advantageous split between good and bad pipes when making a feature analysis.

Comparison of feature analysis from 33 different utilities showed a relatively high variance in the number of features contributing to the performance, which features contributed, and the performance obtained by the models. These variations were especially high for utilities with fewer than 6000 pipes in bad condition. It is worth noting that the number of bad pipes depends on the definition of bad pipes. When solely considering pipes in CS four as “bad”, the high variations were primarily present for utilities with fewer than 1000–1500 pipes in bad condition.

No feature was considered relevant in more than 69% of the utility specific models; however, when only considering utilities with more than 1000 bad pipes there was a higher consensus on which features were relevant (up to 78%). For these utilities, the features that contributed to the performance most of the time were ground level (78%), age (71%), groundwater level (73%), wastewater type (61%), length (57%), dimension (57%), year of construction (46%), and year of rehabilitation (39%). As there is a high redundancy between year of construction, year of rehabilitation, and age, removing one of these as a possible predictor variable would most likely induce the others to contribute to the performance in more cases. In 26 out of 33 cases the most important feature was related to either age, year of construction, or year of rehabilitation. On average 6.5 features contributed to the utility specific models.

The overall trends in feature importance found in this work showed consensus with the findings in a review by Mohammadi et al. [24]; however, due to variations in study design of the articles reviewed by Mohammadi et al. the two papers are not comparable on a detailed level.

The added value of this paper is a better understanding of the underlying parameters influencing sewer deterioration modeling and knowledge of feature importance when encountering the statistical variations between utilities. The exact results related to feature importance are specific to the condition measure used in the study, however, the overall trends are comparable to findings in the literature and can be used to assist the feature selection for sewer deterioration modeling, which is important because feature extraction is a labor intensive process.

Author Contributions: Conceptualization, B.D.H., M.U., T.B.M. and D.G.J.; methodology, B.D.H., T.B.M., D.G.J.; software, B.D.H.; validation, B.D.H.; formal analysis, B.D.H. and S.H.R.; investigation, B.D.H., S.H.R., M.U., T.B.M. and D.G.J.; resources, B.D.H.; data curation, B.D.H. and S.H.R.; writing— original draft preparation, B.D.H.; writing—review and editing, B.D.H., S.H.R., M.U., T.B.M. and D.G.J.; visualization, B.D.H.; supervision, S.H.R., M.U., T.B.M. and D.G.J.; project administration, B.D.H. and D.G.J.; funding acquisition, B.D.H., M.U., T.B.M. and D.G.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovation Fund Denmark.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] F. Tscheikner-Gratl, N. Caradot, F. Cherqui, J. P. Leitão, M. Ahmadi, J. G. Langeveld, Y. Le Gat, L. Scholten, B. Roghani, J. P. Rodríguez, M. Lepot, B. Stegeman, A. Heinrich-

sen, I. Kropp, K. Kerres, M. d. C. Almeida, P. M. Bach, M. Moy de Vitry, A. Sá Marques, N. E. Simões, P. Rouault, N. Hernandez, A. Torres, C. Werey, B. Rulleau, and F. Clemens, "Sewer asset management – state of the art and research needs," vol. 16, no. 9, pp. 662–675.

[2] D. Marlow, L. Pearson, D. H. MacDonald, S. Whitten, and S. Burn, "A framework for considering externalities in urban water asset management," vol. 64, no. 11, pp. 2199–2206.

[3] B. Roghani, F. Cherqui, M. Ahmadi, P. Le Gauffre, and M. Tabesh, "Dealing with uncertainty in sewer condition assessment: Impact on inspection programs," vol. 103, pp. 117–126.

[4] J. Dirksen, F. H. Clemens, H. Korving, F. Cherqui, P. Le Gauffre, T. Ertl, H. Plihal, K. Müller, and C. T. Snaterse, "The consistency of visual sewer inspection data," vol. 9, no. 3, pp. 214–228.

[5] J. B. Haurum and T. B. Moeslund, "A survey on image-based automation of CCTV and SSET sewer inspections," vol. 111, p. 103061.

[6] A. Altarabsheh, M. Ventresca, and A. Kandil, "New approach for critical pipe prioritization in wastewater asset management planning," vol. 32, no. 5, p. 04018044.

[7] M. Elmasry, T. Zayed, and A. Hawari, "Multi-objective optimization model for inspection scheduling of sewer pipelines," vol. 145, no. 2, p. 04018129.

[8] S. M. Ghavami, Z. Borzooei, and J. Maleki, "An effective approach for assessing risk of failure in urban sewer pipelines using a combination of GIS and AHP-DEA," vol. 133, pp. 275–285.

[9] M. Elmasry, A. Hawari, and T. Zayed, "An economic loss model for failure of sewer pipelines," vol. 14, no. 10, pp. 1312–1323.

[10] G. Carvalho, C. Amado, R. S. Brito, S. T. Coelho, and J. P. Leitão, "Analysing the importance of variables for sewer failure prediction," vol. 15, no. 4, pp. 338–345.

[11] B. D. Hansen, S. H. Rasmussen, T. B. Moeslund, M. Uggerby, and D. G. Jensen, "Sewer deterioration modeling: The effect of training a random forest model on logically selected data-groups," vol. 176, pp. 291–299.

[12] B. D. Hansen, D. Getreuer Jensen, S. H. Rasmussen, J. Tamouk, M. Uggerby, and T. B. Moeslund, "General sewer deterioration model using random forest," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 834–841, IEEE.

[13] T. Laakso, T. Kokkonen, I. Mellin, and R. Vahala, "Sewer condition prediction and analysis of explanatory factors," vol. 10, no. 9, p. 1239.

[14] H. Park, S. H. Ting, and H. D. Jeong, "Procedural framework for modeling the likelihood of failure of underground pipeline assets," vol. 7, no. 2, p. 04015023.

[15] Y. Le Gat, "Modelling the deterioration process of drainage pipelines," vol. 5, no. 2, pp. 97–106.

[16] M. Elmasry, A. Hawari, and T. Zayed, "Defect based deterioration model for sewer pipelines using bayesian belief networks," vol. 44, no. 9, pp. 675–690.

[17] A. Hawari, F. Alkadour, M. Elmasry, and T. Zayed, "Condition assessment model for sewer pipelines using fuzzy-based evidential reasoning," vol. 16, no. 1, pp. 23–37.

[18] G. Kabir, N. B. C. Balek, and S. Tesfamariam, "Sewer structural condition prediction integrating bayesian model averaging with logistic regression," vol. 32, no. 3, p. 04018019.

[19] M. M. Rokstad and R. M. Ugarelli, "Evaluating the role of deterioration models for condition assessment of sewers," vol. 17, no. 5, pp. 789–804.

[20] D. Fuchs-Hanusch, M. Günther, M. Möderl, and D. Muschalla, "Cause and effect oriented sewer degradation evaluation to support scheduled inspection planning," vol. 72, no. 7, pp. 1176–1183.

[21] M. O'Reilly, Transport, and R. R. Laboratory, *Analysis of Defects in 180 Km of Pipe Sewers in Southern Water Authority.* Research report (Transport and Road Research Laboratory), Ground Engineering Division, Structures Group, Transport and Road Research Laboratory.

[22] X. Yin, Y. Chen, A. Bouferguene, and M. Al-Hussein, "Data-driven bi-level sewer pipe deterioration model: Design and analysis," vol. 116, p. 103181.

[23] J. P. Davies, B. Clarke, J. Whiter, R. Cunningham, and A. Leidi, "The structural condition of rigid sewer pipes : a statistical investigation," vol. 3, pp. 277–286.

[24] M. Malek Mohammadi, M. Najafi, S. Kermanshachi, V. Kaushal, and R. Serajiantehrani, "Factors influencing the condition of sewer pipes: State-of-the-art review," vol. 11, no. 4, p. 03120002.

[25] A. Hawari, F. Alkadour, M. Elmasry, and T. Zayed, "A state of the art review on condition assessment models developed for sewer pipelines," vol. 93, p. 103721.

[26] J. Guzmán-Fierro, S. Charry, I. González, F. Peña-Heredia, N. Hernández, A. Luna-Acosta, and A. Torres, "Bayesian network-based methodology for selecting a cost-effective sewer asset management model," p. wst2020299.

[27] B. Laden, DANVA, and Fotomanualgruppen, *Fotomanualen: TV-inspektion af afløbsledninger.* DANVA. OCLC: 769296284.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM.

[29] L. Breiman, "Random forests," vol. 45, no. 1, pp. 5–32.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," vol. 12, pp. 2825–2830.

[31] R. R. Harvey and E. A. McBean, "Predicting the structural condition of individual sanitary sewer pipes with random forests," vol. 41, no. 4, pp. 294–303.

[32] "Regression methods for predicting rate and type of failures of water conduits,"

Paper D

Prediction of the Methane Production in Biogas Plants Using a Combined Gompertz and Machine Learning Model

Bolette D. Hansen, Jamshid Tamouk, Christian A. Tidmarsh, Rasmus Johansen, Thomas B. Moeslund, and David G. Jensen

The paper has been published in the
Lecture Notes in Computer Science Vol. 12249, pp. 734–745, 2020.

© Springer Nature Switzerland AG 2020
The layout has been revised.

Abstract

Biogas production is a complicated process and mathematical modeling of the process is essential in order to plan the management of the plants. Gompertz models can predict the biogas production, but in co-digestion, where many feedstocks are used it can be hard to obtain a sufficient calibration, and often more research is required in order to find the exact calibration parameters. The scope of this article is to investigate if machine learning approaches can be used to optimize the predictions of Gompertz models. Increasing the precision of the models is important in order to get an optimal usage of the resources and thereby ensure a more sustainable energy production. Three models were tested: A Gompertz model (Mean Absolute Percentage Error (MAPE) = 9.61%), a machine learning model (MAPE = 4.84%), and a hybrid model (MAPE = 4.52%). The results showed that the hybrid model could decrease the error in the predictions with 53% when predicting the methane production one day ahead. When encountering an offset in the predictions the reduction of the error was increased to 66%.

D.1 Introduction

Climate changes and increasing energy demands have increased the focus on renewable energy in the recent years. The biogas industry contributes to this by e.g., producing energy from wastewater and reducing pollution from agriculture. The biogas industry has had a significantly progress in Europe in the recent years, where the capacity was almost tripled in the period from 2007 to 2017 [1].

Biogas is produced by anaerobic digestion of organic materials such as urban waste for example food, sludge and garden waste, animal manure, industrial waste, lignocellulosic materials such as various types of straw and the biomass of microalgae [2]. How fast a feedstock can be transformed to methane depends on the composition of the feedstock. Carbohydrates, protein and fat are easy digestible, whereas e.g. lignocellulosic materials are much harder to digest [2].

In addition to this, several parameters influence the process. These are, among others, the pH, carbon/nutrient rate, organic loading rate, temperature, the microorganisms and enzymes present, presence of some types of fatty acids, and usage of substrates. Over or under representation of some of the parameters can even lead to failure of the system [3, 4].

Mono-digestion of some feedstocks can be problematic, as it might bring the system out of balance, for instance changing the carbon/nutrient rate will lead to a too high concentration of problematic fatty acids or change the pH. Therefore, much research has focused on co-digestion of feedstocks, which can lead to an increase in the biogas production of 25–400% [3]. However, co-digestion complicates the digestion process further and therefore, mathematical modelling of the process is essential in order to keep

the balance and avoid failure [3]. In 2002 [5] published the IWA Anaerobic Digestion Model No.1 (ADM1). The ADM1 can simulate an average trend in the different parameters, however, it cannot simulate the immediate variations [6]. Since the ADM1 was published plugins with modifications have been added and in 2015 [7] states that despite new models have been developed, there is readily justification for developing a new ADM2. In order to develop an ADM2 it is necessary to have a uniform approach to the mechanisms and challenges which will require further research. After 2015, other models and optimizations of the ADM1 have been performed [8].

Where the ADM1 model focuses on modelling the whole process, another model type, called Gompertz models, is often used for prediction of the gas production. This type of model can give very accurate predictions of the methane production [9], however, like the ADM1 model, this model requires a precise calibration. In order to calibrate a Gompertz model, Gompertz functions need to be set up for all the biomasses used in the model. Several studies have been made in order to improve Gompertz models by experimentally finding the kinematic parameters describing the methane production for a wide range of feedstocks [9–11]. For this reason, the literature can provide the kinematic parameters for a large variety of feedstocks. If the Gompertz functions for some of the feedstocks in a biogas plant is not available in the literature, they can be found experimentally. However, as this often takes several months, they are often based on expert knowledge instead.

It is worth noticing that despite Gompertz variables are available for one feedstock, there might be local variations in the composition of the feedstocks influencing the gas production. This is complicated further from co-digestion and other parameters influencing the digestion of each material. For this reason and despite making several tests in the laboratory, the results might not fit well in a real case scenario. Despite that the mathematical models are essential in order to ensure a stable process, and avoid failure, they are very hard to calibrate, especially when the parameter complexity is high. In these cases further research in parameter characterization is required [3].

Machine learning models have been developed in order to predict and optimize the methane production. In wastewater treatment plants Neural Networks have for instance been used to find the optimal settings for as high a methane yield as possible [12]. In agricultural biogas plants a combination of Genetic Algorithm and Ant Colony Optimization has been used to predict the present gas production based on measured process variables [13]. In controlled laboratory scale experiments Neural Networks have shown precise predictions of the biogas production [14]. Likewise, the machine learning algorithms Random Forest and XGBoost have shown efficient future predictions of the methane yield in an industrial-scale co-digestion facility [15].

Being able to predict the future biogas production is essential in order to plan the operation of the biogas plant. In some cases, the goal is to produce as much methane as possible, as there is an unmet demand. In other cases, it is essential to keep as stable a production as possible in order to meet the demand while avoiding overproduction.

In case of overproduction, the process can be artificially inhibited by chemicals or the surplus methane can be burned off. In both cases resources are not optimally used. This has led to development of software tools used for planning of the biogas production. The software has a framework to take in different machine learning models for prediction of the biogas production. In this case the machine learning models were used to predict the production in three categories: Low, medium, and high [16].

The scope of this article is to investigate if machine learning approaches can be used to optimize the predictions of Gompertz models in industrial settings. This is done by comparison of a Gompertz model, a machine learning model, and a combined model.

D.2 Method

D.2.1 Biogas Plant

The biogas plant has a capacity at approximately 220,000 t biomass/year and produces almost 10 mio. Normal cubic meter (Nm^3 methane/year, corresponding to 99,7 GWh heat per year. The main feedstocks used in the plant are seaweed, manure, eulat and pectin. However, in total 18 specific feedstocks were used. It is worth noticing that it is an industrial setting and the available feedstocks changes over time. For this reason, some of the feedstocks were only used in the first half of the measurement period, while others were only used in the second half of the period. Fortunately, the amount of these temporary feedstocks is limited, and the models needs to be tolerant to this type of changes.

In this biogas plant a Gompertz model is used to plan the infeed to the plant in order to obtain as constant a biogas production as possible.

D.2.2 Data Set

The data set obtained from the biogas plant contained consecutive measures from 818 days. In total 18 different feedstocks were used of which one was only used in the first half of the data set and four were only used in the last half of the data set.

D.2.3 Gompertz Model

For each biomass the expected methane production over time can be described by Gompertz functions [17, 18]. The formula for Gompertz functions can be seen in Eq. D.1.

$$P(t) = P_{max} \cdot exp\left(-exp\left(\frac{R_{max} \cdot e}{P_{max}} \cdot (\lambda - t) + 1\right)\right) \tag{D.1}$$

Where P is the culminative methane production, P_{max} is the maximal total methane production R_{max} is the maximal methane production rate, t is the time measured in

days and k is the delay before any gas is produced from the specific biomass. Hereafter the methane production for a biomass on a given day can be calculated as seen in Eq. D.2.

$$p_{day} = P(t_{day}) - P(t_{day} - 1) \tag{D.2}$$

Where P_{day} is the amount of produced methane on the specific day and t_{day} is the day for prediction. The expected daily methane production per ton of each biomass can be seen in Fig. D.1.

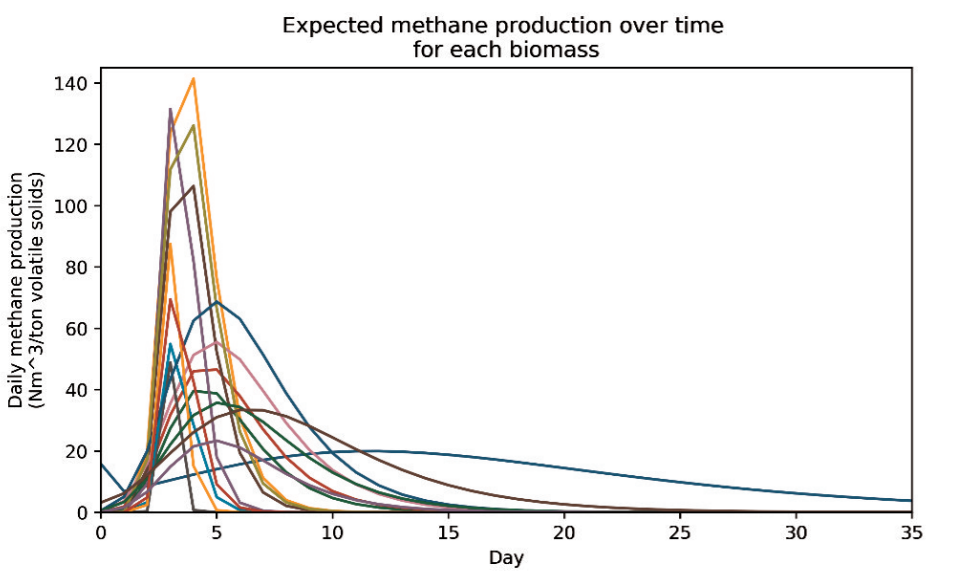


Fig. D.1: Expected methane produced by each type of biomass over time

The methane production was then calculated as the sum of the added methane per biomass per day, including biomasses added up to 60 previous to the precent day. The model encountered the retention as a percentage removal of biomass per day.

The parameters in the Gompertz functions were initially based on empirical numbers where these were available. For the biomasses where empirical data was not available, they were estimated based on knowledge about similar biomasses. Hereafter the model was calibrated using the previous 90 days as calibration data. The calibration was done by minimizing the root mean square error.

In addition to the initial parameters each biomass had a span they were forced to stay within.

D.2.4 Machine Learning Model

Preprocessing

Before training the machine learning model, all the data points were normalized by subtracting the mean value and scaling to unit value using the Python library scikitlearn [19]. Hereafter zeros were replaced with values close to zero. This was done in order to account for algorithms being sensitive to zeros and as it showed better results in some cases. As biogas production is a time-consuming process which takes several days, eight additional features were added to each data point. These features were the measured gas production from the previous six days and the mean and standard divination calculated for the present and previous nine days. As the feature creation in this case implies inclusion of the input parameters nine days before the measurement, the first nine days of the measurement data was excluded due to missing data.

As the last part of the data set was collected while the model was developed, only the first 430 datapoints were used for training, while the last 350 datapoints were used for testing. Thereby the training set contains one biomass which is not present in the test set and the test set contains four parameters which are not present in the training data. An overview of the preprocessing pipeline can be seen in Fig. D.2

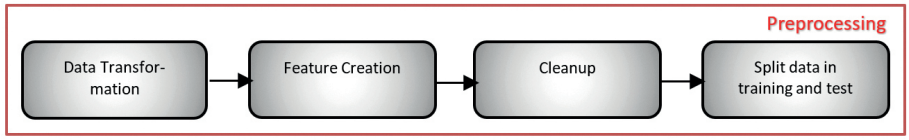


Fig. D.2: The preprocessing pipeline

Model Development and Training

The training set was split into 25 folds for folding where 23 folds were used for training, one fold was used for validation and one fold was used for test.

Initially 15 commonly used machine learning algorithms from the Python library scikit-learn [19] were tested. These were uniform k-nearest neighbors (kNN), distance kNN, Bagging [20] with Decision Tree, AdaBoost [21] Regressor with Decision Tree, Random Forest Regression [22], Bagging with Random Forest, recursive feature elimination (RFE) with the core of linear Ridge, Recursive Feature Elimination by using Gradient Boosting [23], Principal Components Regression, Quadratic Discriminant Analysis, Lasso Regression, Multilayer Perceptron (MLP) Regressor, Naive Bayes [24], Extra Trees [25], and Support Vector Machine [26]. Based on this initial test seven methods were selected. These algorithms were uniform kNN, distance kNN, recursive feature elimination (RFE) with the core of linear Ridge, MLP Regressor, Ada Boost Regressor with Decision Tree Regressor, RFE with the core of Gradient Boosting Regressor, and Random Forest Regressor.

For each fold, the best three models were ensembled whereby the prediction would

be the mean of the three models. If the ensemble score was better than the score of the single models this was selected. Otherwise the best single model was selected. If the selected model did not obtain a sufficient precision, no model from that fold was saved. Lastly all the saved models were applied to the test set and the prediction of the models were averaged in order to predict the methane production one day ahead. An overview of the setup can be seen in Fig. D.3.

D.2.5 Hybrid Model Based on the Gompertz Model and the Machine Learning Model

In order to make a hybrid model based, the error of the Gompertz model was found by subtracting the Gompertz predictions from the measurements. Hereafter a machine learning model similar to the model described in Sect. 2.4 was trained to predict the error of the Gompertz model. Subsequently the predictions from the machine learning model were added to the predictions from the Gompertz model in order to predict the amount of methane produced.

D.3 Results

The predictions from respectively the Gompertz model, the machine learning model, and the hybrid model for the full test set can be seen in Fig. D.4, and a zoomed version can be seen in Fig. D.5. Likewise, the correlation between the observations and the predictions was found as seen in Fig. D.6.

From Fig. 6 it was observed that there was an offset in the predictions according to the correlation line. A similar offset was observed for prediction on the training set. If calculating the mean of the measurements and the predictions and adjusting for this by adding the difference in mean values to the predictions the error of the predictions is decreased. The Mean Absolute Percentage Error (MAPE) for each of the models with and without adjustment according to the mean prediction can be seen in Table D.1.

Table D.1: The mean absolute error for each of the three models

Model	↓ MAPE	↓ MAPE after adjustment
Gompertz model	9.61 %	9.00 %
ML model	4.84 %	3.78 %
Hybrid model	**4.52 %**	**3.06 %**

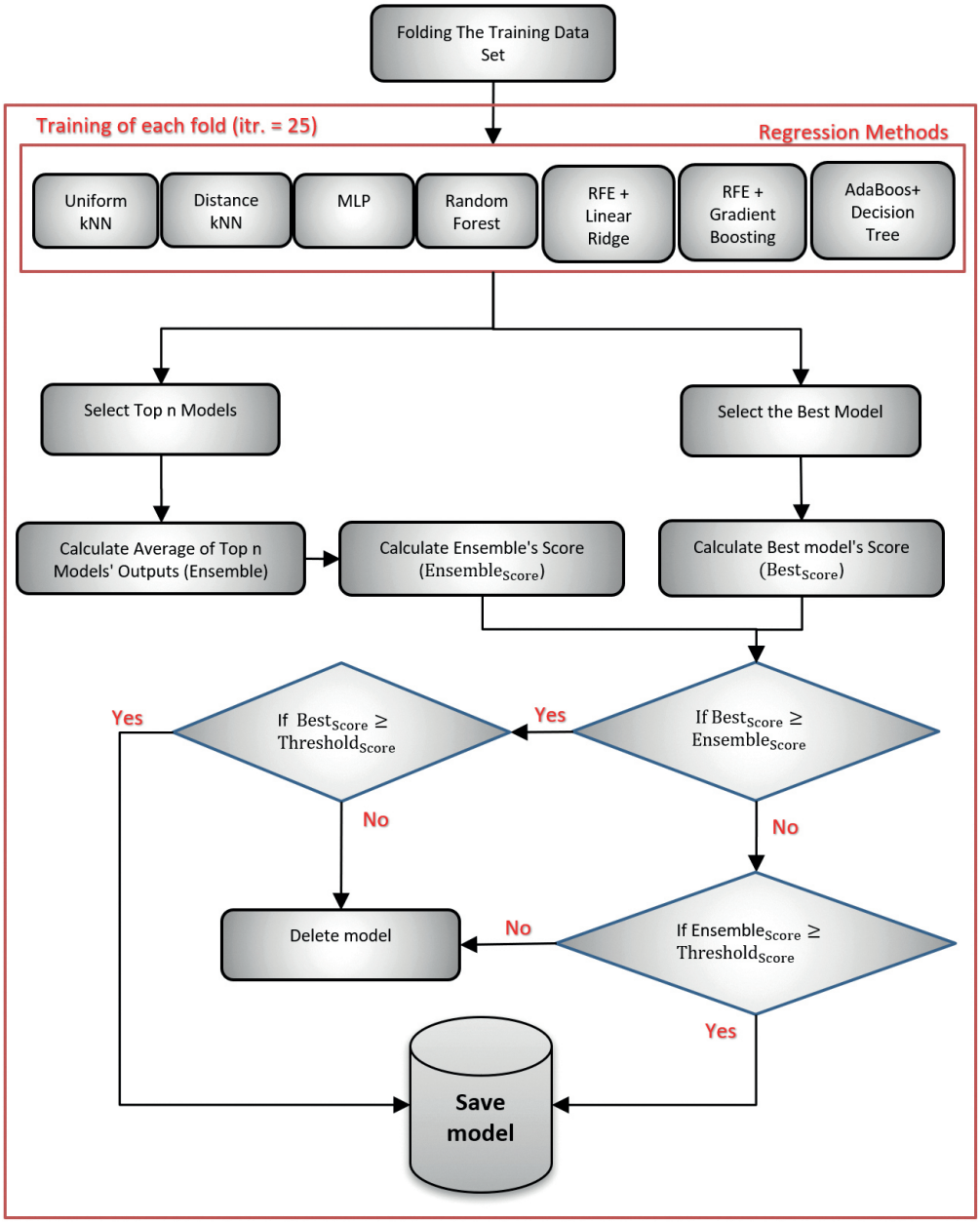


Fig. D.3: Flowchart over the method for developing the machine learning model

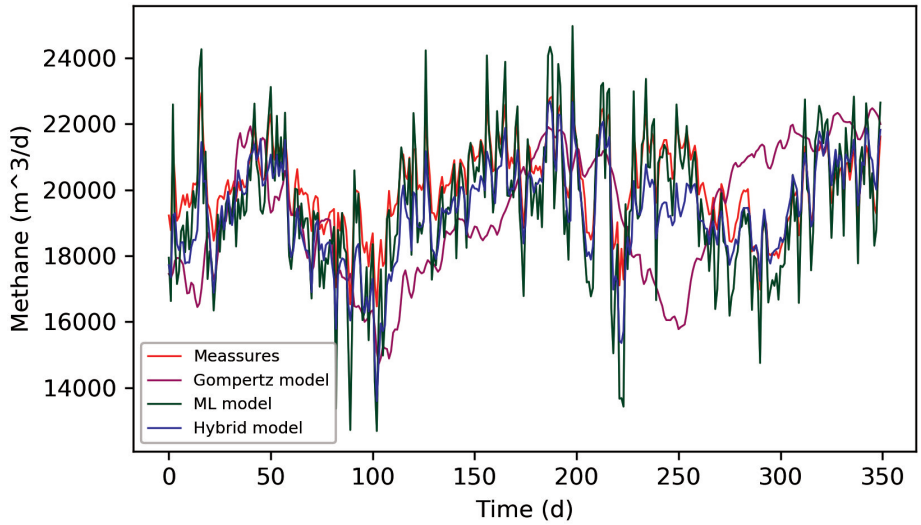


Fig. D.4: Comparison of the measured biogas production and the production predicted one day ahead by the Gompertz model, the machine learning model and the hybrid model.

D.4 Discussion

Modelling of biogas plants is essential in order to keep an optimal operation of the plants. Gompertz models can be used to plan the infeed to biogas plants in order to keep a constant output. However, these models can be hard to calibrate, especially when the number of feedstocks is high. The Gompertz model presented in this paper can estimate the methane production with a MAPE at 9.61%. From Fig. D.6 it can be seen that the correlation between the predicted and observed methane production for the Gompertz model is not very corelated in the data available. This is because the model is used to plan the infeed to the biogas plant in order to ensure a constant production and it tells us that the model is used to its limits. Due to the complexity in co-digestion scenarios more research is required in order to calibrate the Gompertz model further. As this would require several tests each lasting for several days a machine learning and a hybrid model were proposed in order to optimize the predictions further.

From Figs. D.4, D.6 and Table D.1 it can be seen that the machine learning and combined model can improve the prediction one day ahead despite the usage of additional feedstocks in the test set. When comparing the machine learning model with the measurements it is clear that it is able to find the relationships between the input

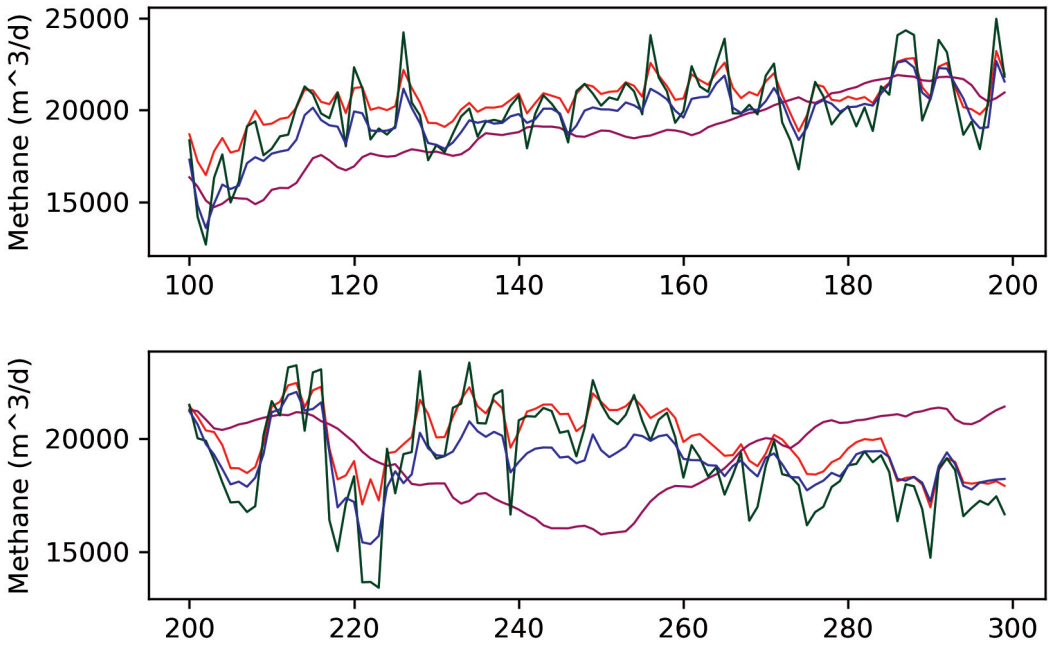


Fig. D.5: Zoomed versions of Fig. D.4

feedstocks and the methane production. However, it typically overestimates the changes in the production: When the production increases the model predicts the increase to be bigger than it is and when the production decreases the model predicts the reduction in the production to be bigger than it is. The hybrid model does not have this problem as the Gompertz model generates a baseline for the production. In other industries it have been found beneficial to develop hybrid models as it can make the model more generalizable and requires less training data [27].

The issue with changes of input data is a typical issue in industrial cases. In this case the quantity of the additional biomasses was relatively low, but if the amount of these feedstocks was increased it could either be added to other feedstocks with similar compositions or the model could be retrained. However, retraining would require several datapoints with usage of the new feedstock.

As it appears from the results the hybrid model can optimize the predictions with 53% when not encountering the offset and with 66% when encountering an offset in the predictions. This is important as surplus methane will be burned off or the gas production will be inhibited with chemistry, as it is too expensive to build storage facilities.

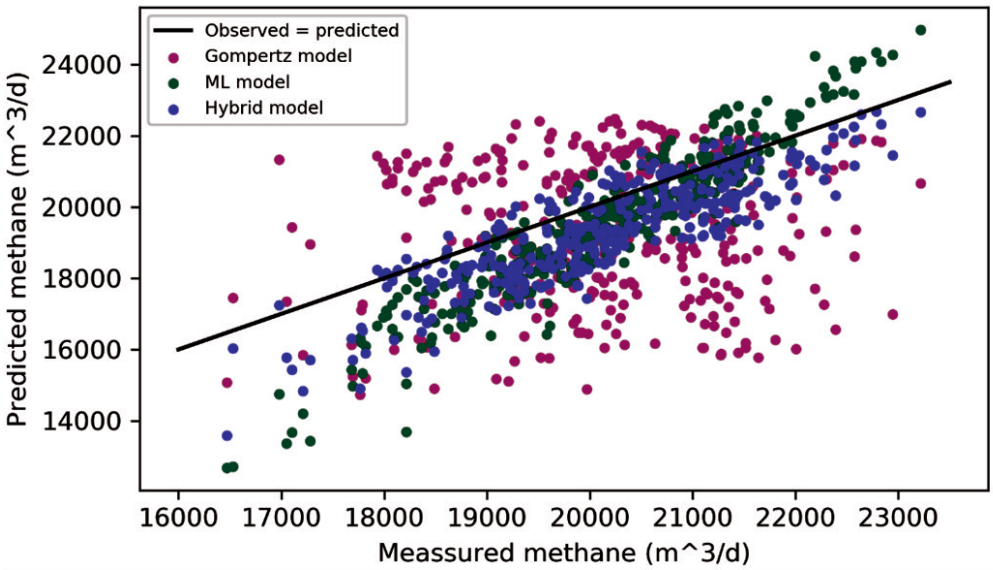


Fig. D.6: Comparison between the observed and predicted values for the three models.

Likewise, if not enough methane is produced the demand is not met. Despite several studies deals with optimization of Gompertz functions in order to develop more precise Gompertz models, they are not suitable for comparison with our model. This is because they are based on experimental studies. The contribution of this paper is to show that we can increase the predictions of the methane production in an industrial setting were some of the parameters are based on expert knowledge as the exact parameters for each of the Gompertz functions are not available.

Two other articles focusing on machine learning based predictions of the methane production in industrial scale biogas plants were found [15, 16]. When compared to [16] the predictions from the machine learning model and the hybrid model presented in this article are quite accurate. [16] replaced the numeric values for the biogas production with values of 0, 1 and 2 for low, medium and high production respectively and obtain an accuracy at 87%. However, due to the low resolution in the output prediction it is much easier to obtain a high accuracy. In our case, the methane production fluctuates between 16,000 Nm^2 and 22,000 Nm^2, which corresponds to one of the categories in [16]. [15] used Random Forest to predict the methane production for time horizons between one and 40 days and obtained R^2 values between 0.88 and 0.82. As the Gompertz model used in our study is used to plan the infeed, in order to obtain as constant a biogas yield as possible, it would not be fair to use R^2 for comparison. This is because optimal

planning entails a lower R^2.

D.5 Conclusion

In this work we have shown that combining a Gompertz model with a machine learning model can optimize the prediction of the methane production one day ahead with up to 66% according to using a Gompertz model alone. This is important as prediction of the methane production is essential in order to keep a constant production of methane, and thereby ensure that the demand is met while avoiding overproduction. If the demand is not met the costumers will have to go elsewhere, which could lead to usage of none sustainable energy sources. If too much methane is produced the surplus will either be burned off or chemistry will be added in order to inhibit the production.

References

[1] S. Xue, J. Song, X. Wang, Z. Shang, C. Sheng, C. Li, Y. Zhu, and J. Liu, "A systematic comparison of biogas development and related policies between china and europe and corresponding insights," vol. 117, p. 109474.

[2] L. M. Colla, A. C. F. Margarites, A. Decesaro, F. G. Magro, N. Kreling, A. Rempel, and T. S. Machado, "Waste biomass and blended bioresources in biogas production," in *Biofuel and Biorefinery Technologies*, pp. 1–23, Springer International Publishing, 2019.

[3] K. Hagos, J. Zong, D. Li, C. Liu, and X. Lu, "Anaerobic co-digestion process for biogas production: Progress, challenges and perspectives," vol. 76, pp. 1485–1496.

[4] T. Scapini, A. F. Camargo, F. S. Stefanski, N. Klanovicz, R. Pollon, J. Zanivan, G. Fongaro, and H. Treichel, "Enzyme-mediated enhanced biogas yield," in *Biofuel and Biorefinery Technologies*, pp. 45–68, Springer International Publishing, 2019.

[5] D. J. Batstone, J. Keller, I. Angelidaki, S. V. Kalyuzhnyi, S. G. Pavlostathis, A. Rozzi, W. T. M. Sanders, H. Siegrist, and V. A. Vavilin, "The IWA anaerobic digestion model no 1 (ADM1)," vol. 45, no. 10, pp. 65–73.

[6] K. Derbal, M. Bencheikh-lehocine, F. Cecchi, A.-H. Meniai, and P. Pavan, "Application of the IWA ADM1 model to simulate anaerobic co-digestion of organic waste with waste activated sludge in mesophilic condition," vol. 100, no. 4, pp. 1539–1543.

[7] D. J. Batstone, D. Puyol, X. Flores-Alsina, and J. Rodríguez, "Mathematical modelling of anaerobic digestion processes: applications and future needs," vol. 14, no. 4, pp. 595–613.

[8] J. A. Arzate, M. Kirstein, F. C. Ertem, E. Kielhorn, H. Ramirez Malule, P. Neubauer, M. N. Cruz-Bournazou, and S. Junne, "Anaerobic digestion model (AM2) for the description of biogas processes at dynamic feedstock loading rates," vol. 89, no. 5, pp. 686–695.

[9] V. Ripoll, C. Agabo-García, M. Perez, and R. Solera, "Improvement of biomethane potential of sewage sludge anaerobic co-digestion by addition of "sherry-wine" distillery wastewater," vol. 251, p. 119667.

[10] L. A. d. Santos, R. B. Valença, L. C. S. d. Silva, S. H. d. B. Holanda, A. F. V. d. Silva, J. F. T. Jucá, and A. F. M. S. Santos, "Methane generation potential through anaerobic digestion of fruit waste," vol. 256, p. 120389.

[11] V. C. Hernández-Fydrych, G. Benítez-Olivares, M. A. Meraz-Rodríguez, M. L. Salazar-Peláez, and M. C. Fajardo-Ortiz, "Methane production kinetics of pretreated slaughterhouse wastewater," vol. 130, p. 105385.

[12] H. Akbaş, B. Bilgen, and A. M. Turhan, "An integrated prediction and optimization model of biogas production system at a wastewater treatment facility," vol. 196, pp. 566–576.

[13] T. Beltramo, M. Klocke, and B. Hitzmann, "Prediction of the biogas production using GA and ACO input features selection method for ANN model," vol. 6, no. 3, pp. 349–356.

[14] F. Tufaner and Y. Demirci, "Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models," vol. 22, no. 3, pp. 713–724.

[15] D. De Clercq, Z. Wen, F. Fei, L. Caicedo, K. Yuan, and R. Shang, "Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion," vol. 712, p. 134574.

[16] D. De Clercq, D. Jalota, R. Shang, K. Ni, Z. Zhang, A. Khan, Z. Wen, L. Caicedo, and K. Yuan, "Machine learning powered software for accurate prediction of biogas production: A case study on industrial-scale chinese production data," vol. 218, pp. 390–399.

[17] B. Gompertz, "XXIV. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. f. r. s. &c," vol. 115, pp. 513–583.

[18] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. van 't Riet, "Modeling of the bacterial growth curve," vol. 56, no. 6, pp. 1875–1881.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," vol. 12, pp. 2825–2830.

[20] L. Breiman, "Bagging predictors," vol. 24, no. 2, pp. 123–140.

[21] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," vol. 55, no. 1, pp. 119–139.

[22] L. Breiman, "Random forests," vol. 45, no. 1, pp. 5–32.

[23] J. H. Friedman, "Stochastic gradient boosting," vol. 38, no. 4, pp. 367–378.

[24] H. Zhang, "The optimality of naive bayes."

[25] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," vol. 63, no. 1, pp. 3–42.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," vol. 2, no. 3, pp. 1–27.

[27] A. Kloss, S. Schaal, and J. Bohg, "Combining learned and analytical models for predicting action effects,"

Paper E

Data-Driven Drift Detection in Real Process Tanks: Bridging the Gap between Academia and Practice

Bolette D. Hansen, Thomas B. Hansen, Thomas B. Moeslund, and David G. Jensen

The paper has been published in the
Water Vol. 14, 926, 2022.

© 2022 by the authors.
The layout has been revised.

Abstract

Sensor drift in Wastewater Treatment Plants (WWTPs) reduces the efficiency of the plants and needs to be handled. Several studies have investigated anomaly detection and fault detection in WWTPs. However, these solutions often remain as academic projects. In this study, the gap between academia and practice is investigated by applying suggested algorithms on real WWTP data. The results show that it is difficult to detect drift in the data to a sufficient level due to missing and imprecise logs, ad hoc changes in control settings, low data quality and the equality in the patterns of some fault types and optimal operation. The challenges related to data quality raise the question of whether the data-driven approach for drift detection is the best solution, as this requires a high-quality data set. Several recommendations are suggested for utilities that wish to bridge the gap between academia and practice regarding drift detection. These include storing data and select data parameters at resolutions which positively contribute to this purpose. Furthermore, the data should be accompanied by sufficient logging of factors affecting the patterns of the data, such as changes in control settings.

E.1 Introduction

With increased focus on the United Nations Sustainable Development Goals (SDGs) and increasing energy prices, there has been an increased interest in optimizing the performance of Wastewater Treatment Plants (WWTP) and Wastewater Recovery Facilities, prospectively referred to as WWTPs in this article. Optimizing the operation of WWTPs is the topic of several studies [1] where some of the more complex control systems include the energy usage and economics [2, 3]. Several companies offer software for real time control of WWTP; however, few focus on the data quality involved [4]. This is problematic as sensor drift can induce decreased total N removal or over aeration, entailing a large increase in energy consumption. Bias in sensor data can easily counteract the energy reduction and cost savings obtained by advanced automatic control [5].

Drift in sensors is a commonly known problem. Due to fouling, optical Dissolved Oxygen (DO) sensors can easily be biased with one mg/L within a month [6]. In calibration data from two WWTPs, examples of drift with more than one mg/L can be found for both ammonia and potassium sensors. In plants, where the NH_4 level is typically below 3 mg/L in the outlet, 1 mg/L is a large deviation. Especially a positive drift in ammonia sensors is subject to increased costs at the utilities, and from an industrial perspective it should be of high priority to detect these faults.

Faulty sensor data is a problem in several different sectors. Teh et al. [7] reviewed 57 papers on sensor faults and methods used for detection and correcting faulty data. The faults included outliers, missing data, bias, drift, noise, constant values and the sensor being stuck at zero. The methods used for drift detection in the reviewed papers

were Principal Component Analysis (PCA), Artificial Neural Network (ANN), Ensemble Classifiers and Dempster-Shafer Theory and Mathematical Modelling (only one paper from 2008). Furthermore, PCA, calibration-based methods, PCA based methods and Kalman filter-based methods were used for drift detection and correction [7].

Several data-driven approaches for wastewater treatment operation have been developed and presented in the literature; however, in a review, Corominas et al. [4] report that only 16 percent of the developed solutions resulted in a commercial product, and seven percent were commercialized without full-scale testing. Within fault detection the most popular approach is PCA (27 papers) followed by ICA (9 papers) and clustering (7 papers) [4].

Within fault detection in process tanks, several studies have been made. Baklouti et al. [8] used univariate statistics to detect bias, drift and variating magnitudes in Dissolved Oxygen (DO) sensors. They introduced bias of two mg/l, drift with a slope of 0.005 and variating signal magnitude of three standard deviations. The calculation of 700 datapoints was made without information on the actual corresponding sampling frequency. Despite stating that it was in general problematic that models did not encounter seasonal changes, etc., the authors tested their model in simulated dry weather data [8].

In 2002 Thomann et al. [9] suggested using control charts to make it easier for WWTP staff to detect drift, outliers and shifts based on four months of collected data. Newhart et al. [10] stated that control charts are well suited for monitoring single variables which only contain a low degree of noise, having been measured on a daily to monthly basis.

Baggiani and Marsili-Libelli [11] used PCA combined with moving windows, T^2 and Q statistics, as well as threshold, to detect spikes and sensor faults in data from a real plant and obtained performances of 100% and 84% depending on the window size used [11]. It is worth noticing that the spikes and faults exemplified in the paper are very distinctive compared to the signal amplitude.

Alferes et al. [12] used PCA over six days and found two deviations in the PCA analysis. The first was explained by a high unusual discharge and the second was related to a turbidity sensor. It is worth noticing that when observing the turbidity data, another case is eye catching; however, it is found in the PCA analysis.

Cheng et al. [13] used kernel PCA (KPCA) and one-class support vector machine to detect anomalies in the inflow components of a real plant over seven years and obtained better results than when using linear PCA and K-nearest-neighbours.

Huang et al. [14] proposed a method for anomaly detection in a WWTP at a paper mill; however, only one case with faulty behaviour was available and the process was in a more closed and controlled environment than a normal treatment plant. This is indicated by specific time slots for different processes to take place.

In 2020 and 2021, several methods for fault detection in WWTPs were proposed in the literature. Ba-Alawi et al. [15] used stacked denoising autoencoders for detection

of drift, bias, precision degradation and complete failure. The method was evaluated on simulated dry weather data and the authors state that the method was superior to existing methods and can reduce operating costs and improve the monitoring of the influent [15]. Kazemi et al. [16] showed that incremental PCA was able to distinguish between time varying events and faults in simulated data, while Kazemi et al. [17] investigated a number of technics including Support Vector Machine, Ensemble Neural Network and Extreme Learning and found that they performed better than a PCA based method after testing on simulated data. Luca et al. [18] applied PCA and statistic for fault detection in DO sensors in simulated data and stated that the method was successful in detecting the faults. Mali and Laskar [19] proposed an optimized Monte Carlo deep neural network and were able to detect faults of low magnitude in simulated data. Xu et al. [20] proposed a version of ICA called complex-valued ICA. The method was both evaluated for simulated data and for data from a real plant. In the real case, the authors had 213 samples of which 45 were from normal operation, and these were used for training; however, these samples were also included in the test set. The authors stated that this method could obtain more accurate, intuitive and efficient fault detection. Klanderman et al. [21] proposed a method based on auto correlation and Fused Lasso. The method was trained on an in-control data set and tested on a simulated data set with introduced faults and data from a real plant, which contained one fault that they were able to detect. Mamandipoor and Majd [22] possessed 11 months of data from 12 sensors in a real plant. The data were classified according to faulty NH_4 data by an expert and a Long Short-Term Memory Network was developed and outperformed PCA-SVM. Cecconi and Rosso [23] used ANN to predict the NH_4 concentration and used PCA along with Shewhart monitoring charts for detection of the variation between measured values and predicted values. This study was based on more than one year of data from a real plant. Six sensors were installed in the plant including two NH_4 sensors. The sensors were cleaned on a weekly basis and calibrated if there was a difference detected of more than 15% between the sensor and the reference. The faults considered in the study were sensor faults caused by wrong calibration, process anomaly and drift. For testing, three types of faults were introduced in real data. The suggested approach was able to detect the faults and the ANN prediction could be used for process control when a fault was detected [23]. Anter et al. [24] used fuzzy swarm intelligence and chaos theory to detect faults in a real data set from 1993 available at the UCI Machine Learning Repository [24]; however, details on the fault types detected are not described.

Except for Cecconi and Rosso [23] and Mamandipoor and Majd [22], none of the solutions proposed in 2020–2021 reflect contemporary conditions met at WWTPs, and while several papers acknowledge that there is a gap between the solutions in academia and in the real world [4], there is a lack of knowledge when it comes to implementing data-driven approaches in real WWTPs.

The aim of this paper is to bridge the gap between academia and practice by applying

different approaches for machine learning to real-world data sets, and thereby identify challenges hindering implementation of data-driven drift detection at normal operating WWTPs. The main contribution of this work is identification of the shortcomings between academia and practice together with recommendations for future data usage and management obtained in collaboration between data scientists and water professionals. To ensure that the recommendations are as relevant as possible for both researchers and managers, this work is based on data available from operating WWTPs, and no extra data acquisition was made. This entails the data being of lower quality than if it is acquired with the specific purpose of developing algorithms for drift detection.

The remainder of this paper is structured as follows. Section E.2 contains information on the data and approaches investigated in this study. Section E.3 contains the results and description of how to interpret these. Section E.4 is a discussion of the results and Section E.5 contains perspectives on drift detection from both academic and practical perspectives. These perspectives are accompanied by recommendations for the future. The paper is concluded in Section E.6.

E.2 Materials and Methods

This section contains an overview of the available data and the applied methodology for anomaly and fault detection. With inspiration from the literature, several methods for anomaly and fault detection were initially considered; however, it became clear that many of the considered methods were not practically applicable. As the purpose of this paper is to bridge the gap between academia and research, descriptions of the unsuccessful methods have been included in this section, together with a description of why they were not successful in this case. Lastly, a description of how the detected anomalies are accessed is included.

E.2.1 Data

Data from three plants were available for this study. The resolution of the data was one sample per minute, and two of the WWTPs had a log accessible with calibration information. One of the WWTPs had one process tank (PCT) while the remaining two WWTPs had two PCTs. An overview of the PCTs can be seen in Table E.1.

The control strategy for aeration of WWTP2 PCT1 was based on alternating operation where the air pumps turned on and off based on ammonia set points. The remaining PCTs were controlled by PID controllers. A PID controller tries to obtain a constant NH_4 level which is defined by a set point. The PID controller adjusted the amount of aeration based on the difference between the NH_4 concentration and the set point for the NH_4 concentration. How fast the PID adjust the aeration depends on three constants. This control strategy is beneficial as it allows for a more constant concentrations in the PCT.

Table E.1: . Overview of the available data.

Process Tank	Data Period	Log *
WWTP 1 PCT 1	From 25 January 2021	4 measurements
	To 14 September 2021	1 calibration
WWTP 2 PCT 1	From 13 June 2020	Not available
	To 14 September 2021	
WWTP 2 PCT 2	From 13 June 2020	Not available
	To 14 September 2021	
WWTP 3 PCT 1	From 1 February	3 measurements
	To 14 September 2021	3 calibrations
WWTP 3 PCT 2	From 1 February 2021	3 measurements
	To 14 September 2021	2 calibrations

* Logs were available until 22 May 2021.

Multiple parameters were available for the three plants including flow to the plant, NH_4, NO_3, DO, K and SS, while other parameters variated between the plants such as information on the aeration, if N_2O was measured, etc. The parameters flow, NO_3 and DO are highly related to the NH_4 level in the plant. Furthermore, plots of the data did not indicate that the remainder of the parameters, which were available for all the PCTs, should be included. Therefore, it was decided to focus on the parameters flow, NH_4, NO_3 and DO.

For two of the plants, lab measurements and calibration logs were kept for the NH_4 sensor and the NO_3 sensor. From the logs it could be seen that a drift of the NH_4 sensor of 0.5 mg/L was accepted, while a drift of 1 mg/L was accepted for the NO_3 sensor. In the log calibration, events were noted down; however, this was done manually. In some cases, it was stated that a sensor was adjusted, but it was not stated which sensor.

E.2.2 Machine Learning Approaches

As described in Section E.1, several different data-driven approaches for drift and fault detection in WWTPs exist; however, these methods cannot be directly applied to the data available for this study.

A characteristic for almost all the methods presented in the literature is that they have been developed and tested on data sets where the faults are already known, either because the faults have been simulated or because the data come from well monitored WWTPs. Such labelled data sets are rarely available for normal WWTPs, which is also the case for the data available for this study. Therefore, it was sought to obtain a labelled data set for drift by manually labelling the data in the PCT with alternating operation, as this was the easiest PCT to assess. To do this, an interactive software tool for systematic labelling of each aeration cycle was made. Each aeration cycle could

then be labelled as OK or as a fault type. However, during the labelling process it was observed that it was hard to label the data without introducing several faults. Reasons for this included the operators changing the control settings instead of calibrating the sensors and the utility accepting the NO_3 sensor to drift with up to 1 mg/L without considering it as an anomaly.

The lack of labelled data entails that it is not possible to use traditional supervised learning. Another approach initially tested was predicting each parameter based on one class learning. Thereby the variations between the prediction and the measurement would be the fault. However, this task was complicated by the fact that the immediate previous measurement could not be used as input for the predictive machine learning algorithm, as drift develops over time. Thereby most of the drift would also be present in the immediate previous measurement and consequently, the algorithm would predict the measured value and not the real value. Therefore, experimenters tried to train a Random Forest model, which is an ensemble method, on the first 80% percentage of the data and test it on the remaining 20% for each PCT. For this task, daily average values were used to neglect normal variations during a day such as increased flow in the rush hours, rainfall and when the aeration pump was activated. This approach showed low performance of the algorithm and the main bottleneck for obtaining better results was the large variations in control setting at the PCTs. Therefore, experimenters decided to use unsupervised learning.

E.2.3 Unsupervised Learning Algorithms

As described in Section E.1, a commonly used unsupervised method for fault detection in WWTPs is PCA. Therefore, the data sets were normalized according to the standard deviation and examined through PCA. All combinations of Principal Components (PC) were then plotted per day and visually inspected. It was observed that the patterns changed over time, and especially changes in control settings caused the patterns to change. Changes in patterns when plotting principal components were also observed by Alferes et al. [12] who only looked at a few days of data. However, when considering several months of data this approach is not efficient, as the evaluation is based on visual inspection. Furthermore, it was found to be much simpler to interpret the data and changes by simply plotting all combinations of parameters per day. It was also investigated if using PCA on daily values could be used to detect anomalies. In this connection, it was tested if faults and anomalies could be removed by removing the least contributing PC; however, the anomalies were present in all principal components and this approach did not work.

More complex solutions such as deep auto-encoders were considered; however, based on the results with one class learning it was not expected that this approach would be efficient. Therefore, for the purpose of this study, it was found more relevant to use a simpler and more transparent approach.

The last approach considered was to use the Local Outlier Factor (LOF) [25] on daily values. Initial results showed that this method gave the most promising results, for which reason it was chosen to use LOF.

Local Outlier Factor

LOF is an unsupervised learning algorithm which measures the distance to a certain number of nearest neighbours and uses this distance as a measure of anomaly.

For the LOF it was decided to use daily data. This was done to neglect the large variations in inflow, wastewater composition and aeration periods during a day. After averaging the data to daily signals, the data were scaled according to the standard deviation. In the specific implementation of the LOF, the distance to the 20 nearest neighbours was used to calculate the LOF. To ensure that the method can be applied in real time, the LOF was implemented as a Moving LOF filter, where the LOF for a given day was based on the 99 previous days.

A threshold of two was applied to the Moving LOF, and all datapoints exceeding the threshold were considered as abnormal. All periods of abnormal behaviour were subsequently assessed.

E.2.4 Assessment of Anomalies

Several different types of anomalies were present in the data. For gaining an overview, the anomalies were categorized into five general groups, namely missing data, increased presence (referring to increased flow or increased presence of NH_4, NO_3 or DO), change in control settings, sensor drift or over aeration and other. In some cases, multiple anomalies were present, and in these cases it was evaluated, which was the primary reason for the detection. For instance, there could be a scenario where a sensor has drifted but nothing is detected until an increase in flow is present and after the next day, nothing is detected again. In such a case the anomaly is annotated as an "increased presence", even though the reason for the anomaly to be detected might be a combination of the drift and the increase in flow.

Plots were made for each PCT showing the anomaly category and relevant data examples were plotted. Additionally, examples of longer periods of anomalies not reaching the threshold were plotted.

E.3 Results

This section contains a description of the results of the anomaly detection. The results for each of the PCTs are presented in Sections E.3.1–E.3.5. For each of the PCTs, examples of anomalies have been highlighted. The examples have been selected so that as many different scenarios as possible are shown, to give insight into as many scenarios

as possible. Details on all observations are presented in Table E.2. Furthermore, general observations are described in Section E.3.6.

E.3.1 WWTP1 PCT1

The data available for WWTP1 PCT1, calibration and lab measurements, Moving LOF and the anomalies detected using the Moving LOF and thresholding can be seen in Figure E.1. The detected anomalies are colour coded according to the anomaly type observed. In the figure, it is worth noticing that several different control settings have been used in the first period for which the data was available. Consequently, the algorithm does not consider this type of control setting as an anomaly if it is strongly present in the LOF window. This might be the reason that changes in control settings in the middle of May 2021 were not detected. As seen in the figure, most of the detected anomalies were related to increased presence of one or more of the parameters.

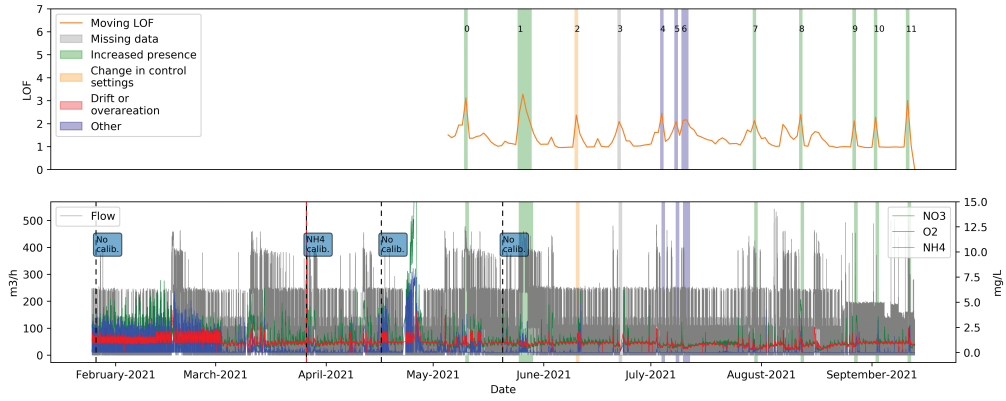


Fig. E.1: WWTP1 PCT1. In the upper graph, the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 11, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4 and the calibration data available from the plant.

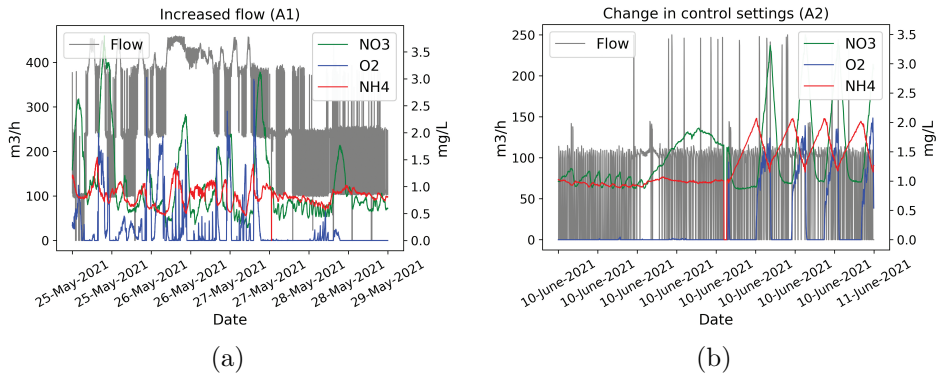
Examples of an anomaly caused by increased flow and an anomaly caused by change in control settings are presented in Figure E.2. Further details on the anomalies can be found in Table E.2.

E.3.2 WWTP2 PCT1

Figure E.3 shows the data, Moving LOF and detected anomalies for WWTP2 PCT1. The control strategy for WWTP2 PCT1 is based on alternating operation and the figure

Table E.2: Overview of the anomalies detected in the five PCTs when using Moving LOF and a threshold of two.

WWTP 1 PCT 1	WWTP 2 PCT 1	WWTP 2 PCT 2	WWTP 3 PCT 1	WWTP 3 PCT 2
0. Increased flow	0. Missing data	0. Increased flow combined	0. Change in PID. There is an	0. Increased flow
1. Increased flow	1. Increased flow + NH4	with changed control	increased flow starting the	1. Change in
2. Change in PID	drift (up)	settings the previous day	day before and continuing	control settings
3. Missing data	2. Increased flow + NH4	1. Increased flow	two days after the anomaly	2. Increased flow
4. Other	drift (up)	2. Increased flow	was detected. In the period	3. Increased flow
5. Other	3. NH4 drift (up)	3. Increased flow	of the anomaly, the pattern	4. Increased NO3,
6. Other	4. NH4 drift (up)	4. Increased flow and	of the sensors changed,	low DO
7. Increased flow	5. Increased flow	increased NO3 unrelated	indicating change in	5. Increased flow,
8. Increased NH4,	6. Increased flow	to flow	control settings.	increased NO3,
NO3 and DO	7. High concentrations of	5. Increased NO3	1. Increased NH4	low DO
9. Increased NO3	NH4 and NO3	6. Increased NO3	concentration due to	
and DO	8. High concentrations of NO3	7. Increased flow	missing aeration	
10. Increased NO3	present or NO3 sensor	8. Data shows low flow, very	2. Increased flow, inducing	
and DO	drifted (up)	large amounts of DO and	high NH4, NO3 and	
11. Increased flow	9. Increased flow	increasing NH4.	DO concentrations	
	10. Increased flow	9. Increased flow	3. Increased flow, inducing	
	11. Increased flow	10. NO3 and NH4 the first day,	high NH4 and NO3	
	12. NH4 drift (up)	increased flow the	concentrations	
	13. This anomaly starts as NH4	second day	4. Increased NH4	
	drift (up). The second day	11. Increased concentrations of	concentrations inducing	
	data is missing for almost	NH4, NO3 and DO.	high NO3 concentrations	
	13 h. Hereafter, the lower	Possible because the other	5. Increased NH4	
	setpoint seems to be slightly	PCT at the WWTP was out	concentrations inducing	
	increased with 0.1, which	of operation, see anomaly	high NO3 concentrations	
	handles the problems with	14 for WWTP2 PCT1	6. Increased flow, inducing	
	over-aeration. The last day	12. Increased flow	high NH4 and	
	of the anomaly is due to an	13. Increased concentrations of	NO3 concentrations	
	increased flow.	NH4, NO3 and DO. One		
	14. All parameters are low	day with increased flow.		
	except for the flow. Maybe	Possible because the other		
	this PCT has been out of	PCT at the WWTP is out of		
	operation or experiments	operation, see anomaly		
	had been performed.	17–18 for WWTP2 PCT1.		
	15. Increased flow	14. Increased concentrations of		
	16. Increased flow	NH4, NO3 and DO. One		
	17. Low parameters, see 14	day with increased flow,		
	18. Low parameters, see 14	see 13.		
	19. Increased flow	15. Increased flow		
	20. High levels of NH4 present	16. Increased flow		
	day the first day, increased	17. Increased flow		
	flow the second day	18. Missing data		
	21. NH4 drift (up)	19. NH4 sensor drifted (up)		
	22. NH4 drift (up)	20. Increased flow		
	23. NH4 drift (up)	21. Increased flow		
	24. Change in setpoint. In the	22. Increased flow		
	period up to the detection	23. Increased flow		
	of this anomaly the			
	setpoints were increased			
	multiple times. This also			
	happened two days before			
	this anomaly was detected.			
	The day before this			
	anomaly was detected, the			
	setpoints were decreased			
	inducing over aeration. The			
	day after the setpoint was			
	increased again, which was			
	the case for the remainder			
	of the anomaly. The LOF			
	decreased over time as it			
	learnt the new behaviour			
	25. Missing data			
	26. Increased flow and NH4			
	concentration			
	27. Increased flow and NH4			
	concentration			



(a) (b)

Fig. E.2: Examples of anomalies detected in WWTP1 PCT1. (a) increased flow, (b) change in control settings.

shows that several anomalies caused by sensor drifts or over aeration were found by the algorithm.

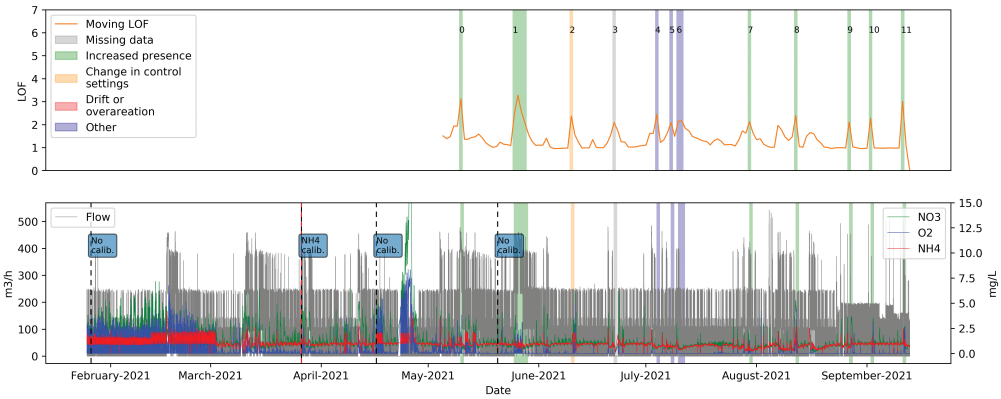


Fig. E.3: WWTP2 PCT1. In the upper graph the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 27, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4. No calibration data were available from the plant.

For WWTP2 PCT 1, in which alternating operation was used, 27 anomalies were detected. Of these, 15 were primarily detected due to increased presence of flow, NH_4, NO_3 or DO. However, in two of the cases the NH_4 sensor had already drifted but an increase in flow was the factor which made it exceed the threshold (Anomaly 1–2 in Figure E.3).

Examples of missing data, NH_4 sensor drift, high NO_3 levels, increased flow, increased presence of NH_4, an anomaly categorized as other (which most likely is caused by the PCT being out of operation or experiments performed at the plant) and change in control settings are presented in Figure E.4. Further details on all the anomalies detected in the PCT are presented in Table E.2.

In addition to the anomalies exceeding the threshold, some longer time periods with increased LOF were observed for the PCT with alternating operation. The increase in LOF was associated with a NH_4 sensor drift, which the operator compensated for by changing the set points. An example of this can be seen in Figure E.5. It is worth noticing that the NO_3 and DO levels gradually increased in the period before a change in setpoints for NH_4 and suddenly decreased after the changes. This is especially clear in the period from 23 June 2021 to 1 July 2021.

E.3.3 WWTP2 PCT2

Figure E.6 shows the data, the Moving LOF and the detected anomalies for WWTP2 PCT2. As seen in the figure, increased presence of the different parameters is the most common reason for anomalies; however, increased presence of some of the parameters can be caused by other factors, such as change in the usage of the plant. For instance, anomalies 11, 13 and 14 coincide with anomalies 14, 17 and 18 in WWTP2 PCT1, which are most likely caused by PCT1 being out of operation and thereby cause an increased pressure on this PCT. Figure E.6 also shows that several different control settings were used in the beginning of the data collection. However, as this was within the first 99 days of the data collection, the Moving LOF could not give the outlier score of the data for this period. When considering the anomalies detected by the Moving LOF, anomaly eight differs from previously elaborated anomalies. It has been classified as 'other', and the anomaly is most likely caused by a fault in the DO sensor as a constant increase in NH_4 concentration and low NO_3 concentration indicate a lack of DO in the PCT. A detailed plot of anomaly eight is shown in Figure E.7. Further details on the anomalies detected in WWTP PCT2 can be found in Table E.2.

E.3.4 WWTP3 PCT1

The data, lab measurements and calibrations, Moving LOF and detected anomalies for WWTP3 PCT1 are shown in Figure E.8. For this plant two anomalies distinguish themselves. These are the anomalies zero and one. Anomaly zero is observed during a longer period of increased flow. In the parallel PCT the full period of increased flow has been detected as an anomaly; however, for this PCT only one day during the increased flow was detected. During this day, changes in patterns indicated that the control settings were changed, possibly to deal with the increased flow. For anomaly one, it was observed that there were several hours with no DO, constantly increasing NH_4 levels and

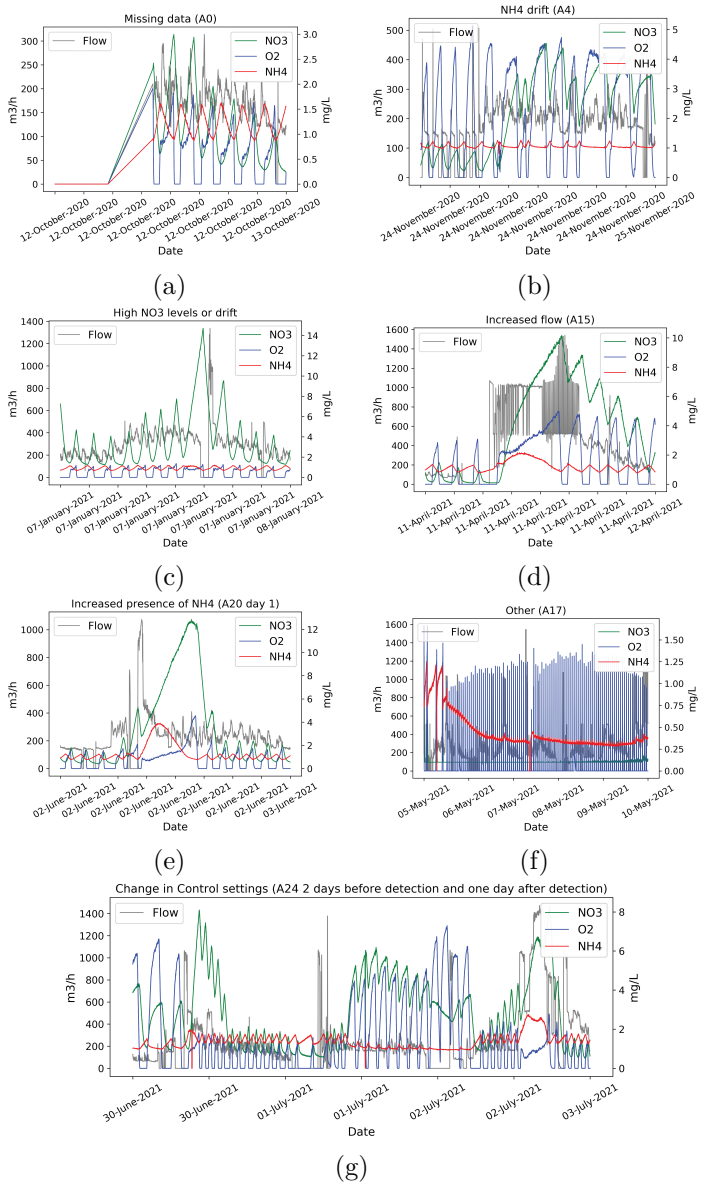


Fig. E.4: Examples of anomalies detected in WWTP2 PCT1. (a) missing data (W2P1A0), (b) NH_4 drift (W2P1A4), (c) High NO_3 levels or drift, (d) Increased flow (W2P1A15), (e) Increased presence of NH_4, (f) Other (W2P1A17), (g) Change in control settings (two days before onset of W2P1A24 until 1 day after detection).

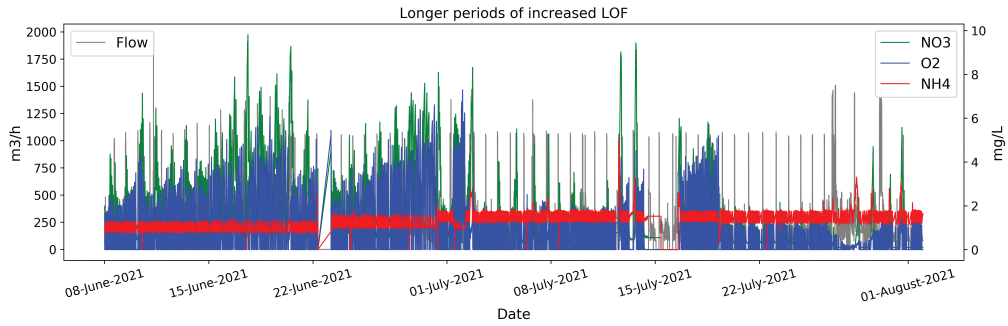


Fig. E.5: Long period of increased LOF in WWTP2 PCT1. The pattern in the data indicates that the NH_4 sensor had drifted and that the operator of the plant subsequently adjusted the setpoint instead of calibrating the sensor.

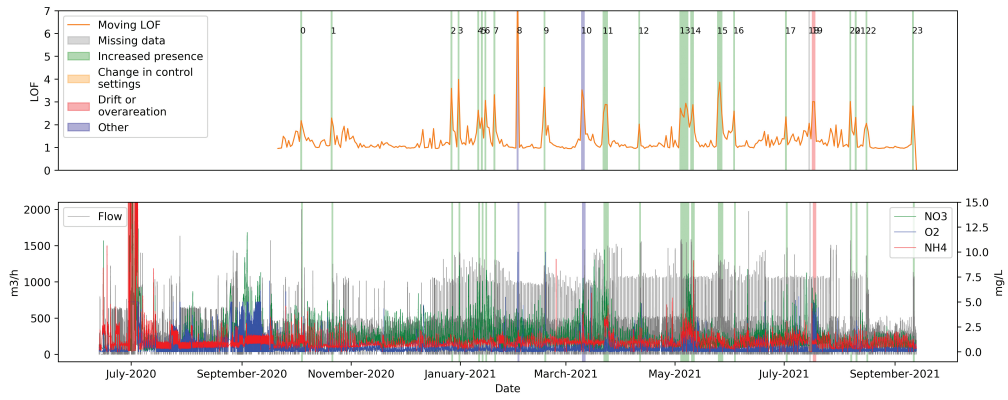


Fig. E.6: WWTP2 PCT2. In the upper graph the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 23, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4. No calibration data were available from the plant.

low NO_3 levels, indicating that the aeration pump had been out of operation. Detailed plots of anomaly zero and one can be found in Figure E.9.

E.3.5 WWTP3 PCT2

The data, measurements, and calibration as well as Moving LOF and detected anomalies for WWTP3 PCT2 are shown in Figure E.10. In this PCT, anomaly one differs from previous observations. The pattern of the data indicates that the control settings were changed to reduce the NH_4 concentration in the outlet; however, for a while this entails

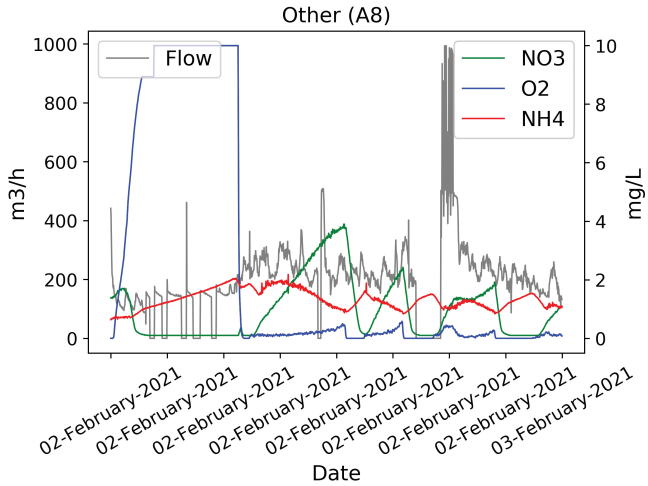


Fig. E.7: Example of an anomaly detected in WWTP2 PCT2. The anomaly is categorized as other. The anomaly is most likely caused by a fault in the DO sensor.

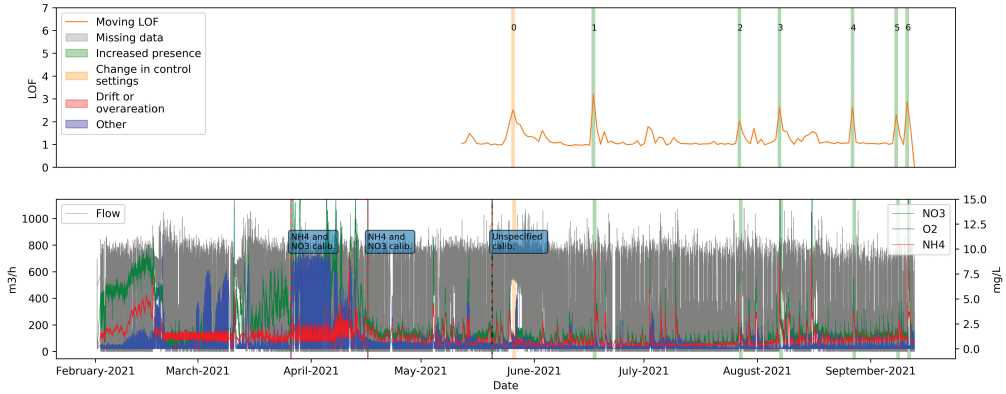


Fig. E.8: WWTP3 PCT1. In the upper graph the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 6, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4 and calibration data available from the plant.

that the air pump is constantly active as the NH_4 level does not decrease. Hereafter, a more normal pattern is observed again. A detailed plot of anomaly zero is presented in Figure E.11. Another observation made for this PCT is a low concentration of DO, which is positive, as it indicates that all the DO has been used.

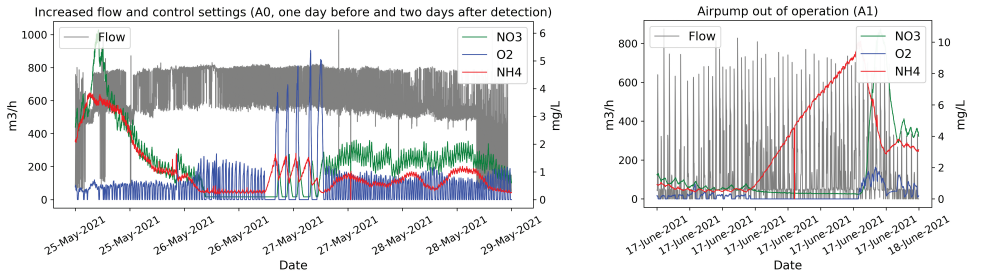


Fig. E.9: Examples of anomalies detected in WWTP3 PCT1. (a) Anomaly zero, increased flow prompting the operator to change the control settings. Only the 26 of May is detected as an anomaly, (b) Increased NH_4 levels due to lack of aeration

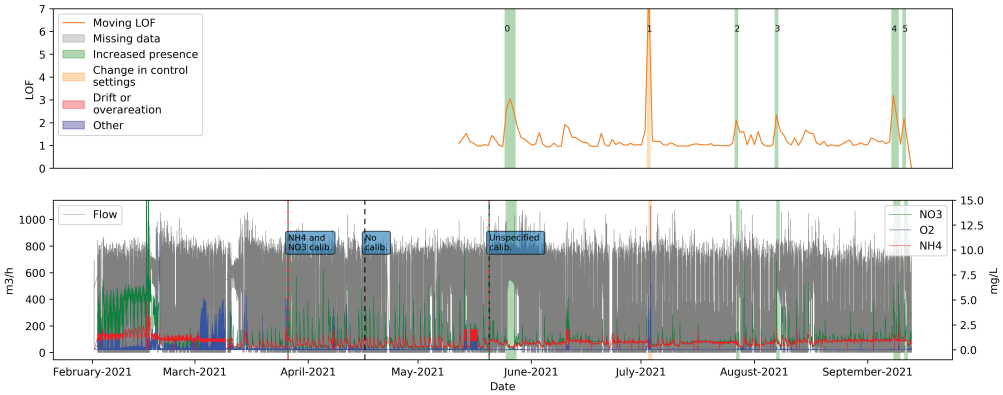


Fig. E.10: WWTP3 PCT1. In the upper graph the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 5, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4 and calibration data available from the plant.

E.3.6 General Observations

An overview of the observations and detailed descriptions for each of the PCTs is presented in Table 2.

Generally, it is worth noticing that the easiest drift to detect was the NH_4 sensor measuring too high values during alternating operation. In these cases, indications of drift could be visually observed in the data before the threshold was exceeded. However, reducing the threshold would also introduce more anomalies due to increased flow, which can be considered as false positives.

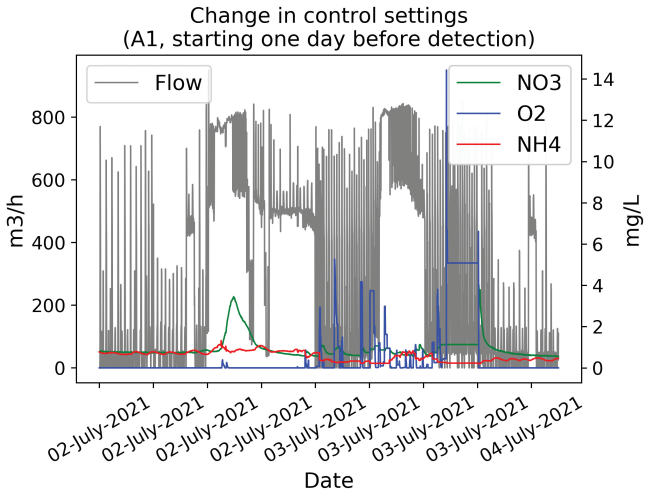


Fig. E.11: Example of an anomaly detected in WWTP3 PCT2. The anomaly is most likely caused by change in control settings.

In several cases, faults such as drift were included in the data window used by the Moving LOF. Thereby faults would not be as abnormal for the Moving LOF as if a clean data set was available.

In some cases, increased amounts of NO_3 were detected indicating that the NO_3 sensor measured too high values. However, it was not possible to evaluate if the sensor measured within the accepted range of ± 1 mg/L.

In general, the algorithm did not detect drifts when the sensors measured too low of values. However, when reviewing the data manually there were indicators of the NO_3 sensor measuring too low values. Generally, drift towards low concentrations is harder to detect than drift towards high concentrations, as there is a natural limit in how much a drift towards zero can be distinguished from normal behaviours. Furthermore, reaching a low number of particles in the outlet of the plant is also an indicator of optimal operation of the plant.

Several cases of changes in control settings were detected as anomalies. Changes in control settings are not faults; however, they change the basis for any type of data-driven algorithm significantly.

Regarding missing data, it is worth noticing that this type of anomaly can easily be detected using rule-based methods. This type of anomaly was present several times but was not removed before applying the LOF algorithms, as daily values were based on average values for a given date.

Other Observations

The problem with NH_4 sensors measuring too high values is that this can entail plants over-aerating, which is expensive. An increase in multiple data parameters was observed when the NH_4 sensor measured too high values in alternating operation. Thereby it is possible that an increase in the daily average of concentrations could indicate NH_4 drift and that NH_4 sensor drift hereby could be detected by utilization of a simple rule-based algorithm, such as alarming, if a threshold is exceeded for a longer period. An overview of the average values per day for WWTP2 PCT1 can be seen in Figure E.12.

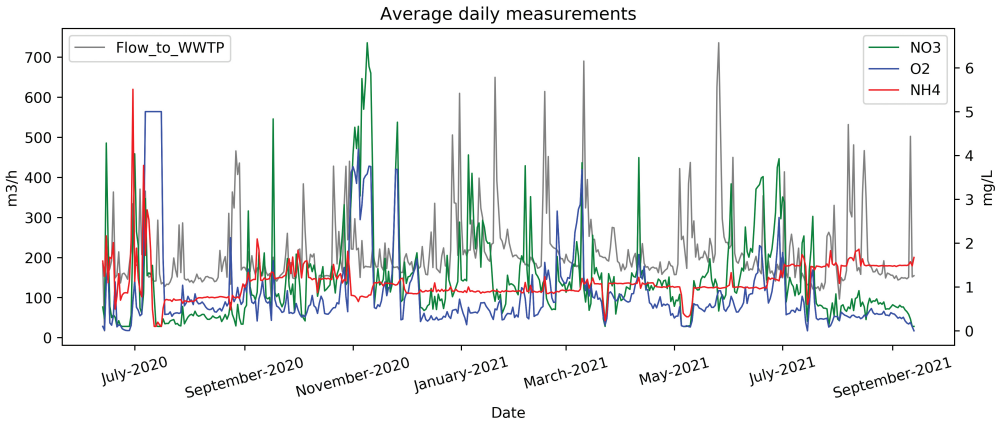


Fig. E.12: Average daily values of Flow, NH_4, NO_3 and DO in WWTP2 PCT1.

E.4 Discussion

This section contains a discussion of the results.

Based on initial tests described in Section E.2, it was chosen to use LOF combined with a threshold. LOF was chosen despite several more complex methods having been previously presented in the literature, as it was not possible to apply the more complex methods to the data, due to low data quality in real WWTPs. The results showed that it was possible to detect anomalies; however, the most detected anomalies were related to increased presence of flow or substances. Increased presences of flow and substances are not faults. It is problematic that the most detected fault is increased in different parameters as previously published papers primarily focus on dry weather data, simulated data or in-control data, because this means that the developed methods do not encounter the challenges met at real plants. It is important to be aware of this

shortcoming as it entails that automatically detecting outliers as faults most likely will entail that valid datapoints are considered as faults whereas actual faults are overlooked.

The results presented in this paper are highly dependent on the used threshold and the window length. If one lowers the threshold more anomalies would be detected; however, it would also entail a larger number of anomalies caused by increased flow or substances. For plants with more than one PCT, the number of anomalies caused by increased flow or increased presence of substances could be reduced by removing detections which are present in both PCT simultaneously, as can be seen in Tables E.3 and E.4. It is important to mention that increased presence of substance can both be caused by external factors, such as industrial discharges, which are anomalies, and internal factors such as sensor drifts, which are faults. This makes it complicated to distinguish between anomalies and faults.

Table E.3: Overview of anomaly types detected in WWTP2 for the two PCTs. The table shows how many of the different types of anomalies are detected in total for each of the PCTs. Furthermore, it shows how many anomalies are detected if anomalies present in both PCTs are removed. The numbers in parentheses are the number of anomalies which have some overlap with the other PCT, but the period of the detections is not similar.

Anomaly type	PCT1, All Detections	PCT1, Overlapping Detections Removed	PCT2, All Detections	PCT2, Overlapping Detections Removed
Increased flow	8	1(1)	13	2(4)
Increased presence incl. substances	5	2(2)	6	3(3)
Missing data	2	1(1)	1	0(1)
Drift	6	6	1	1
Change in control settings	1	(1)	0	0
Combinations of multiple types	3	1(1)	1	1
Other 3 (3) 2 1 (1)	3	(3)	2	1(1)

Table E.3 shows that if anomalies present in both PCTs are removed for WWTP2, the anomalies caused by increased flow and increased presence of other substances will change from 0.48% of the detected anomalies to 27% of the anomalies for PCT1 and from 79% to 63% for PCT2. For PCT1 sensor drift increases from 21% to 56% and for PCT 2 the percentage of anomalies caused by drift would increase from 4% to 17%. Due to the low number of anomalies for WWTP3, it would be misleading to make similar calculations for this plant.

It is important to mention that drifts primarily were detected in the PCT with alternating operation and for this plant several of the cases with drift could be detected earlier by visual inspection of the data, if good visualization tools were provided.

Table E.4: Overview of anomaly types detected in WWTP3 for the two PCTs. The table shows how many of the different types of anomalies are detected in total for each of the PCTs. Furthermore, it shows how many anomalies are detected if anomalies present in both PCTs are removed. The numbers in parentheses are the number of anomalies which have some overlap with the other PCT, but the period of the detections is not similar.

Anomaly type	PCT1, All Detections	PCT1 Overlapping Detections Removed	PCT2, All Detections	PCT2 Overlapping Detections Removed
Increased flow	0	0	3	(1)
Increased presence incl. substances	6	2(1)	2	(1)
Change in control settings	1	(1)	1	1

E.5 Perspectives and Recommendations

Drift in sensors, especially in NH_4 sensors, causes non-optimal operation at WWTPs, which can induce inefficient N removal, increased resource usage and extra economic costs. Sensor drift is common in most WWTPs and this needs to be handled for resource optimization. Several data-driven algorithms for drift detection have been developed in academia; however, they often remain as academic projects entailing a gap between academia and the real world. This gap was investigated by applying different data-driven solutions on real WWTPs. The results showed a number of significant challenges in real data, which have not been handled in the current academic solutions.

This study showed that it was not possible to obtain valid, consistent and precise labelling of sensor drift in the data after its collection. Only extreme sensor drifts and sensor drifts induing over aeration could be identified in PCTs with alternating operation. The study also showed that NH_4 drift, to some extent, can be identified using unsupervised learning. It also shows that more anomalies were detected in the WWTP with alternating operation than in the plants with PID control systems. However, the performance does not meet the needs at the WWTPs.

The challenges described in the Method and the Result sections clearly illustrate that action needs to be taken if optimal operation at the plants should be widespread among WWTPs. In the following, the challenges met at the plants are described and discussed in Sections E.5.1-E.5.5. Section contains a discussion of whether it is feasible to acquire a sufficient data set for data-driven drift detection. Section contains perspectives on other ways of handling drift while Section contains a discussion of why well considered data acquisition from the plants is still important. Section contains perspectives on how the data available today can still create value.

E.5.1 Variations between Plants

There is a large variance between plants. For instance, there is a large variation in the design of WWTPs, the sensors installed at the plants, the composition of raw wastewater, the control strategies and the data stored from each plant. It is worth noticing that factors such as the composition of the wastewater can induce the NH_4 sensor to drift earlier in one plant than in another. The control methods and settings vary largely between different plants, due to factors such as variation in the discharge requirements of the plants.

From the perspective of a data scientist, it would be plausible to gain more knowledge on changes and abnormalities in the PCTs by comparing two PCTs in the same plant. This approach is not feasible from the perspective of water professionals as the tanks often have different control settings. However, it might be beneficial to compare the results of the anomaly detection. If an anomaly is detected in both PCTs simultaneously, the anomaly is most likely caused by surrounding factors and not faults in the sensors at the plant.

Like in other fields, such as maritime image recognition [26], large variation in the environment prevents formulation of specific general requirements; however, it is possible to discuss the main factors which need to be considered.

E.5.2 Control Settings

In situ changes in control settings at the WWTPs were largely observed in the data. Some of the changes were detected as anomalies but it was not always the case. In some cases, the changes were a consequence of drift in the NH_4 sensor. This is a practical solution at the plant and solves the present problem; however, it also introduces a bias in the data and makes faulty data normal. Furthermore, it was observed that sometimes, when the conditions at a plant using a PID-controller changed, the control settings were changed. This could, for instance, be due to increased flow. Some of the more extreme cases were visible for a human observer while it is uncertain if changes of less extreme character were present in the data. Change in control settings largely affects the patterns in the data, complicating development of data-driven solutions and in cases where it is found necessary to change the settings, it is essential that the changes are logged.

E.5.3 Logging Strategies

Missing and insufficient logging was a large challenge met in this study. In the cases where a log was available it solely contained information on measurement and calibration, and in some cases, it was not clear which sensors were calibrated due to unprecise documentations. Changes in control settings were never mentioned, despite being essential for the patterns in the data. The lag of logs at the plants is not solely a problem

from a data science perspective, but it also makes it hard for newcomers to understand the plant, as they cannot see what has been done previously.

From a data science perspective, all lab tests, calibrations and change in control settings should be documented in a software system with constrained input parameters selected either from dropdown menus or check boxes, leaving solely numbers for manual entering. However, from the operator's point of view this can easily be considered as unnecessary bureaucracy. Therefore, the logging software should be as simple as possible while still providing sufficient data, and the operators should be included in the design and implementation processes and be able to see a benefit.

E.5.4 Data Quality

Multiple definitions of data quality can be found; however, the key element is that data is of high quality if it is 'fit for use' for the given purpose. Thereby data can have a high quality in one perspective while being of low quality from another perspective. Data need to contain a certain level of completeness, consistency, validity and timeliness, which all depend on the particular purpose [27, 28].

This study shows that the information in the data available was insufficient for comprehensive drift detection in multiple sensors. Furthermore, due to a combination of missing logs and low resolution in the data, four out of seven data sets available for this study were not used.

Insufficient data quality is a problem in multiple other industrial cases. Despite companies collecting data with the purpose of using it, there is a high amount of data, which are collected without being actionable [28].

Prospectively, the authors suggest that data owners at utilities and municipalities consider what they wish to gain from their data and, based on this, select which data to store and what the resolution should be for the data to contain sufficient information. It is possible that other factors not directly connected to the content of the plant such as energy usage and cost at a given time could be relevant factors for benchmarking the performance of the plant. Generally, it is important that the pattern in the data is relevant. Changes in patterns can occur by change in control settings, sensor drift, change in the catchment area etc. From a data quality perspective, the changes in patterns should be minimized, and when they occur, they should be well documented. In case a lot of information needs manual entry, it could be considered to use a well-defined user interface, to reduce faults in the manual documentation and increase the precision of the data. Generally, data should be easily accessible and interpretable [27]. In this connection it is important to ensure coherent naming of parameters, etc. For more information on data quality, please refer to Mahanti [27].

E.5.5 Learning Algorithms

Development of data-driven drift detection in treatment plants is complicated since the control systems are based on feedback loops. Thereby the system is automatically adjusted to the drift, minimizing the changes in faulty data compared to correct data. Furthermore, constant concentration levels in the outlet, where the NH_4 and NO_3 sensors are placed, are considered optimal; however, a constant value further decreases the level of information in the data. As there is a large uncertainty in the composition of the wastewater arriving at the plant, it can be difficult to distinguish between natural variations and sensor drift from a data perspective. A solution to this could be sensors located at the inflow. This would give the possibility of performance evaluation, etc.; however, it would also result in more sensors to maintain.

Due to the costs of sensors, it is often not feasible to implement additional sensors. Therefore, when selecting sensors in WWTPs and deciding which parameters to store, it is important to consider the indirect information in sensors and potential use cases. For instance, Thürlimann et al. [29] suggested a soft sensor using the pH in the inlet and the outlet to detect NH_4 peak load events. Another parameter worth considering in the future is the airflow. The correlation between the airflow and the DO most likely contains usable information of the processes in the plant.

E.5.6 Data-Driven Drift Detection–Is It Worth It?

Due to large variations between plants, it is necessary to acquire a high-quality data set for each plant and subsequently adjust the model to the plant. Acquiring the data set entails that the operators systematically measure and calibrate sensors. Furthermore, the control settings should not be changed and if they need change due to external factors, this should be documented. With such a high-quality data set it is possible to detect faults in the plant [23]. If the catchment area of the plant is changed or if the control system needs updates, for instance due to better algorithms, the data acquisition needs to be remade. This means that the operators need to be systematic in the operation of the plant for several months, or preferably a year, every time a change is made. Lab measurements are easy to perform, and the biggest obstacle is to obtain a culture among the operators where lab measurements are performed instead of ad hoc adjustment of the control settings. A utility that can acquire the needed data set might already have obtained a culture of high-quality sensor maintenance, making data-driven drift detection redundant.

E.5.7 Other Ways of Handling Drift

The above statement yields a need for higher quality in sensor data at wastewater plants. This is especially relevant for the sensors which record data that are used by the plant's control system. Data quality can be obtained by regular monitoring, calibration and

cleaning. Other approaches include self-calibrating sensors and soft sensors. It could also be argued that in some cases, multiple sensors of the same type could be used for drift detection; however, as the sensors would be in the same environment, they would also be affected by the same environmental factors such as fouling or drift after heavy rain or high NH_4 levels. Contemporary, ion-selective sensors are widely used as they are cheap to operate. Another solution could be to use sensors based on gas chromatography for quality control. This sensor uses chemicals and measure once an hour; however, this would be an expensive solution.

A different approach to manage drift could be to include more rules in the control strategy, for instance by stopping aeration if the NO_3 level does not increase or by finding the actual NH_4 level by aerating until the NH_4 level does not decrease more during night-time.

E.5.8 Why Well Considered Data Acquisition from Plants Is Still Necessary

Increased focus on the SDGs emphasizes that the utilities optimize the operations at the plants by reducing energy usage and lowering greenhouse gas emissions while ensuring a high degree of N removal. However, to benchmark performance of experiments performed to optimize the performance, the general performance of the plant needs to be known. Newhart et al. [10] stated that it is essential to a define the problem scope and desired goals when integrating data-driven control at WWTPs. This can be generalized to other tasks involving data-driven solutions.

E.5.9 Can Low Quality Data Still Create Value?

Data quality is a relative concept, and it is related to the purpose of the data [27]. Therefore, the data can be of high quality if used for other purposes. For instance, comparing the available parameters for a given day with average values of the previous days, days with similar flow or similar weekdays can give information to water professionals and help them evaluate the operation of the plant. If available, the energy usage can give information on the effectiveness of the operation of the plant. Furthermore, comparing the average price of the energy used at the plant a given day to the average energy price the same day can give information on how sustainable the energy usage is, as low energy prices are often related to a surplus production of green energy. This is relevant as it can help operators evaluate and optimize the control strategy of the plant and thereby contribute to a more holistic cross sectorial optimization, which is essential to obtain smart cities.

E.6 Conclusions

Sensor drifts are widely present in WWTPs and can result in less efficient operation at the plants. Several approaches for solving this problem have recently been proposed and documented in academia; however, the studies rarely reflect the conditions at real treatment plants and thereby remain as academic projects. The aim of this study was to investigate this gap between academia and practice by applying algorithms suggested in academia on data from real WWTPs. The results showed that obtaining a robust and valid model for fault detection is challenged by several factors such as low data quality, missing logging and in situ changes of control settings. The most often detected anomalies were related to increased flow or increased concentrations, which can be hard to distinguish from sensor drift. It is the author's interpretation that better algorithms and results could be obtained by increased focus on the data quality by including well-considered data management, logging strategies and consistency in the control settings of the WWTP. However, if a utility can obtain such a data set, the problems with drift might already have been solved. Other solutions to handle sensor drift include implementation of improved sensors for quality control, self-calibrating sensors and soft sensors based on informative parameters.

While the data quality might not be sufficient for automatic drift detection, the quality might be sufficient for statistical purposes, which can contribute to information for water professionals and help them evaluate the performance of the plant.

Author Contributions: Conceptualization B.D.H., T.B.M. and D.G.J.; methodology, B.D.H., T.B.M. and D.G.J.; software, B.D.H.; validation, T.B.H., T.B.M. and D.G.J.; formal analysis, B.D.H.; investigation, B.D.H.; resources, B.D.H.; data curation, B.D.H.; writing—original draft preparation, B.D.H.; writing—review and editing, B.D.H., T.B.H., T.B.M., and D.G.J.; visualization, B.D.H.; supervision, T.B.H., T.B.M. and D.G.J.; project administration, B.D.H., and D.G.J.; funding acquisition, B.D.H., T.B.M. and D.G.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovation Fund Denmark.

Acknowledgments: The authors wish to thank Lars Lading from EnviDan A/S for his assistance with accessing data and supervision.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] W. Zhang, N. B. Tooker, and A. V. Mueller, "Enabling wastewater treatment process automation: leveraging innovations in real-time sensing, data analysis, and online controls," vol. 6, no. 11, pp. 2973–2992.

[2] I. Santín, C. Pedret, R. Vilanova, and M. Meneses, "Advanced decision control system for effluent violations removal in wastewater treatment plants," vol. 49, pp. 60–75.

[3] P. A. Stentoft, L. Vezzaro, P. S. Mikkelsen, M. Grum, T. Munk-Nielsen, P. Tychsen, H. Madsen, and R. Halvgaard, "Integrated model predictive control of water resource recovery facilities and sewer systems in a smart grid: example of full-scale implementation in kolding," vol. 81, no. 8, pp. 1766–1777.

[4] L. Corominas, M. Garrido-Baserba, K. Villez, G. Olsson, U. Cortés, and M. Poch, "Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques," vol. 106, pp. 89–103.

[5] O. Samuelsson, G. Olsson, E. Lindblom, A. Björk, and B. Carlsson, "Sensor bias impact on efficient aeration control during diurnal load variations," vol. 83, no. 6, pp. 1335–1346.

[6] O. Samuelsson, A. Björk, J. Zambrano, and B. Carlsson, "Fault signatures and bias progression in dissolved oxygen sensors," vol. 78, no. 5, pp. 1034–1044.

[7] H. Y. Teh, A. W. Kempa-Liehr, and K. I.-K. Wang, "Sensor data quality: a systematic review," vol. 7, no. 1, p. 11.

[8] I. Baklouti, M. Mansouri, A. B. Hamida, H. Nounou, and M. Nounou, "Monitoring of wastewater treatment plants using improved univariate statistical technique," vol. 116, pp. 287–300.

[9] M. Thomann, L. Rieger, S. Frommhold, H. Siegrist, and W. Gujer, "An efficient monitoring concept with control charts for on-line sensors," vol. 46, no. 4, pp. 107–116.

[10] K. B. Newhart, R. W. Holloway, A. S. Hering, and T. Y. Cath, "Data-driven performance analyses of wastewater treatment plants: A review," vol. 157, pp. 498–513.

[11] F. Baggiani and S. Marsili-Libelli, "Real-time fault detection and isolation in biological wastewater treatment plants," vol. 60, no. 11, pp. 2949–2961.

[12] J. Alferes, S. Tik, J. Copp, and P. A. Vanrolleghem, "Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection," vol. 68, no. 5, pp. 1022–1030.

[13] T. Cheng, A. Dairi, F. Harrou, Y. Sun, and T. Leiknes, "Monitoring influent conditions of wastewater treatment plants by nonlinear data-based techniques," vol. 7, pp. 108827–108837.

[14] F. Huang, W. Shen, and Z. Liu, "Applications of sub–period division strategies on the fault diagnosis with MPCA for the biological wastewater treatment process of paper mill," in *2019 Chinese Control Conference (CCC)*, pp. 5138–5143, IEEE.

[15] A. H. Ba-Alawi, P. Vilela, J. Loy-Benitez, S. Heo, and C. Yoo, "Intelligent sensor validation for sustainable influent quality monitoring in wastewater treatment plants using stacked denoising autoencoders," vol. 43, p. 102206.

[16] P. Kazemi, J. Giralt, C. Bengoa, A. Masoumian, and J.-P. Steyer, "Fault detection and diagnosis in water resource recovery facilities using incremental PCA," vol. 82, no. 12, pp. 2711–2724.

[17] P. Kazemi, C. Bengoa, J.-P. Steyer, and J. Giralt, "Data-driven techniques for fault detection in anaerobic digestion process," vol. 146, pp. 905–915.

[18] A.-V. Luca, M. Simon-Várhelyi, N.-B. Mihály, and V.-M. Cristea, "Data driven detection of different dissolved oxygen sensor faults for improving operation of the WWTP control system," vol. 9, no. 9, p. 1633.

[19] B. Mali and S. H. Laskar, "Incipient fault detection of sensors used in wastewater treatment plants based on deep dropout neural network," vol. 2, no. 12, p. 2121.

[20] C. Xu, D. Huang, D. Li, and Y. Liu, "Novel process monitoring approach enhanced by a complex independent component analysis algorithm with applications for wastewater treatment," vol. 60, no. 38, pp. 13914–13926.

[21] M. C. Klanderman, K. B. Newhart, T. Y. Cath, and A. S. Hering, "Fault isolation for a complex decentralized waste water treatment facility," vol. 69, no. 4, pp. 931–951.

[22] B. Mamandipoor, M. Majd, S. Sheikhalishahi, C. Modena, and V. Osmani, "Monitoring and detecting faults in wastewater treatment plants using deep learning," vol. 192, no. 2, p. 148.

[23] F. Cecconi and D. Rosso, "Soft sensing for on-line fault detection of ammonium sensors in water resource recovery facilities," vol. 55, no. 14, pp. 10067–10076.

[24] A. M. Anter, D. Gupta, and O. Castillo, "A novel parameter estimation in dynamic model via fuzzy swarm intelligence and chaos theory for faults in wastewater treatment plant," vol. 24, no. 1, pp. 111–129.

[25] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," vol. 29, no. 2, pp. 93–104.

[26] M. Pedersen, N. Madsen, and T. B. Moeslund, "No machine learning without data: Critical factors to consider when collecting video data in marine environments," vol. 16, no. 3.

[27] R. Mahanti, *Data quality: dimensions, measurement, strategy, management, and governance*. ASQ Quality Press.

[28] A. Scarisbrick-Hauser and C. Rouse, "The whole truth and nothing but the truth? the role of data quality today," vol. 1, no. 3, pp. 161–171.

[29] C. M. Thürlimann, D. J. Dürrenmatt, and K. Villez, "Soft-sensing with qualitative trend analysis for wastewater treatment plant control," vol. 70, pp. 121–133.

Part III

Appendices

Appendix F

Lookup Tables for Assessment of Use Cases

Bolette Dybkjær Hansen

This appendix provides lookup tables used for assessing the use cases presented in Chapter 2.

F.A Lookup tables

Table F.1: Intervals for estimating the customers' willingness to pay in 1000 DKK and corresponding values used in formula.

Willingness to pay	Value used in formula
<25	12.5
25 - 100	62.5
100 - 300	200
300 - 500	400
500 <	500

Table F.2: Intervals for estimating the the number of customers after five years and corresponding values used in formula.

Expected number of customers after 5 years	Value used in formula
<10	5
10 - 25	17.5
25 - 75	62.5
75 - 200	137.5
200 <	200

Table F.3: Intervals for estimating the savings obtained by the solution for EnviDan and the corresponding value used in formula.

Expectations to savings for EnviDan after 5 years	Value used in formula
None	0
Unspecified potential saving	100
<100	50
100 - 300	200
300 - 500	400
500 - 1000	750
1000 <	1000

Table F.4: Where to access the data and values for how hard it is to access.

Data accessibility	Value used in formula
EnviDan, GIS etc.	0.05
Utility	0.25
Local meter	0.4
Need for data collection	1

Table F.5: Number of input parameters and corresponding values inserted in the formula.

Number of input parameters	Value used in formula
1-5	0.05
6-10	0.16
10-20	0.3
20-50	0.7
50 <	1

Table F.6: Number of output parameters and corresponding values inserted in the formula.

Number of output parameters	Value used in formula
1	0.05
2	0.1
3 - 5	0.2
5-10	0.375
10 - 20	0.75
20 <	1

Table F.7: Level of correlation between input and output parameters and corresponding values inserted in the formula.

Is there a direct correlation between input and output?	Value used in formula
Yes, direct	0.1
Yes, in short terms (within a day)	0.1
Yes, in medium terms (<a month)	0.3
Yes, in long terms (1 month<)	1
Yes, but complicated	1
Indirect correlation	1

Table F.8: Type of input parameters and corresponding values inserted in the formula.

Type of input parameters	Value used in formula
Data points, time series etc.	0.1
Images	1

Table F.9: Machine learning type and values inserted in the formula.

Method	Value used in formula
Supervised	0.25
Supervised and/or unsupervised	0.25
Unsupervised	0.4
Reinforcement learning	1

Table F.10: Required adjustment per sale and values inserted in the formula.

Required adjustment per sale	Value used in formula
No adjustment	0
Simple adjustment	0.4
Multiple adjustments	1

Table F.11: Minimum requirements for precision of the solution and values inserted in the formula. If a skewed dataset is used, the true positive and false negative rates can be considered instead of the precision.

Minimum requirement for precision	Value used in formula
>0.9	1
0.7-0.9	0.25

Table F.12: Overview of the α used in the formulas.

Alpha	Description	Value
α_1	Tuning of $time_{dev}$	0.741
α_2	Tuning of $time_{adj}$	0.4
α_3	Tuning of $data_{access}$	0.2
α_4	Tuning of *complexity* for time calculation	1.205
α_5	Tuning of n_{in}	0.225
α_6	Tuning of n_{out}	0.225
α_7	Tuning of $corr_{in_out}$ for time calculation	0.225
α_8	Tuning of $type_{in}$	0.225
α_9	Tuning of *method*	0.225
α_{10}	Tuning of $corr_{in_out}$ for risk calculation	0.4
α_{11}	Tuning of *complexity* for risk calculation	0.904

Appendix G

Supplementary to paper C, Correlation between topographically connected pipes

Bolette Dybkjær Hansen

This appendix provides unpublished supplementary material to Paper C

G.A Correlation between topographically connected pipes

As described in Paper C, a selective survival bias is present in the inspected data. This has motivated an investigation of the correlation between different defect types in topographically connected pipes. To investigate this, the different defect types observed in the CCTV inspections were binarized according to whether a defect was present or not, not encountering the severity of the defect. For each defect type present in one pipe, the probability of any defect type being present in the pipe, the upstream pipe, and the second upstream pipe was found. The results can be seen in Figure G.1.

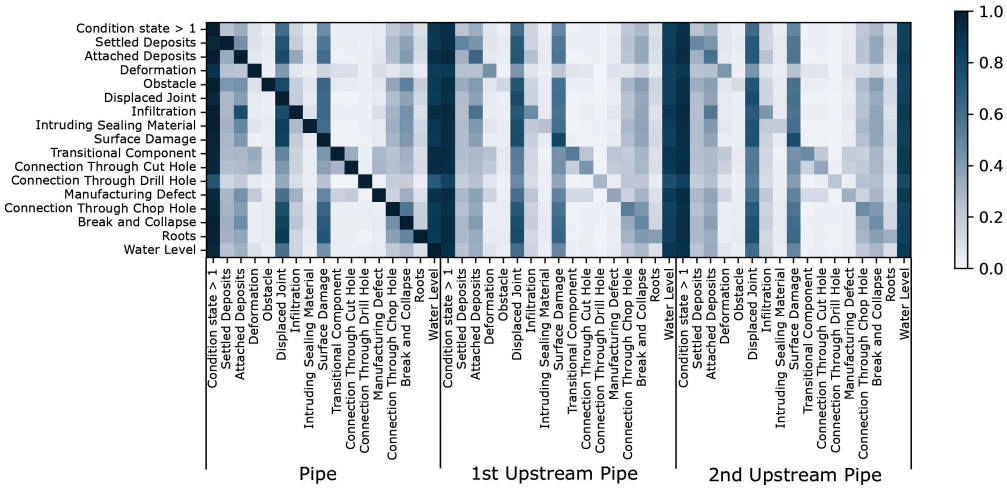


Fig. G.1: The defect types which can be observed in a given pipe are listed to the left. In the bottom the defect types for the given pipe, an upstream pipe, and the second upstream pipe are listed. The colors indicate the probability of one of the defects listed in the bottom being present given the defect to the left is present..

In Figure G.1 it can be seen, that if any defect is present in a given pipe there most likely also is a displaced joint, a surface damage, and water in the pipe itself, in the upstream pipe, and in the second upstream pipe. Likewise there is an increased probability of settled deposits, attached deposits, and connection through chop holes in the pipe, and the up stream pipes. However, it is also worth mentioning that the occurrence of a displaced joint, surface damage or water in the pipe, does not entail a large increase in other fault types. This is likely because these three fault types are present in a large percentage of the pipes, which is not the case for many of the other fault types. When solely considering the pipes categorized as being in condition state

four or having defect of degree three or four, the same patterns can be observed as shown as shown in Figure G.2.

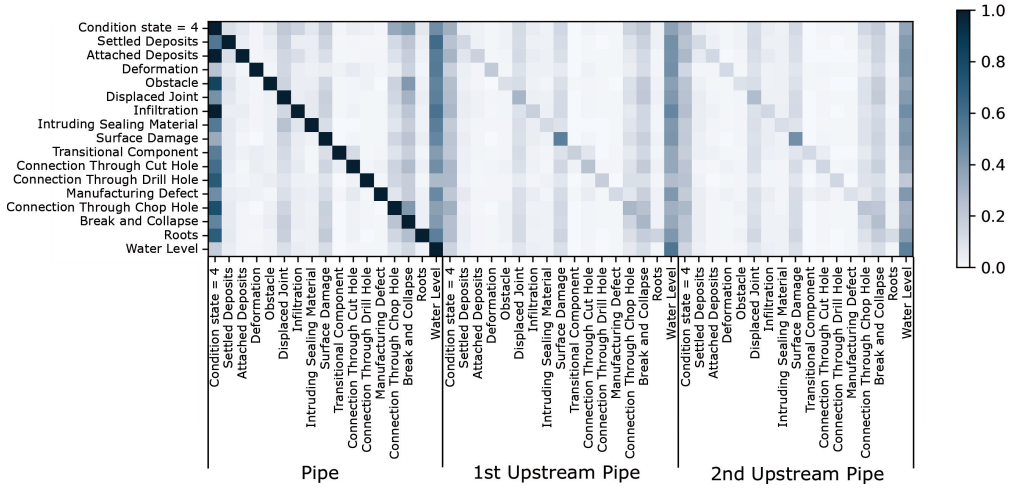


Fig. G.2: Correlation between pipes and upstream pipes with defect types of severity three or four or condition state four. The defect types which can be observed in a given pipe are listed to the left. In the bottom the defect types for the given pipe, an upstream pipe and the second upstream pipe are listed. The colors indicate the probability of one of the defects listed in the bottom being present given the defect to the left is present..

As seen in Figure G.2, similar patterns can be seen when considering the pipes with severe defects as when considering all defects, however, the correlations are generally weaker.

To see how input from the surrounding pipes can contribute the the performance of the deterioration models, models were trained with input from the upstream pipes. As it is clear that the distribution of the data set is important for the performance, the models were trained for both cases. Furthermore, when upstream pipes are included as prediction parameter, the size of the dataset decreases, as fewer pipes have a second or third upstream pipe which have been inspected. Therefore, to ensure a sufficiently large dataset it was it was decided to include pipes with missing data. For this reason it was decided to test XGBoost, a tree based method, which can handle missing data. Consequently, four tests were carried out: one using the full dataset including data points with missing values, one using using an approximately equally distributed dataset including data points with missing values, one using the full dataset except for data points with missing values, one using using an approximately equally distributed dataset except for data points with missing values. The four combinations can be seen in Figure G.3.

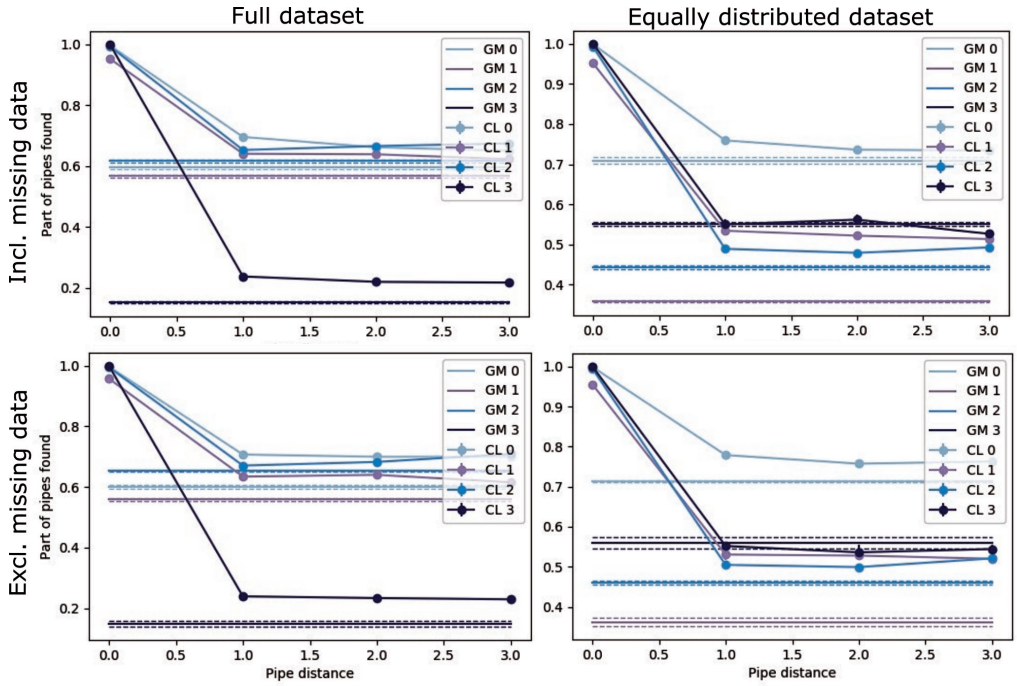


Fig. G.3: Performance of an XGBoost deterioation model when applied to the full dataset incl and excl missing data and an equally distributed dataset incl and excel missing data. The figure show the performance of the general models (GM), which does not include observations from CCTV inspections, and the models which include observations from CCTV inspections (CL) from the pipe itself, 1st, 2nd, and 3th upstream pipes respectively. Please notice that the 2nd axis is different for the models using the full dataset and the equally distributed dataset.

From Figure G.3 it can be seen that input from the upstream pipes can help improving the performance, of the model. It is not important if it is the third of the first upstream pipe which is included, and the performance is only slightly affected by the whether data points with missing data is included or not. When training on an equally distributed dataset the performance change, however, which model is the best depends on the purpose.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-927-1

AALBORG UNIVERSITY PRESS