Master Thesis
TVVR25/5019

Evaluating flow cytometry for investigating slow sand filter function in drinking water treatment

Exploratory data analysis and implementation of machine learning on pre-existing datasets from three Swedish drinking water treatment plants

Agnes Westberg

Division of Water Resources Engineering
Department of Building and Environmental Technology
Lund University

Evaluating flow cytometry for investigating slow sand filter function in drinking water treatment

Exploratory data analysis and implementation of machine learning on pre-existing datasets from three Swedish drinking water treatment plants

By: Agnes Westberg

Master Thesis

Division of Water Resources Engineering
Department of Building & Environmental Technology
Lund University
Box 118
221 00 Lund, Sweden

Water Resources Engineering
TVVR-25/5019
ISSN 1101-9824

Lund 2026
www.tvrl.lth.se

Swedish title:	Utvärdering av flödescytometri för kvalitetssäkrande av långsamfilter i dricksvattenproduktion
English title:	Evaluating flow cytometry for investigating slow sand filter function in drinking water treatment
Author:	Agnes Westberg
Supervisor:	Catherine J Paul, Isabella K Erb, Iria Feijóo Rey, Caroline Schleich, Sandy Chan
Examiner:	Kenneth M Persson
Language	English
Year:	2026
Keywords:	<Flow cytometry; Machine learning; Slow sand filtration; fluorescence fingerprinting; drinking water production>

Abstract

Quality assurance is crucial for producers of drinking water, and several steps are taken to ensure a safe and aesthetically acceptable drinking water. Monitoring the microbiological quality of drinking water, both finished and within the drinking water treatment plants, is done through cultivation-based techniques. These techniques can take up several days before results can be seen, and quicker results would allow drinking water producers to act on potential quality issues before the water reaches the distribution system.

Flow cytometry offers quick and high-throughput results, measuring total and intact cell counts along with other parameters depending on which reagents are used. It also allows for mapping the cells detected to two-dimensional histograms based on, for example, fluorescence. This project studies the possibility of relating traditional cultivation-based techniques to results from flow cytometric measurements on water effluent of slow sand filters, first through exploratory data analysis and then with machine learning. Machine learning was used to find patterns in the fingerprints to determine if heterotrophic plate counts or coliform content were above or below a set threshold.

Data was provided by three different drinking water treatment plants: Norsborg's drinking water treatment plant (Stockholm vatten och avfall), Ringsjöverket (Sydvatten) and Berggården (Tekniska verken). Correlations were found between flow cytometric data and heterotrophic plate counts at Norsborg's drinking water treatment plant. The machine learning models also showed promise at determining if the heterotrophic plate count would be above or below 20 colony forming units/mL by analysing the fingerprints. Correlations were also found between total cell counts and head loss in the filters at Norsborg's drinking water treatment plant and Ringsjöverket. No relationships between flow cytometric data and cultivation-based techniques were observed at Berggården or Ringsjöverket, and the machine learning models were less effective at these sites.

Acknowledgements

Many people and organizations have played into this project, and each of them have my sincerest gratitude. First, I would like to thank my supervisors, Catherine Paul, Isabella K. Erb, Iria Feijóo Rey and Caroline Schleich for their support, insights and feedback throughout my thesis work. Many thanks are reserved for Sandy Chan, lab supervisor at Sydvatten, who trusted me with my first opportunity to work with slow sand filters and flow cytometry, without whom I would most likely never have ended up with this project at all. I would also like to thank Ludvig Klotz, a colleague in the water group at TMB, and most importantly a great friend, who has gone above and beyond to help me and cheer me up when the work felt too hard.

This project would not have been possible without the cooperation from the drinking water producers Stockholm vatten och avfall, Sydvatten and Tekniska verken. Therefore, I want to thank Anders Lindström (Sydvatten), Lars-Erik Hägglund (SVOA), Anders Svensson (SVOA) and Felicia Larsson (Tekniska verken) for providing me with data and for answering every question I had about the datasets and your treatment processes. Finally, I want to thank everyone at these organisations who have had part in producing the data, sampling and running analyses, of whom there are too many to mention here.

Popular summary

Flödescytometri som verktyg i dricksvattenproduktionens kvalitetssäkrande arbete

Dricksvattenproducenter har flera metoder för att undersöka och kontrollera vattnets kvalitet. Den mikrobiella kvaliteten kan ta allt ifrån 18h till sju dygn att fastställa. Flödescytometri är en metod som har föreslagits som ett potentiellt alternativ eller tillägg till dessa metoder, och som erbjuder korta analystider och hög potential för parallella analyser. I det här projektet har vi undersökt hur flödescytometriska mätningar kan jämföras med mer traditionella metoder.

Dricksvatten måste vara rent och säkert för att kunna distribueras till konsumenter. Detta kontrollerar dricksvattenproducenter genom olika kemiska och mikrobiella mätningar. Resultat för de kemiska parametrarna kan oftast inhämtas inom en arbetsdag, medan de mikrobiella sällan kan kontrolleras förrän tidigast dagen efter provtagning. Eftersom dricksvatten inte kan återkallas efter att det nått distributionsnätet, är snabba resultat av stor vikt. Flödescytometri kan ge resultat för flera prover och mikrobiologiska parametrar mycket snabbt, ofta under en timma. Att kunna koppla flödescytometriska resultat till dricksvattnets kvalitet skulle vara värdefullt för dricksvattenproducenter, då det skulle göra kvalitetssäkring och felsökning mycket snabbare.

I det här projektet har vi undersökt om, och i så fall hur, de flödescytometriska parametrarna totalantal celler, antalet intakta celler och fluorescens-histogram kan kopplas till de mer traditionella analysmetoderna odlingsbara mikroorganismer (3-dygns) och antalet koliforma bakterier. Datan som vi undersökt kom från tre olika svenska vattenverk, ett i Stockholm, ett i Linköping och ett i Stehag. I projektet såg vi stor variation i vilka flödescytometriska parametrar som kunde kopplas till de mer traditionella metoderna, både i den mer utforskande delen av projektet, och i maskininlärningen.

Slutsatserna som vi drog av detta var att mer ingående studier behövs för att fastställa huruvida totalantal eller intakta celler korrelerar med odlingsbara mikroorganismer eller antalet koliforma bakterier. Resultaten från den förutsägande maskininlärningen var också diffusa, och mer studier behövs för att kunna bestämma tydligare samband. Vi drog även slutsatsen att flödescytometri kan vara ett kraftfullt verktyg i sig, från vilket mycket information kan hämtas, och att det kan vara begränsande att se det från samma perspektiv som de mer traditionella metoderna.

Table of Contents

List of abbreviations

FCM	Flow CytoMetry
FCS	Flow Cytometry Standard
HNA	High Nucleic Acid
HPC	Heterotrophic Plate Count
ICC	Intact Cell Count
LDA	Linear Discriminant Analysis
LNA	Low Nucleic Acid
MPN	Most Probable Number
PCA	Principal Component Analysis
PI	Propidium Iodide
SHAP	SHapley Additive exPlanations
SSF	Slow Sand Filtration / Slow Sand Filter
TCC	Total Cell Count

1. Introduction

1.1 Background

1.1.1 Slow Sand Filtration

Safe and clean drinking water is an important factor in the health and well-being of a population, so much so that it is classified as a human right [1]. Excessive microbial growth in the water system can lead to health hazards, increased strain and corrosion of the distribution system, as well as consumer dissatisfaction with the aesthetic qualities of the drinking water. [2], [3] To achieve a safe drinking water several different purification techniques can be applied, one of which is slow sand filtration (SSF). By making use of a bacterial community grown in a sand bed in tandem with the mechanical filtering of the sand, microbial activity is reduced by several orders of magnitude. When in use, biofilm consisting of a diverse community of microbiota forms at the top of the filter, the so called *schmutzdecke*, as well as further down into the sand bed. The schmutzdecke needs to be removed at relatively regular intervals, when the head loss over the filter gets too large. [4] The Swedish Food Agency (Livsmedelsverket) classifies slow sand filtration (SSF) as a microbiological barrier, making it crucial to monitor its function to ensure the quality of the water produced is adequate. [5]

1.1.2 Assessing Drinking Water Quality According to the Swedish Standard

To assess the microbiological quality of drinking water, indicator organisms can be used. This means enumeration or simply detection of microorganisms that are not necessarily pathogenic themselves, but their presence can indicate that other, more problematic organisms also could have survived the cleaning process. [5], [6], [7] Two categories of indicator organisms are coliform bacteria, and culturable microorganisms [8].

Coliform bacteria is a group partially made up of microorganisms that can be found in mammalian faeces, such as *Escherichia coli*. In an individual with and ongoing gastrointestinal infection, the number of coliform bacteria will be much higher than that of the microorganism causing the infection. Detection of coliform bacteria can therefore indicate a faecal contamination. These factors, as well as the relative ease with which many coliform bacteria can be cultivated, has led to the total coliform concentration being widely used as indicator organisms in water quality monitoring. [9] However, it has been questioned if the total coliform count is a meaningful figure in the evaluation of water quality, as coliforms can also be found growing in soil and other natural environments [10], [11]. Still, many countries use total coliform counts as a standard in their work to ensure good drinking water quality. Elevated counts are generally thought to be an indicator of outside contamination and to, and lacking

treatment of, the water. This is the case both if the contamination is faecal in nature or caused by infiltration from the surrounding soil. [11]

The current Swedish standard method for determination of total coliforms is Colilert®-18, developed by IDEXX [12]. It consists of a substrate to be added to 100 mL of water sample, and a plate with 97 wells of two different sizes. The plate is laminated shut using a sealer machine, and the sample is distributed across the wells. This method relies on the fact that coliform bacteria express the enzyme β-D-Galactosidase. This enzyme converts the colourless ortho-nitrophenol galactoside into the yellow ortho-nitrophenol, allowing for visual distinction of positive and negative wells, where the wells positive for coliform bacteria turn yellow, and the negative wells remain colourless. The number of positive large and small wells are then used to find the most probable number (MPN) in a table provided by the manufacturer. The incubation time for Colilert®-18 is 18h-22h [13], [14], meaning that same-day results are not possible for a laboratory staffed for 8h per day.

In the Swedish standard for drinking water, culturable microorganisms are defined as those microorganisms that can form colonies in a pour plate of tryptone yeast extract agar when incubated at 22 ± 2°C over three days[8], [15]. This method of enumeration is referred to as heterotrophic plate count (HPC), as the microorganisms grown on the plate are heterotrophic (requiring an organic source of carbon). The enumeration of culturable microorganisms is used to monitor the function of disinfecting steps in the water treatment process.[16] Only a portion of the bacteria present in a water sample can be cultivated in this way, and HPC may therefore mostly be interesting to monitor trends in the water supply [16], [17]. The Swedish Food Agency standards give no maximum allowed concentration of culturable microorganisms and instead say that there should be no abnormal changes in HPC for the water to be considered safe. There are guidelines suggesting that HPC should be no higher than 1000 CFU/mL (colony forming units/mL) for drinking water to be considered completely safe. There is no guide for water to be considered unsafe to drink. [8], [18]

1.1.3 Flow Cytometry

Flow cytometry (FCM) has become more prevalent in a drinking water context over the last few years. It works by adjusting the pressure of the sample solution and a carrying fluid (called sheath fluid) in such a manner that cells in the sample line up in single file [19]. A laser is then pointed at the line of cells, and the way it scatters is detected. One detector is set up to record the forward scatter (FSC) and another the side scatter (SSC). FSC can be used to measure the size of a cell: a larger cell will cause a higher forward scatter. SSC gives a measure of how complex the cell is. [20]

Flow cytometers are also equipped with sensors to pick up on the fluorescence emitted when cells are hit with the laser. If fluorescent tags or dyes are used, it is therefore possible to classify and enumerate cells with certain characteristics depending on the wavelength emitted. These tags can be fluorophores bound to antibodies, or simply

compounds that bind to nucleic acids, among others. [19], [20], [21] FCM is a powerful tool when it comes to determining total cell counts in samples. More traditional analyses require cells to be cultivated before enumeration, meaning that they are restricted by what microorganisms can proliferate on the media at hand. This can lead to the total cell count in a sample being underestimated.[16] Since no cultivation is required for flow cytometry [21], this issue is mitigated and a more accurate total cell count is provided.

Two often used dyes are SYBR Green I (SG-I) and propidium iodide (PI). SG-I is membrane permeant and can therefore be used to determine the total cell count in a sample. PI, however, is membrane impermeant and when used with SG-I, the intact cells can be enumerated. [22], [23], [24], [25] While the green fluorescing SG-I is still present and binds to the DNA, the wavelengths output by SG-I is re-absorbed by the PI, quenching the green signal and magnifying the red [25], [26], [27].

1.1.4 Machine learning

Machine learning is a useful tool for analysing large datasets [28], like those that can be generated using FCM to continuously monitor water quality. The models applied in this study were logistic regression and random forest. These both fall under supervised learning, where the correct classification is known, and can be compared to the classifications made by the algorithms [29], [30]. The models use the information provided to them to find patterns to classify the samples by [28]. In this case, the flow cytometric fluorescence fingerprints will be used for the input data.

1.1.5 Norsborg's DWTP, Ringsjöverket and Berggården

Three drinking water treatment plants from different parts of Sweden provided data for this project: Norsborg's drinking water treatment plant, run by Stockholm Vatten och Avfall (SVOA); Ringsjöverket, run by Sydvatten; Berggården, run by Tekniska verken. The three of them are located throughout southern Sweden, with Ringsjöverket in Scania, Berggården in Östergötland and Norsborg in Södermanland. The approximate locations of the waterworks are shown in Figure 1.1, where the spread from north to south can be seen.

Figure 1.1: Map showing the approximate locations of the water treatment plants from which data was provided. (Map from Lantmäteriet)

All three waterworks have surface water as their source. Mälaren in an oligotrophic lake that is high in alkalinity [31], from which Norsborg's DWTP source their water. After an initial separation of larger solids, flocculation chemicals are added to the water which then moves on to a flocculation chamber. The water passes through rapid sand filters, is pH-adjusted and then enters their 38 slow sand filters. Before the water moves on to the distribution system it is disinfected with UV and chlorinated using monochloramine.

Ringsjöverket gets its water from the lake Bolmen, oligotrophic and low in alkalinity [31]. After leaving the lake the water travels through an 80km long tunnel, before reaching the plant in Stehag. The process then goes as follows: flocculation, sedimentation, rapid sand filtration, slow sand filtration, UV-disinfection and lastly chlorination with sodium hypochlorite. There are 20 slow sand filters at Ringsjöverket.

Berggården sources its water from a stream leading from the lake Vättern, which is a lake with very low humic content and low alkalinity. This DWTP does not have a flocculation and sedimentation step. After initial removal of larger solids, the water goes through rapid sand filtration, then slow sand filtration, UV-disinfection and chlorination. A final pH-adjustment is done, and then the water enters the distribution system. Berggården has eight slow sand filters, two of which are covered.

1.2 Methodology

The paper "A data-driven early warning system for *Escherichia coli* in water based on microbial community analysis using flow cytometry 2D histograms" by Erb, et.al. shows a method of applying machine learning to predict the outcome of *Escherichia coli* enumeration in environmental bathing waters using flow cytometry data. No specific labelling of *E. coli* was used, and the staining used followed a standard procedure using SYBR Green I and propidium iodide (PI). In this method, quantification of *E. coli* in the environmental water samples was done using Colilert®-18. [32] This can also be used to quantify coliform bacteria in a drinking water sample [13], which is one of the parameters this study aims to investigate.

Since no data collection took place during the project itself, the following analyses was partially determined by what data the drinking water producers have collected and can provide to external researchers. When data had been acquired, exploratory

data analysis was performed to investigate what connections were readily noticeable in two-dimensional representations.

The method used by Erb, et.al. [32] was applied to the data provided by the drinking water production companies, adapting it to take into account the specific parameters in this study's dataset.

1.3 Objectives

This project aims to find possible connections between flow cytometric data and regulatory parameters such as 3-day HPC and coliform bacteria, as well as process parameters such as temperature and head loss. The ultimate goal is to provide a solid foundation for future studies, and for drinking water producers wanting to investigate the use of flow cytometry in their process.

Little is known about the differences in flow cytometric data between slow sand filters at different drinking water treatment plants with different source waters. This project aims to shed some light on this issue by investigating how the relationships between flow cytometric data and other parameters differ between waterworks.

In this project, the machine learning method used by Erb, et.al. (2025) will be applied to data (flow cytometric and coliform MPN) collected on water that has passed through slow sand filters in three different drinking water treatment plants, to investigate the predictability of coliform MPN and heterotrophic plate counts.

2. Materials and Methods

2.1 Data collection

No data was collected in this project and was instead provided by the cooperating water production companies Stockholm Vatten och Avlopp (SVOA), Sydvatten, and Tekniska Verken. SVOA provided data collected throughout the years 2024 and 2025 (January 2024 – September 2025), including flow cytometric results, 3-day heterotrophic plate counts (referred to as HPC or 3-day HPC in this report), coliform and E. coli counts, water temperature and head loss. The same data was provided by Tekniska verken, but from the period January 2025 – October 2025. Sydvatten provided flow cytometric data, coliform counts and head loss. The data from Sydvatten was collected only during the summers (June – August) and had data from 2022 – 2025. All flow cytometric data was collected by analysing 50µL of sample mixed with either SYBR Green or SYBR Green/PI.

Among the data provided by the drinking water producers were FCS-files (Flow Cytometric Standard-files), containing the necessary information for fingerprinting. Some samples had been run in duplicates, and for the machine learning, only one of the duplicates were used.

2.2 Data exploration and Preprocessing

All the data provided were entered into spreadsheets, one for each treatment plant. These served as the platform for exploratory data analysis, as well as the metadata for the machine learning models. To investigate possible relationships between flow cytometric parameters and other microbiological or process parameters, FCM results were plotted against other data such as sampling date, coliforms, and HPCs.

To pair the file names of the FCS-files to their corresponding results in the spreadsheet, the names of all files in the directory housing all FCS-files for a given DWTP were extracted using a python script. The name of one of the replicates of a sample was appended to CSV-file (Comma Separated Variables), which could then be copied into the spreadsheet. To ensure that the right file name ended up with the right analysis results, both the files and the data was first sorted by date. To extract only one replicate, slightly different criteria were needed for the different DWTP, as the naming conventions were different. This was done by learning the naming conventions and implementing inclusion criteria for the file names to be selected.

2.3 Machine learning

Machine learning was used to find possible patterns in the fluorescence fingerprints of the samples. First, a grid was constructed. The size of this grid was determined via trial and error, by testing one size and then seeing how well this separated the different categories, which will be discussed later. A data frame was constructed, with each row corresponding to an FCS-file in a given directory, and each column corresponding to a pixel in the grid. Every pixel in the grid corresponded to a part of the FCM fingerprint, and the pixel's value was decided by the number of events recorded in this part of the fingerprint. A simplified version of this is shown in Figure 2.1. This was done by using the python library FlowCytometryTools to quantify the event count within each pixel and entering the pixel values into the grid data frame.

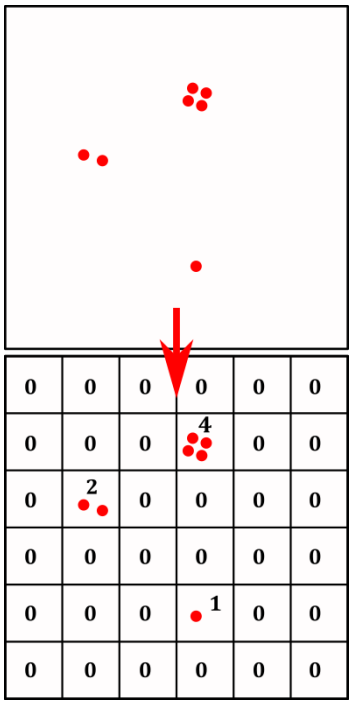


Figure 2.1: Simplified version of a six-by-six grid partitioning a "fingerprint" consisting of seven events.

The meta data for each dataset, the spreadsheets, were then loaded in and using these, a new data frame was constructed. The two data frames were merged by matching the file names. A threshold was set to determine which samples were to be designated positive and negative. This threshold was set, again, via trial and error by studying the separation in the PCA and LDA, and by ensuring that no category got less than 25 samples. The dataset was then downsampled, creating 500 datasets with the same number of positive and negative samples. The number of samples to include of each category was set as the number in the minority group, which in this case was always the positive samples. This meant that the positive samples were the same in all 500 datasets while the negatives were randomised. The datasets were then split into 80% training data and 20% test data.

Only the pixel data was used in the machine learning, meaning that these were the only columns in the data frame that were provided to the models. The column denoting which category a sample belonged to was set as the target to be predicted. The machine learning models logistic regression and random forest were then applied to the 500 downsampled datasets. Both were done using the SciKit learn library in python. The portion of a dataset designated as training data was used by the model to find possible patterns, and the test portion was used to see if the patterns found were able to predict the category of the sample. A mean accuracy and a standard deviation of the accuracies were calculated from the accuracies output by the model. Confusion matrices were constructed using SciKit learn.

2.3.1 Python packages used

- Numpy 1.26.4
- Pandas 2.2.2
- Matplotlib 3.10.8
- FlowCytometryTools 0.5.1
- Scikit-Learn 1.4.2
- Seaborn 0.13.2
- SHAP 0.48.0

3. Results

3.1 SVOA

3.1.1 Exploratory Data Analysis

As a first approach, the total cell counts for every sample was investigated. This can be seen Figure 3.1, where every sample's total cell count (TCC) has been plotted to its corresponding filter. This was done to determine if any filters in particular showed spikes in total cell counts throughout the sampling.

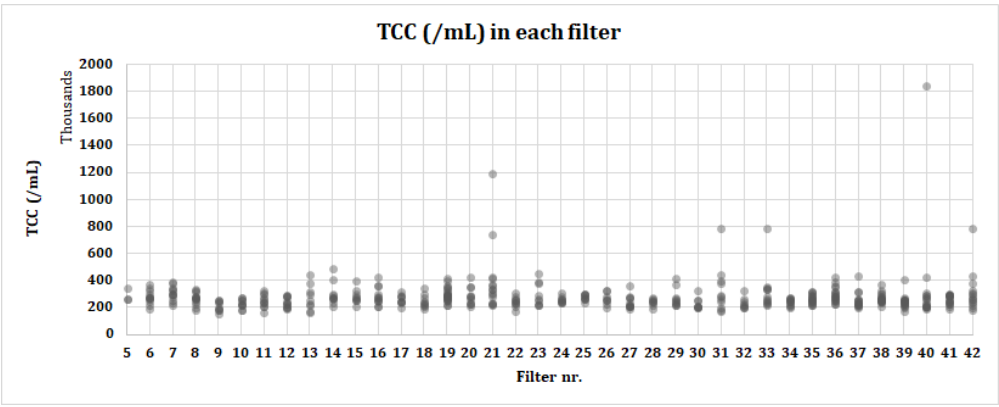


Figure 3.1: Total cell count (TCC) plotted for each individual filter at Norsborg DWTP. Samples are from different dates. Filters 21, 31, 33, 40 and 42 all show some counts that are clearly higher than those for the other filters. The lowest counts start at a little higher than 100 000 cells/mL, and the highest is almost 2 000 000 cells/mL.

Figure 3.2 shows all measured HNA contents (%HNA) of samples for a given filter. There are not as many clear spikes in any filters except for filter 40. Overall, the %HNA seems more varied than the TCC. Since %HNA is describing the fraction of cells that are more complex, having higher nucleic acid content, the variation of

this value could indicate that the composition of the population fluctuates a lot in each filter.

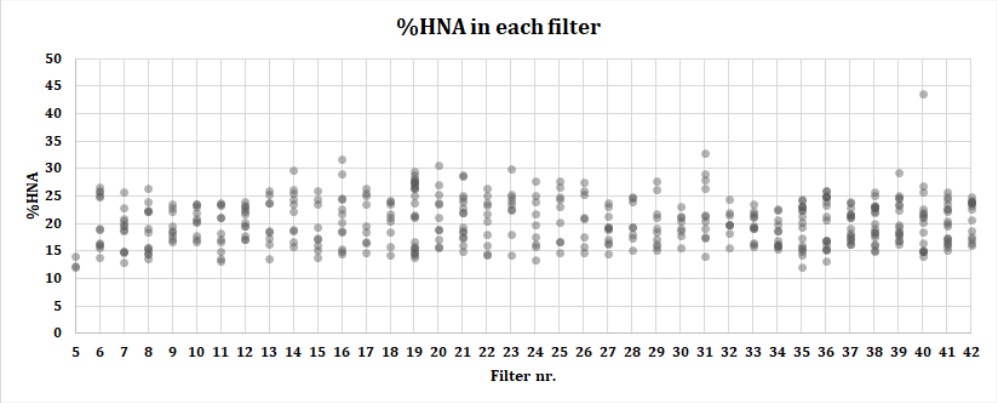


Figure 3.2: High nucleic acid content (%HNA) for each of the filters at Norsborg DWTP. The differences between filters are not as drastic as those for TCC. Filter 40 still shows a peak.

Figure 3.1 shows clear spikes for filters 21, 31, 33, 40 and 42. These all have one or two TCC measurements that seem to stand out significantly from the rest. Filter 40 also has a spike in %HNA at some point during the sampling. To further investigate these high cell counts, the analysis results for filters 21, 31, 33, 40 and 42 were looked at individually, with TCC and %HNA being plotted together with the heterotrophic plate counts (if there is such a result for the given filter on the given day) and the head loss, against the dates on which the samples were taken. This is done to find potential correlations between the TCC, %HNA, HPC and head loss. For SSF 21, these more in-depth results can be seen in Figure 3.3. It seems that a high HPC corresponds to an increase in TCC (3.C A), and that when head loss is low, the TCC is high, at least when the head loss first decreases. The corresponding data for SFF 31, 33, 40 and 42 are shown in “Appendix 1: In-depth plots for filters at Norsborg’s DWTP”.

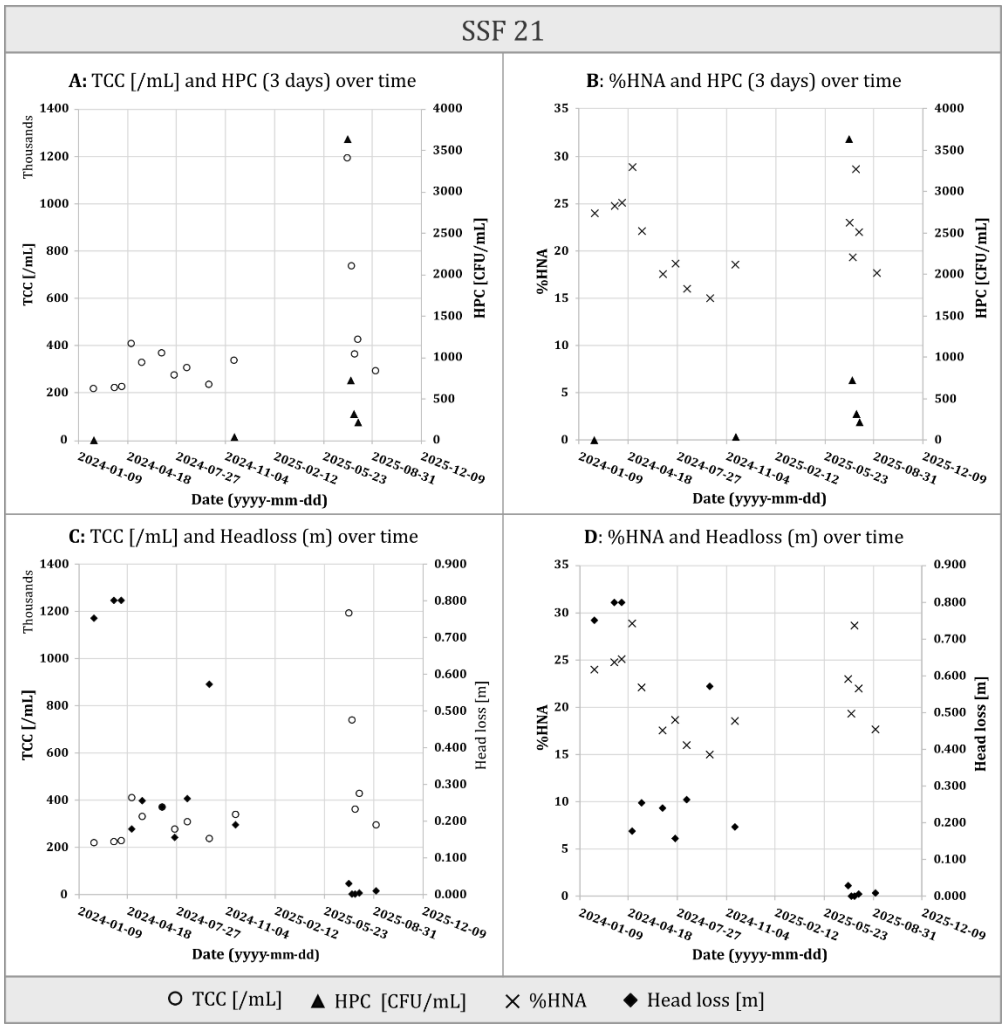


Figure 3.3: In-depth view of different microbiological and process data from SSF 21 at Norsborg DWTP. [A] TCC per mL and HPC (CFU per mL) plotted against the sampling date. The highest TCC and HPC fall on the same day. [B] %HNA and HPC (CFU per mL) plotted against the sampling date. [C] TCC per mL and head loss (m) plotted against the sampling date. The highest head losses fall on the same days as the lowest TCC. The highest TCC is acquired on a day with low head loss. The TCC then decreases, while the head loss stays low. [D] %HNA and head loss (m) plotted against the sampling date.

To further investigate the relationship between head loss and the total cell counts, all TCC measurements were plotted against their corresponding head losses (Figure 3.4). This is of interest since the head loss is a measure of how clogged the filter is, or how developed the schmutzdecke has become since the last scraping [4]. The resulting graph shows the highest TCC:s occurring when the head loss is relatively low. There is also a large variation in TCC on the low end of the head loss. The relationship between %HNA and head loss was also investigated and seemed to be insignificant (not shown).

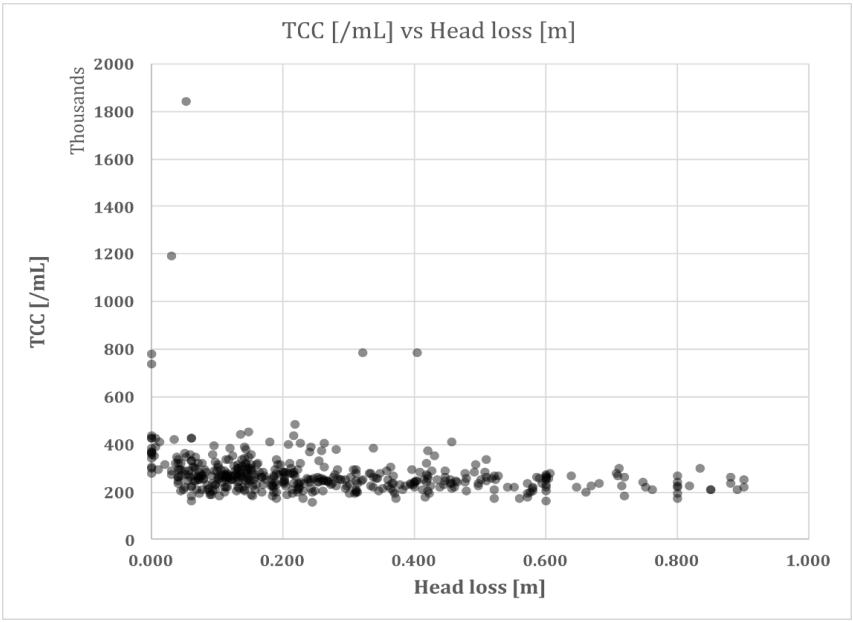


Figure 3.4: TCC per mL for all filters and all samplings plotted to head loss. The higher cell counts are mostly found in the lower half of the head loss values.

Some of the results seen in Figure 3.3 and Appendix 1 suggest that TCC and HPC could have a positive correlation. The data here is more lacking than that for head loss, as HPC was only measured intermittently, and not continuously as was the case for head loss. Only 67 of the samples had corresponding HPC data, and these have been plotted in Figure 3.5. It seems that higher HPC correlates with higher TCCs.

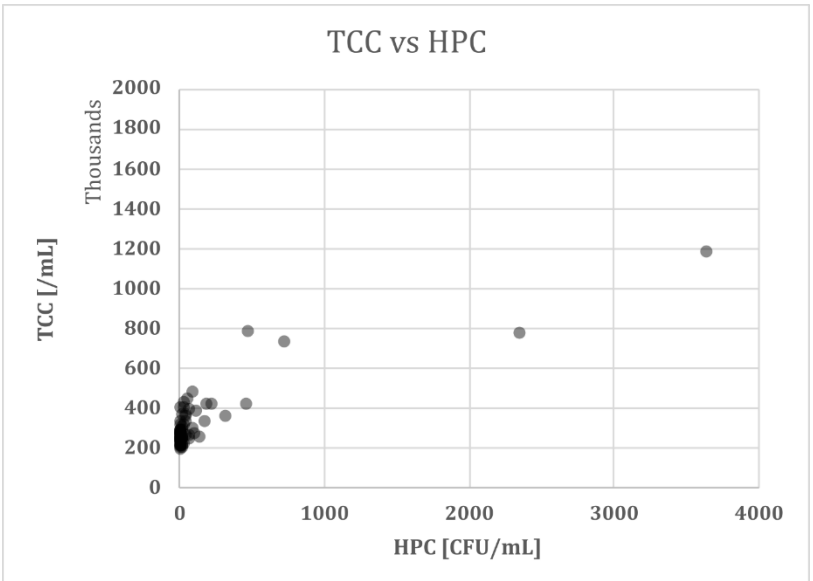


Figure 3.5: TCC plotted to HPC for all data points where flow cytometry and HPC was done on the same day. Higher TCC seems to correlate to higher HPC.

%HNA did not seem to correlate with HPC, as seen in Figure 3.6. This could be an indication that perhaps the population variations do not correlate with the HPC increasing or decreasing.

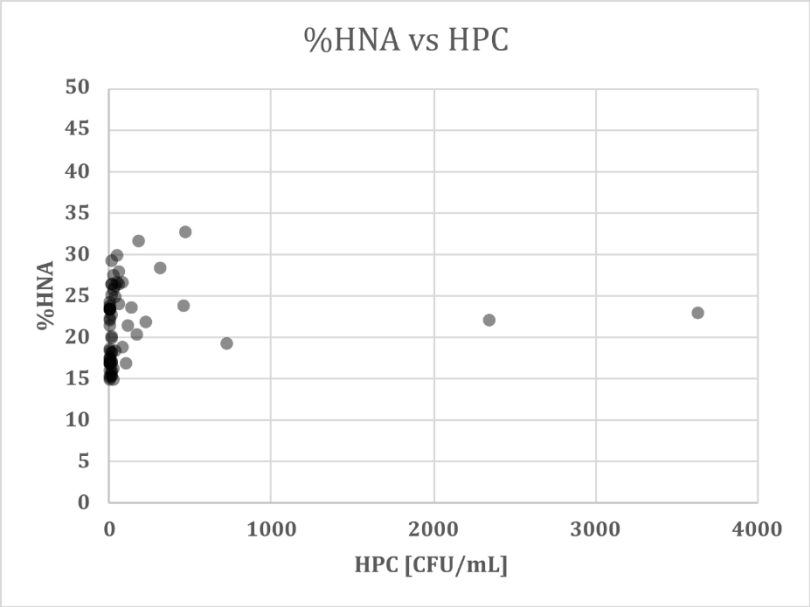


Figure 3.6: %HNA versus HPC. No clear correlation can be seen.

3.1.2 Machine learning

Machine learning was applied to the data to examine the predictability of HPC using the two-dimensional histograms (fingerprints) from flow cytometry. To adapt the data, the fingerprints were divided into a ten-by-ten grid, as seen in Figure 3.7.

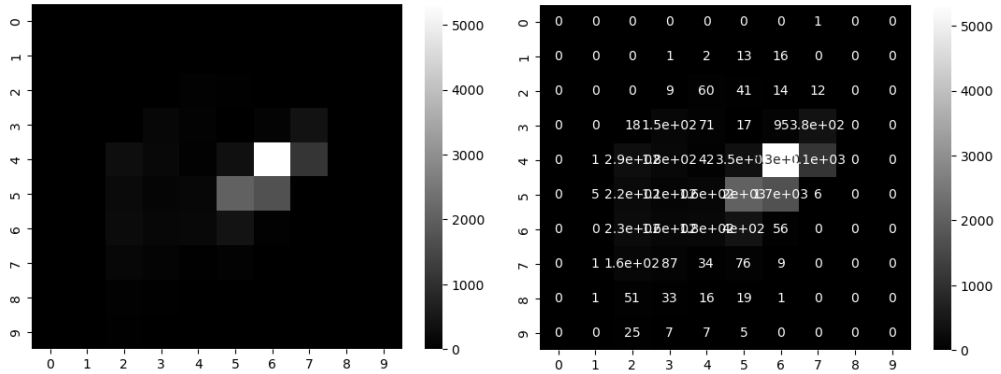


Figure 3.7: Sample from the DWTP divided into 100 squares using a 10x10 grid. Lighter shade indicates higher event count in the square, as shown by the scale to the right of the graph. Shown both without event counts (left) and with event counts (right).

To determine whether a sample was to be considered positive, a limit of >20 CFU/mL was chosen. This limit was chosen as to leave at least 25 positive samples (in the study by Erb, et.al. an accuracy of 80% was reached using 27 samples) and was tested to find a limit where the separation between positive and negative samples was the greatest in the principal component analysis (PCA) and linear discriminant analysis (LDA). 26 samples were positive using the limit >20 CFU/mL. The number of pixels in the grid was also decided through an iterative process where the individual parameters were adjusted one by one, to find the settings which yielded the biggest separation between positive and negative samples (Figure 3.8 and Figure 3.9).

As a first step before machine learning was applied to the data, a PCA was done. Using the values of each pixel as the input components and dividing the data into two categories: positive and negative. This was done to get an idea of how well the samples can be separated using the histogram data. The PCA plot is shown in Figure 3.8, where the positive samples are indicated in red, and the negatives in blue. The positives and negatives overlap, but while the negatives are concentrated towards the lower end of the x-axis, the positives are more scattered across the plot. Component one, represented by the x-axis, explains 56.62% of the overall variance of the samples' fingerprints.

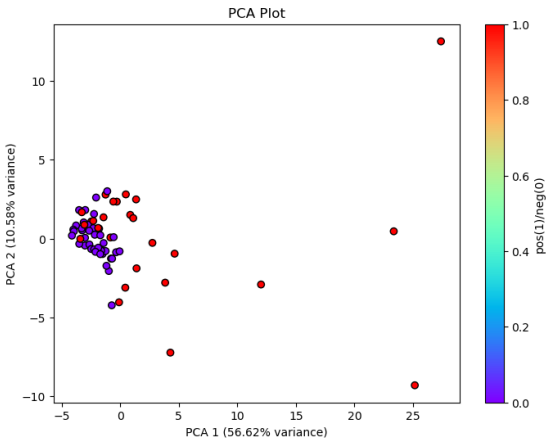


Figure 3.8: PCA for the fingerprints with corresponding 3-day HPC measurements. The distance between the dots represents the differences in the fingerprints. Red dots represent samples with more than 20 CFU/mL, and blue dots represent samples with less than or equal to 20 CFU/mL.

The PCA-plot shows that there is some separation between the positive and negative samples. If there is some difference between the positive and negative fingerprints, it is possible that a machine learning algorithm could discern patterns within the data and determine how a sample should be classified. To further investigate this separation, an LDA was done. The LDA showed a clear separation between the positives and negatives (Figure 3.9).

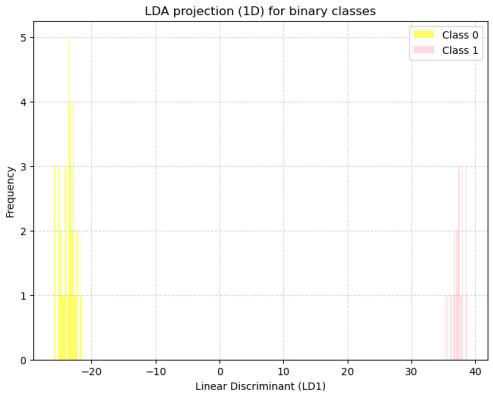


Figure 3.9: LDA for negative (class 0) and positive (class 1) with a 10x10 grid and a threshold of >20 CFU/mL

3.1.2.1 Logistic regression

First, downsampling was done to create a balanced data set. Using the Python library Scikit-learn and the function resample, the dataset was shrunk to 26 positives and 26 negatives. This was done 500 times, creating 500 different datasets. All datasets contain all positives, while the 26 negatives are randomized each time from the 41 total negatives. Each of the 500 balanced sets were divided into a training set and a testing set, where 80% of the set was used for training and 20% for testing.

The model used first was logistic regression, a supervised machine learning algorithm that is well-suited for predicting the probability of a binary outcome. When running the model, the mean accuracy was 83.4%, with a standard deviation of 0.104. This meant that, on average, in 83% of the test cases, the model predicted the correct label for the sample (more than or less than 20 CFU/mL). Figure 3.10 shows confusion matrices generated by the model. A confusion matrix shows in how many cases the model's prediction agrees with the true label of a sample, and ideally the pairings 1/0 and 0/1 would be as low in number as possible. In Figure 3.10 it can be seen that the model seems to be equally likely to make false negatives (1/0) as false positives (0/1).

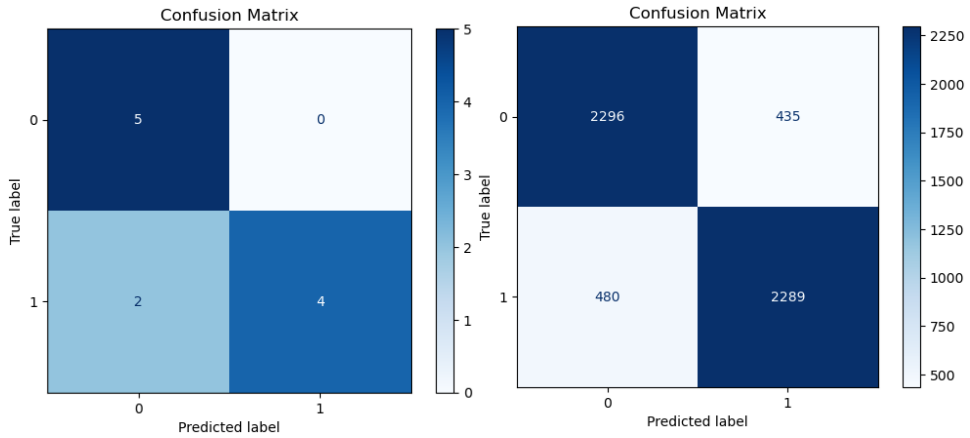


Figure 3.10: Confusion matrices for the logistic regression model. [Left] Example of confusion matrix for one of the 500 datasets. [Right] A cumulative confusion matrix, showing all outcomes of all datasets.

The parameters given as input were the event counts assigned to each of the 100 pixels for every sample. These pixels were of different importance to the model when

assigning an outcome (positive or negative). For logistic regression, this can be represented by the features' coefficients. A positive coefficient means that an increase in the features value indicates an increase of the likelihood of a positive classification. A negative coefficient indicates the opposite. The coefficients corresponding to each pixel have been colour-coded and plotted in Figure 3.11. For example, pixels 58, 18 and 42 seem to have a negative impact on the model outcome, while pixels 77, 8 and 16 have a positive impact.

To further investigate the feature importances a SHapley Additive exPlanations (SHAP) analysis was done (Appendix 4: SHAP plots). This shows feature 76 as the most important feature, and reports it as having a positive effect, i.e. an increase in feature 76 would increase the likelihood of a positive classification. In Figure 3.11 feature 76 has a low positive coefficient, and it would seem that this is in contradiction to the SHAP analysis. Feature 15 is ranked as the fourth most important, and should have a positive effect, according to the SHAP, while the coefficient suggests a negative effect. While this can seem contradictory, it is an artefact of the way the logistic regression model identifies features. While they are numbered 1-100 in the input, the model identifies them by the order in which they were input. This means that the first feature becomes numbered 0, the second 1 and so on [33]. Following this, the feature numbers in the SHAP analysis, as seen in Appendix 4 gets pushed back, and feature 15 equals feature 16 in Figure 3.11.

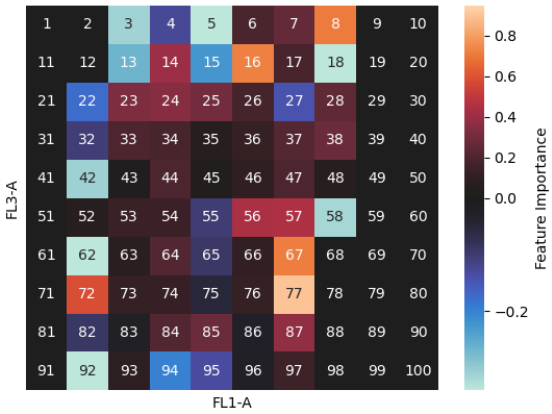


Figure 3.11: Colour-coded map of feature importances for the logistic regression model. Blue shades indicate a negative effect on the model outcome (higher values mean lower likelihood of positive classification). Orange shades indicate a positive effect on model outcome (higher values mean higher likelihood of a positive classification).

3.1.2.2 Random Forest

The other model used was the random forest algorithm. The preparations were the same as for the logistic regression model, with downsampling and creation of the training datasets. The function RandomForestClassifier from the Scikit-learn library, with all parameters set to their default value, was used. This resulted in a mean accuracy of 77%, with the standard deviation 0.12. Figure 3.12 shows the confusion matrices generated by the random forest model. The cumulative confusion matrix has

a higher number of false negatives, predictions in the 1/0 quadrant of the matrix, than false positives. This means that the model might prone to report that a sample is negative, in this case having less than 20 CFU/mL, when it is positive, and has more than 20 CFU/mL.

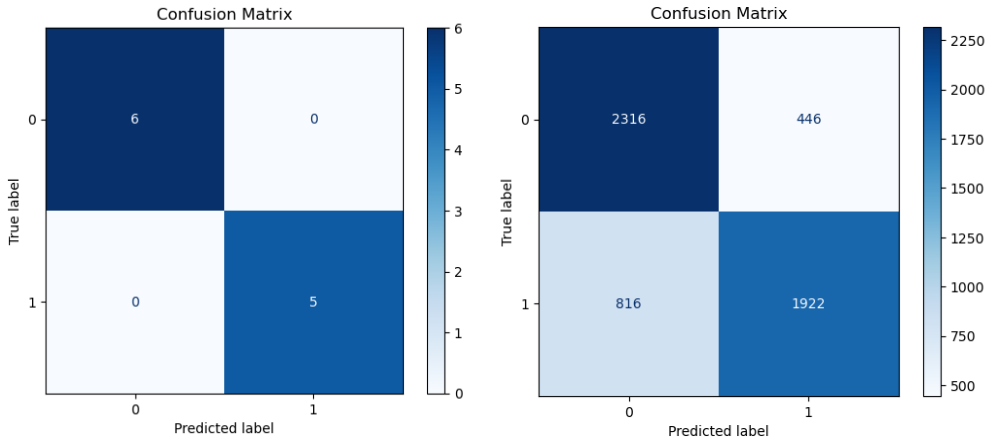


Figure 3.12: Confusion matrices for the random forest model. [Left] Example of confusion matrix for one of the 500 datasets. [Right] A cumulative confusion matrix, showing all outcomes of all datasets.

The model takes into account all the pixels created by the grid applied over the fingerprints (Figure 3.7). This does not mean that every feature is equally important in the decisions made by the model. In Figure 3.13 the importances of the different features are visualized using a heatmap. Feature 57 was most important for deciding whether a sample was to be considered positive or negative. This is a pixel located in the lower part of the HNA population in the fingerprint.

The heatmap in Figure 3.13 shows the importance of the features, and a subsequent SHAP-analysis showed how these features affected the decisions the model made (Appendix 4). Higher counts in feature 57, for example, made it more likely that the model would conclude that the sample was positive.

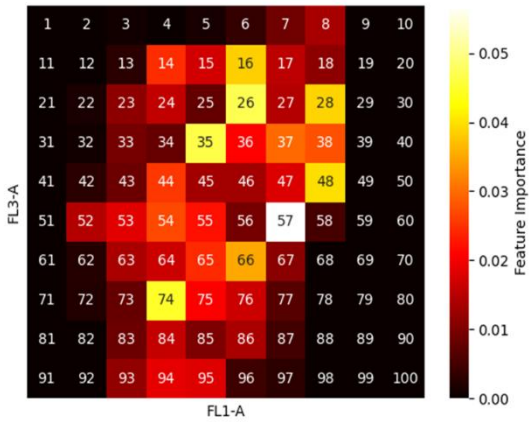


Figure 3.13: Heatmap showing the absolute feature importances for the random forest model with the fingerprints from SVOA.

3.2 Sydvatten

3.2.1 Exploratory data analysis

Initially, the same analysis was done as for the data from SVOA, where every TCC measurement were plotted versus its corresponding filter. This is shown in Figure 3.14. Filters 1, 3, 21 and 22 show more varied results, and to a lesser extent so do filters 9, 13 and 20.

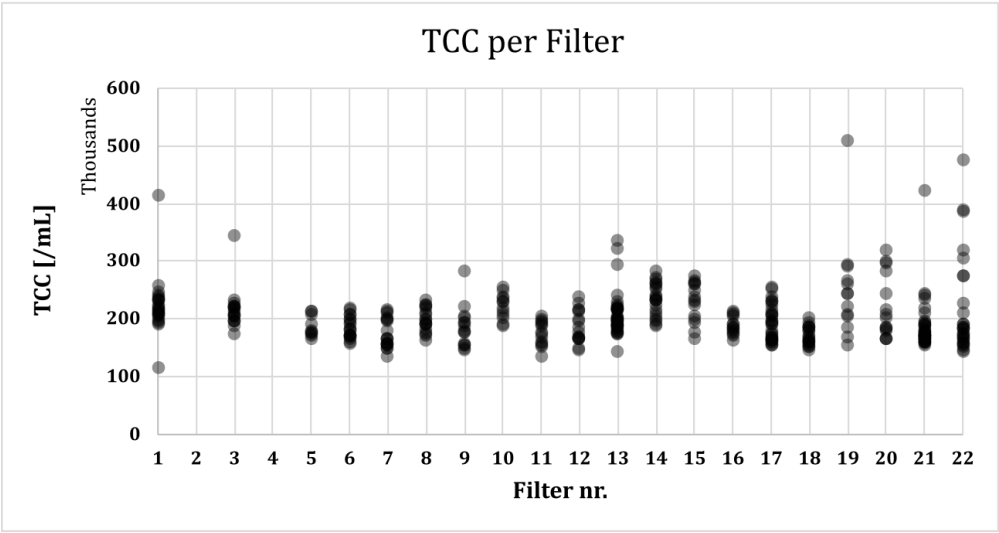


Figure 3.14: All TCC measurements for each filter at the Ringsjöverket DWTP (Sydvatten). Larger deviations can be seen for filters 1, 3, 21 and 22. There are smaller deviations for filters 9, 13 and 20. There are no filters 2 or 4.

The fractions of HNA bacteria were plotted in Figure 3.15 and are quite consistent for each filter. There are no clear peaks or valleys in any of the filters. Some are more varied, e.g. filters 1, 3 and 22. Others are very consistent, e.g. filters 9, 10 and 14. This could suggest that the filters with a wider spread have more fluctuating microbiomes, where different populations are dominant at different points in time.

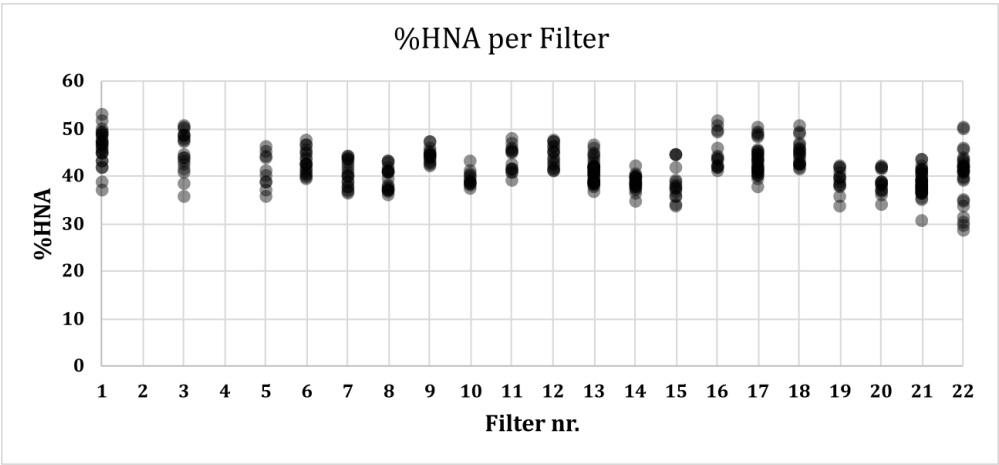


Figure 3.15: All %HNA values for each filter at the Ringsjöverket DWTP (Sydvatten). No filters look particularly scattered compared to one another.

To investigate the filters which showed spikes in TCC in Figure 3.14, these individual filters were plotted as seen in Figure 3.16. There, it can be seen that the highest TCC measured in SSF 1 coincided with the lowest head loss. There seems to be no clear correlation between the other parameters. Filters 3, 9, 13, 20, 21 and 22 can be found in “Appendix 2: In-depth plots for filters at Ringsjöverket”.

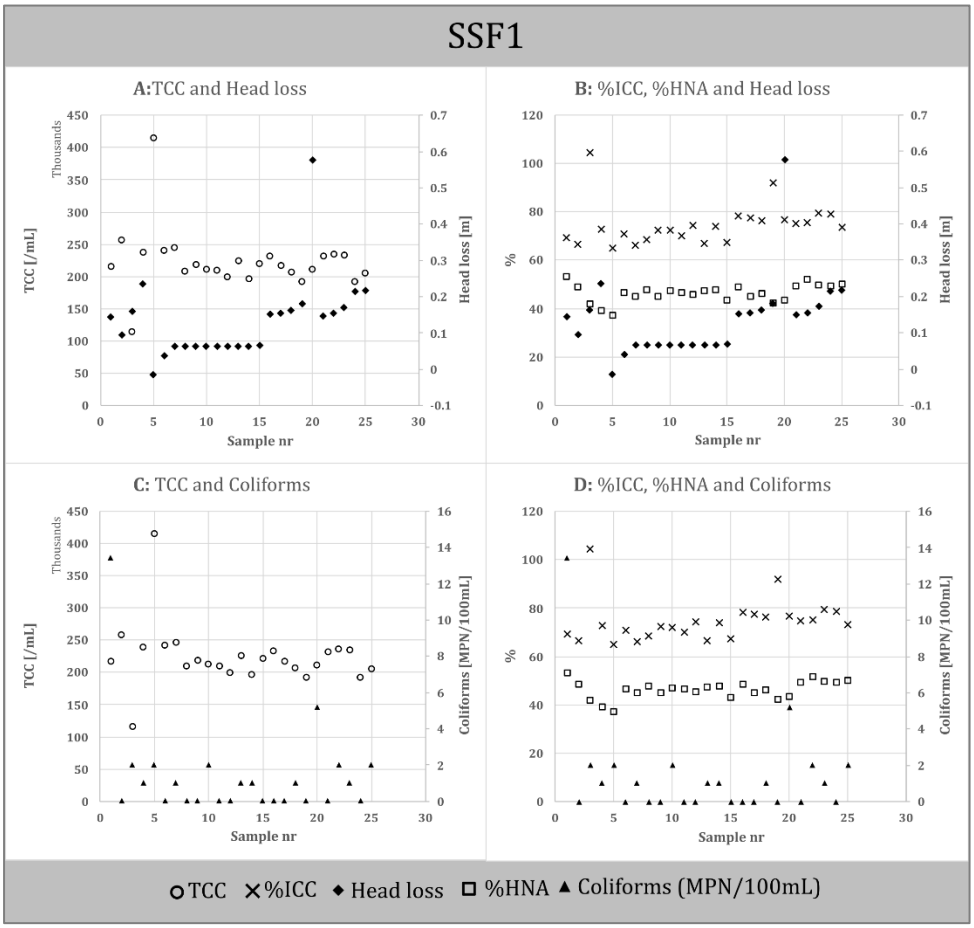


Figure 3.16: In-depth view of results for SSF1 at Ringsjöverket (Sydvatten). Values are plotted in the sampling order, without regard to the exact date. This was for clarity, which was otherwise lost because the results were collected over four summers. [A] TCC /mL and head loss. The highest TCC coincide with the lowest head loss. [B] %HNA, %ICC and Head loss. [C] TCC and coliforms. [D] %HNA, %ICC and coliforms.

Some percentages of intact cells were found to be over 100%, which is not possible. This could be an artifact of ICC and TCC being measured in different volumes of the same sample.

Each flow cytometric measurement was accompanied by Colilert®-18 analysis. %HNA and TCC was plotted against the coliform MPN/100mL and can be seen in Figures 3.I. Neither TCC nor %HNA seemed to follow any specific pattern in regard to coliform content in the water sample.

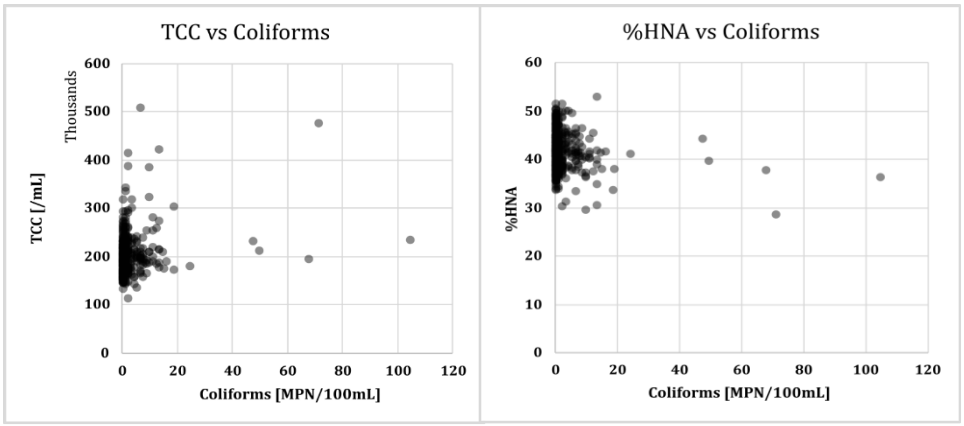


Figure 3.17: [Left] TCC plotted against MPN coliforms/100mL. No clear relationship can be seen. [Right] %HNA plotted against MPN coliforms/100mL. No clear relationship can be seen.

Coliform content was also compared to the ratio of intact cells measured in the FCM, and no relationship was found (not shown).

As could be seen in Figure 3.16 and Appendix 2, there seemed to be some correlation between lower head loss and higher TCC for the filters plotted in-depth. All TCC measurements were plotted against their corresponding head loss values. The result can be seen in Figure 3.18. There is a big variation in TCC in the lower end of the head loss range. Counts of 300 cells/mL were found to only occur when the head loss was lower than 0.1 m.

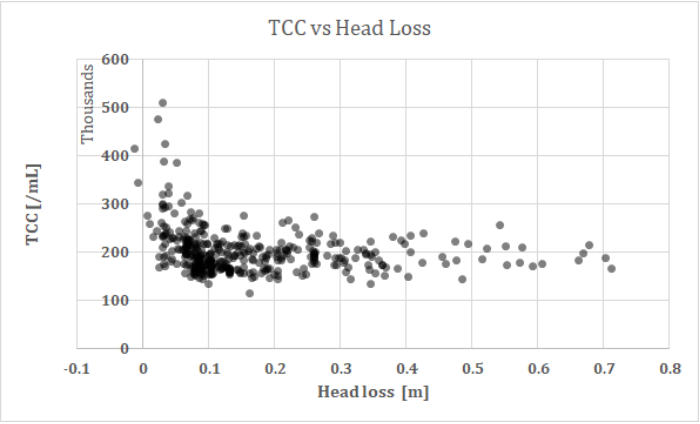


Figure 3.18: TCC /mL against head loss. The variation is largest in the lowest range of

It was also of interest to investigate if there were any correlations between head loss and %ICC, since this could give a view of how the function of the filter changes between scrapings. As can be seen in Figure 3.19, there is no clear correlation for %ICC and head loss. There are perhaps a bit more varied results in the lower range of the head, but no clear trends to neither the positive nor negative. This suggests that

even when the filters let through more total bacteria, the fraction of live bacteria doesn't seem to change.

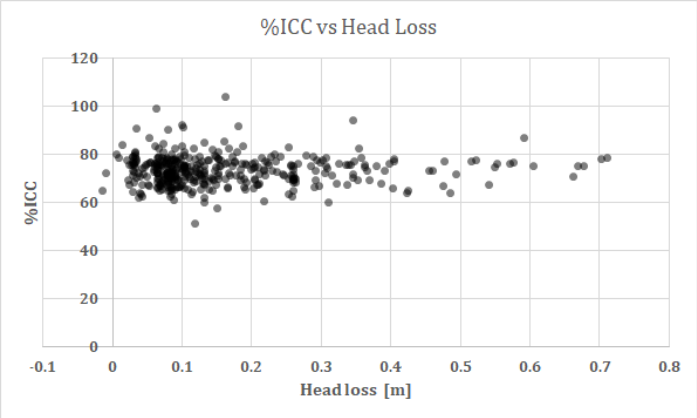


Figure 3.19: The percentage of intact cells plotted against the headloss at the date of sampling.

3.2.2 Machine Learning

In the data from Ringsjöverket, the main principal aim of the machine learning was to investigate the predictability of coliform counts from the two-dimensional histograms from the flow cytometric analysis. The histograms studied are those with red fluorescence on the y-axis and green fluorescence on the x-axis.

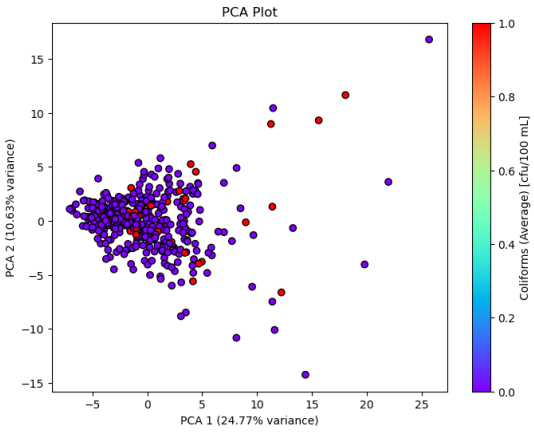


Figure 3.20: PCA showing the positive and nevative samples. The categories are completely overlapping.

To classify samples into positive and negative, a threshold of >8 coliforms/100mL was set. This resulted in 34 positive samples and 364 negative samples. The threshold was decided on using trial and error, running the model with different thresholds to achieve as high accuracy as possible. The separation between positives and negatives in the LDA did not change much between different thresholds, and the trial-and-error method was therefore chosen. To analyse the histograms, they were divided into grids of twelve-by-twelve pixels. This was decided by looking at the separation between positive and negative samples for different pixel counts in the LDA.

Downsampling was then performed, creating 500 datasets. Each dataset consisted of 34 positive and 34 negative samples. This meant that, just as in the case of Norsborg DWTP, every dataset contained the same positives and randomized negatives. Each of the 500 datasets were divided into 80%/20% training/testing sets.

3.2.2.1 Logistic regression

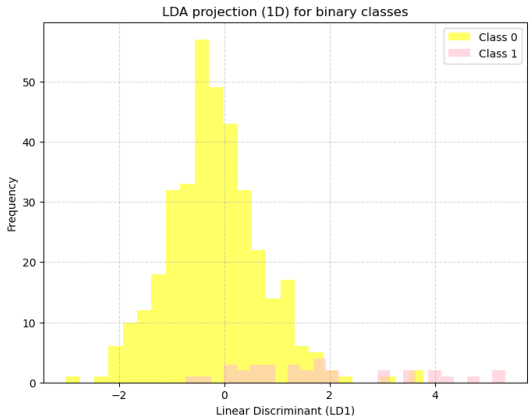


Figure 3.21: LDA plot, showing positive samples in pink and negative in yellow.

The logistic regression model gave a mean accuracy of 56.3% and the standard deviation 0.130. For reference, an accuracy of 50% in a binary classification such as this represents complete random guesses. This model is therefore only five percentage points above completely random predictions. The confusion matrices in Figure 3.22 also show the low accuracy, with 1501 predictions in the false positive quadrant, and 1556 in the false negative quadrant. The confusion matrix example from one of the training datasets shows that in this set, more false negatives than true positives were found. This, in essence, means that this model was not very useful for predicting coliform content in a water sample.

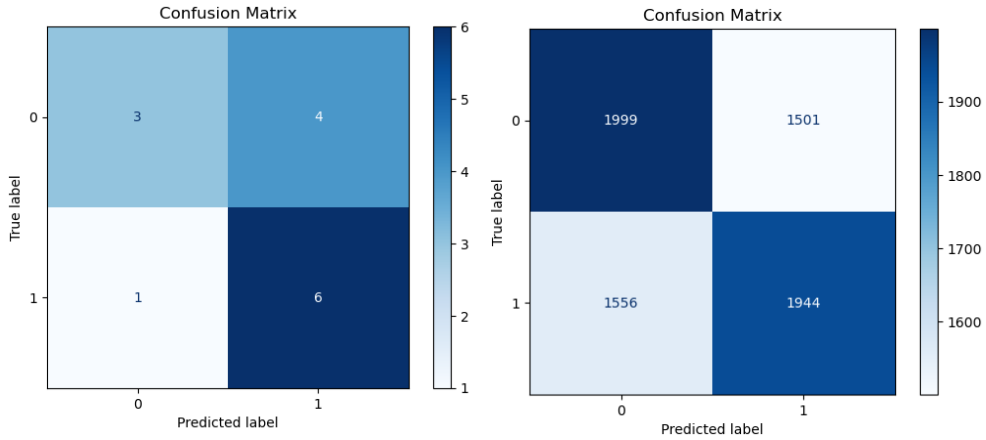


Figure 3.22: Confusion matrices for the logistic regression model. [Left] Example of confusion matrix for one of the 500 datasets. [Right] A cumulative confusion matrix, showing all outcomes of all datasets.

To investigate the importance of each feature, i.e. the pixels in the twelve-by-twelve grid, the feature coefficients from the logistic regression assigned to each pixel were plotted as seen in Figure 3.23. A negative coefficient (shown in shades of blue in Figure 3.23) means that when the feature value increases, the likelihood of the model classifying a sample as having more than the given threshold of coliform bacteria decreases. An increase in the input of a feature with a positive coefficient (orange shades in the figure) increases the likelihood of a positive classification. For example, the coefficients indicate that feature 29 would have a negative effect on the classification and features 87 and 92 a positive effect.

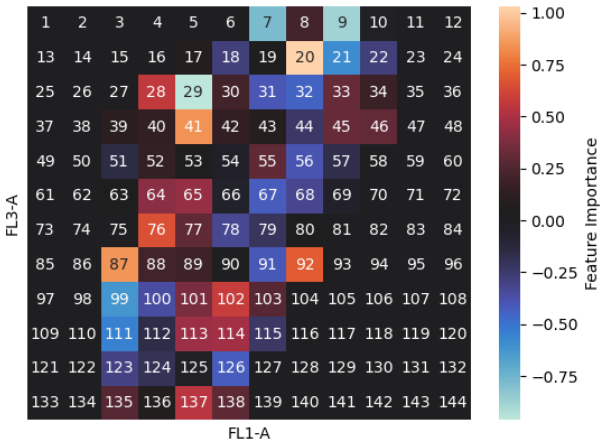


Figure 3.23: Colour-coded map of feature importances for the logistic regression model. Blue shades indicate a negative effect on the model outcome (higher values mean lower likelihood of positive classification). Orange shades indicate a positive effect on model outcome (higher values mean higher likelihood of a positive classification).

A SHAP-analysis was conducted to determine the feature importances for the logistic regression, which can be found in Appendix 4. The feature with the highest predictive power according to the SHAP analysis was feature 91, and that as the value in this

pixel increases, so does the likelihood of a positive classification. This can at first seem like it goes against what can be seen in Figure 3.23, where feature 91 is seen to have a negative coefficient, and should therefore have the opposite effect. However, this is, just as it was for the analysis of the data from SVOA, an effect of how the logistic regression model parses features. It does not store the feature names, and instead just identifies them by the order of which they are input. [33] This means that the first feature in the input will have the number 0 in the order of the logistic regression. This leads to the feature numbers in the SHAP in Appendix 4 being pushed back by one and feature 91 in the SHAP equals feature 92 in Figure 3.23.

3.2.2.2 Random Forest

The random forest model was applied to this dataset as well and was more accurate than the logistic regression. It gave a mean accuracy of 66.3%, with the standard deviation 0.115. The confusion matrices can be seen in Figure 3.24. The model output 1172 false negatives, and 1188 false positives. It does not seem like the model tended more towards either false positives nor false negatives.

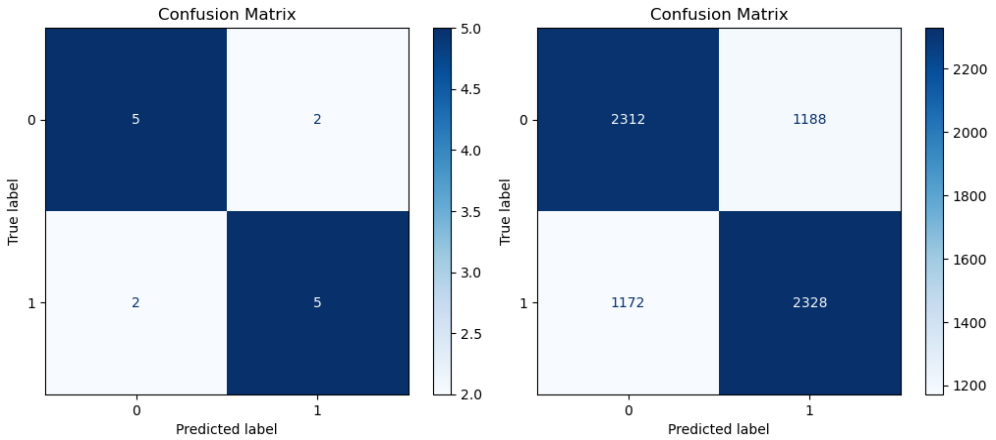


Figure 3.24: Confusion matrices for the random forest model. [Left] Example of confusion matrix for one of the 500 datasets. [Right] A cumulative confusion matrix, showing all outcomes of all datasets.

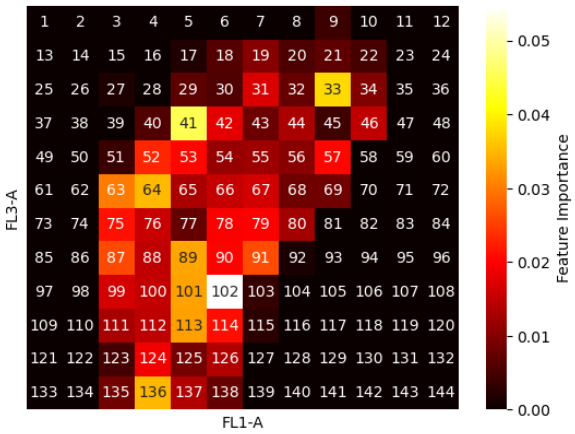


Figure 3.25: Heatmap showing the absolute importance of the features used as input for the random forest. Lighter colours indicate higher importances. Feature 102 is the lightest, and therefore seems to be the most important.

The feature importance has been visualised in Figure 3.25, where it can be seen that pixel 102 was important for classification. In the subsequent SHAP analysis (Appendix 4), pixel 102 was also at the top of the importance list, and reported as having a positive effect on the model outcome (higher counts in the pixel led to a higher likelihood of the sample being classified as positive). Pixel 102 is located in the lower part of the bulk of the population, closer to the LNA region than the HNA region.

3.3 Tekniska Verken

3.3.1 Exploratory data analysis

Data from four filters (SSF1, SSF2, SSF7 and SSF8) at Berggården DWTP was provided by Tekniska Verken. Two of these (SSF1 and SSF8) have been covered by roofs, while the other two are without cover. The roofs minimize the filters' exposure to light and the covered filters need to be scraped at about a fourth of the frequency as the uncovered filters.

Every total cell count value for each of the four filters were plotted, as seen in Figure 3.26. There are large variations within each filter, but the uncovered filters seem to trend slightly higher than the covered filters, though the overlap is large. All filters were plotted more in-depth, and this can be seen in "Appendix 3: In-depth plots for filters at Berggården".

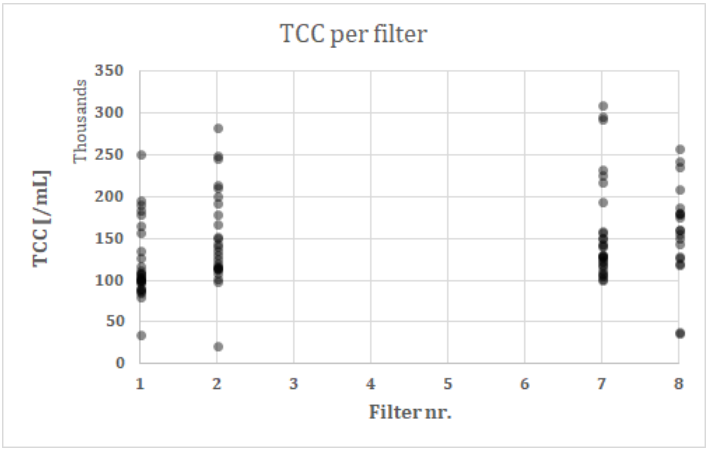


Figure 3.26: All TCCs from Berggården, grouped to their corresponding filter. The counts range from less than 50 000 cells/mL to over 300 000 cells/mL.

To get a closer look at the variation in TCC seen in Figure 3.26, the cell counts were plotted against their corresponding sampling date. This can be seen in Figure 3.27, where the filters have been classified as either covered or uncovered. There definitely is a time dependency in the TCC, with lower counts between April and August. During this time, only filter 1 of the covered filters were sampled. Filter 1 does seem to have slightly lower cell counts during the April-August period, but the significance of this is difficult to gauge.

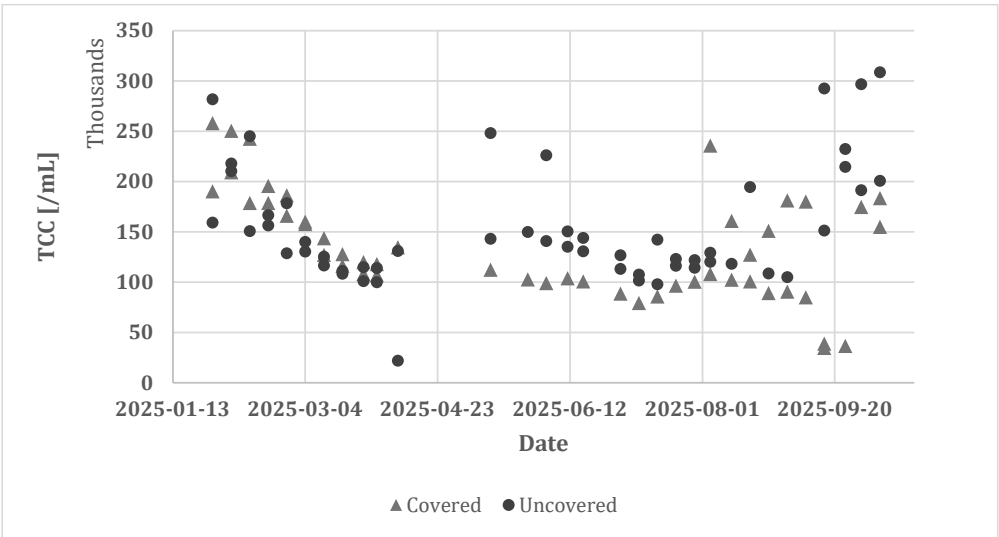


Figure 3.27: Total cell counts plotted against sampling date. Samples from covered filters are marked by triangles, samples from uncovered filters by dots.

The TCC:s were also plotted against the water temperature (Figure 3.28). Here it can be seen that when the temperatures are lower, all the filters are quite closely clustered. When the temperatures rise, the counts become more varied.

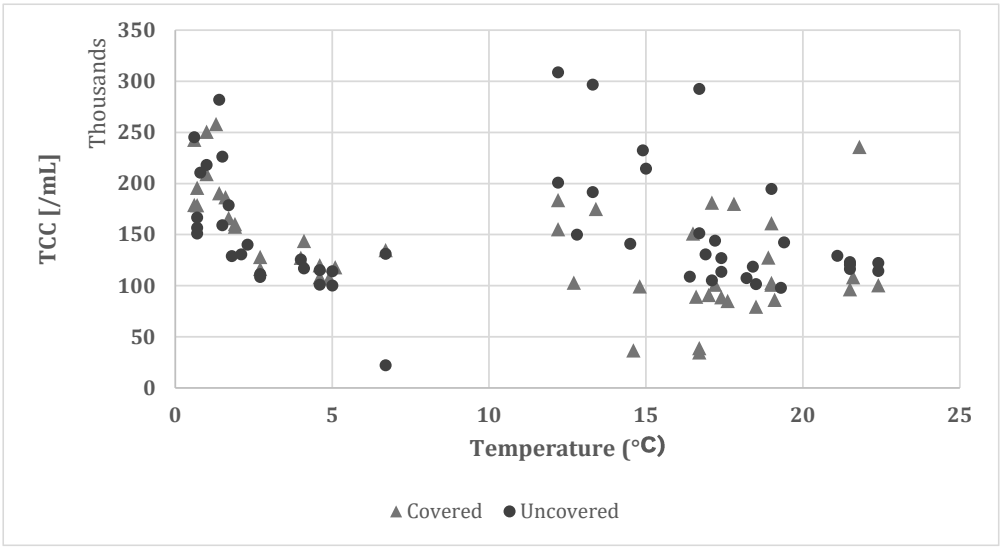


Figure 3.28: Total cell counts plotted against sampling water temperature. Samples from covered filters are marked by triangles, samples from uncovered filters by dots.

The relationship between head loss and TCC was also investigated. TCC plotted against head loss can be seen in Figure 3.29. In this data set, there is seemingly no clear correlation between TCC and head loss.

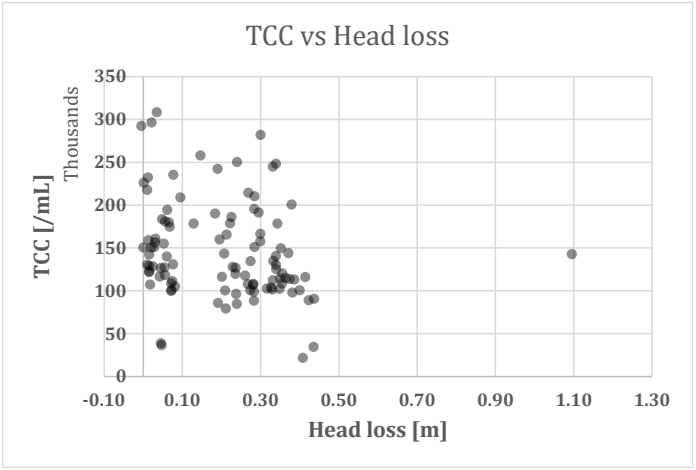


Figure 3.29: TCC plotted against the head loss measured on the sampling date. No correlation can be seen.

3.3.2 Machine learning

The data from Berggården did not contain a lot of samples positive for coliforms, and therefore the relationship between FCM and HPC. A PCA was done to see what the separation between positive and negative samples (Figure 3.30). The negative samples seem to almost completely overlap the positive, but the positives seem to cluster a little higher on PCA 2 than the negatives.

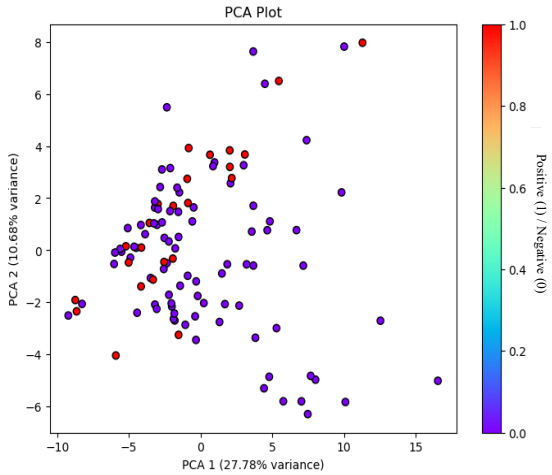


Figure 3.30: PCA showing the variance between positive (red dots) and negative (blue dots) samples.

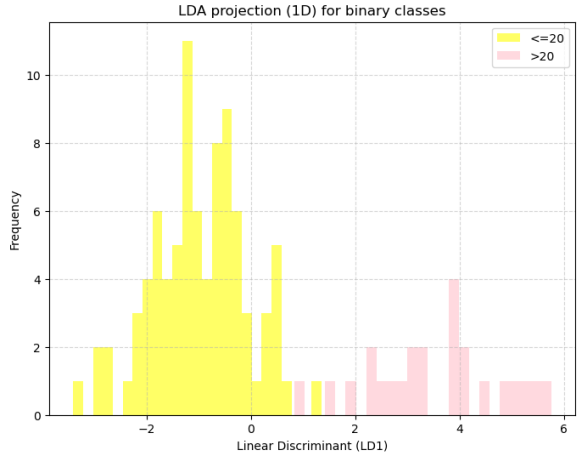


Figure 3.31: LDA plot showing the separation between positive (pink) and negative (yellow) samples. There is overlap, but the peaks for the two categories are separated.

An LDA was also done as to further investigate the possible separation between samples classified as negative and as positive (Figure 3.31). This was also used to set the parameters for the machine learning models. The plot in Figure 3.31 was generated when the grid was set to twelve-by-twelve pixels, and with the threshold >20 CFU/mL, and had the clearest separation between the two classes.

3.3.2.1 Logistic regression

The logistic regression gave a mean accuracy of 54.9%, with the standard deviation 0.146. This is similar to the result of the logistic regression applied to the data from Ringsjöverket. The confusion matrices can be seen in Figure 3.32. The model seems to be overall slightly more likely to output false negatives than false positives. The confusion matrix for only one of the downsampled datasets shows that only one positive sample was correctly identified.

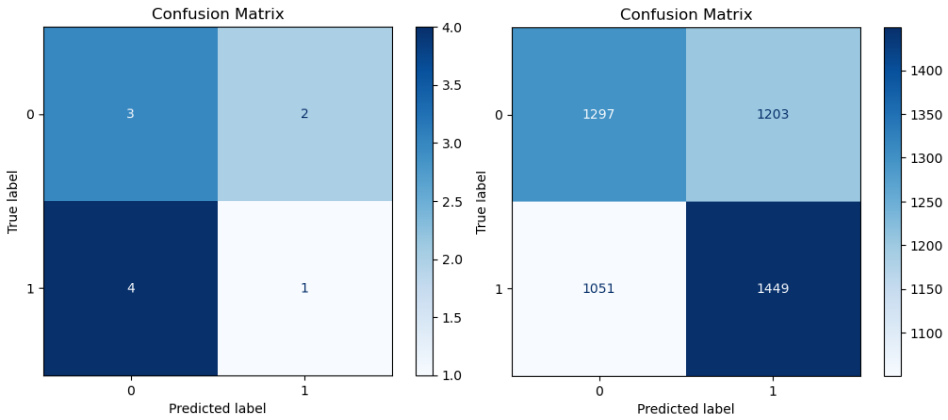


Figure 3.32: Confusion matrices for the logistic regression model. [Left] Example of confusion matrix for one of the 500 datasets. [Right] A cumulative confusion matrix, showing all outcomes of all datasets.

The colour-coded map in Figure 3.33 shows how the pixels affected the model outcome. In the SHAP analysis, which output is shown in Appendix 4, it can be seen that pixel 56 was the most important feature and had a negative effect on the model

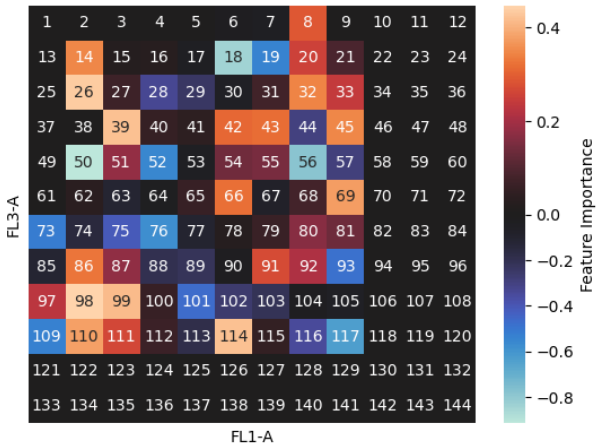


Figure 3.33: Colour-coded map of feature importances for the logistic regression model. Blue shades indicate a negative effect on the model outcome (higher values mean lower likelihood of positive classification). Orange shades indicate a positive effect on model outcome (higher values mean higher likelihood of a positive classification).

outcome. In Figure 3.33, it looks as if pixel 50 and 18 would be more important than pixel 56. This can be explained by the fact that the feature importance in the colourmap is based on the training data, i.e. the patterns the model found when training on the 500 downsampled sets, and the SHAP is based only on the test data. This means that the features in the SHAP data shows which features were used when classifying samples, and the colourmap shows what the model deemed important when training. As with the colourmaps from the logistic regressions for SVOA and Sydvatten, the feature numbers in the SHAP are shifted down by one, so pixel 56 in Figure 3.33 is listed as feature 55 in the SHAP.

3.3.2.2 Random forest

The random forest achieved higher accuracy than the logistic regression, with 60.5% mean accuracy and 0.146 standard deviation. Again, the cumulative confusion matrix showed a tendency for the model to give more false positives than false negatives (Figure 3.34).

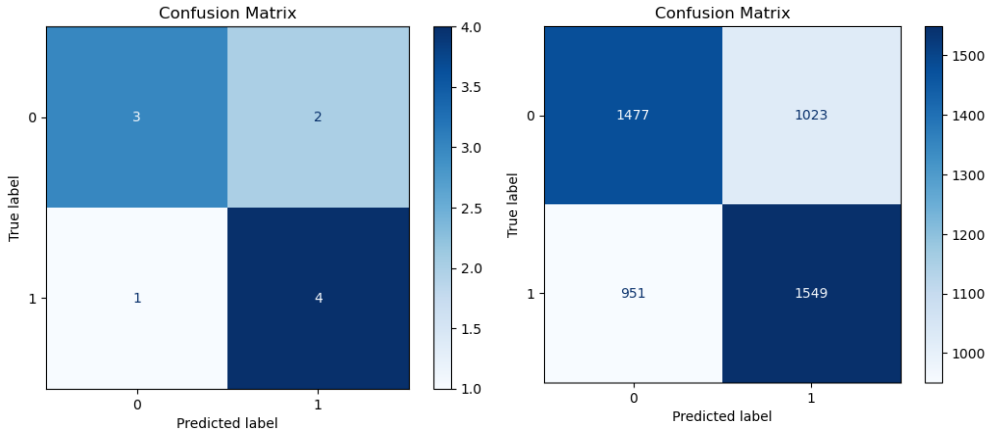


Figure 3.34: Confusion matrices for the random forest model. [Left] Example of confusion matrix for one of the 500 datasets. [Right] A cumulative confusion matrix, showing all outcomes of all datasets.

The heatmap in Figure 3.35 shows where the pixels with the highest predictive power were located in the grid. This showed that pixels 74, 75 and 76 were of high importance. This is also reflected in the SHAP analysis (Appendix 4), where these are also the top three most important features. All three are listed as having a negative effect on the classification, meaning that a higher cell count in these pixels made it less likely to have more than 20 CFU in the 3-day HPC.

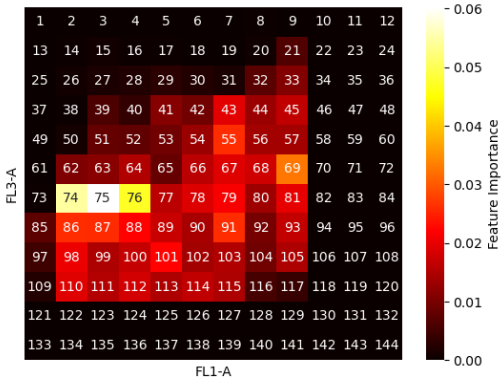


Figure 3.35: Heatmap showing the absolute importance of the features used as input for the random forest. Lighter colours indicate higher importances. Feature 75 is the lightest, and therefore seems to be the most important.

3.4 Comparing the drinking water treatment plants

The data provided by the drinking water treatment plants did not always overlap. All three included total cell counts measured using FCM and at least some coliform concentrations measured using Colilert®-18. The data from SVOA had very few datapoints where both FCM and coliform concentrations were available. Sydvatten had, in almost all cases, coliform data for every FCM measurement. It was the same for Tekniska verken, but with much fewer positive samples. SVOA and Tekniska verken also provided HPC results, while Sydvatten had no such data for individual filters.

While FCM data was provided by all three producers, there were some differences in what parameters had been measured and calculated. Sydvatten had TCC, ICC and %HNA measurements, SVOA had TCC, a very limited number of ICC and %HNA, and Tekniska verken had TCC and ICC. TCC and ICC require different dyes (SYBR green for TCC and SYBR green/PI for ICC), and ICC can therefore not be obtained from the FCS-files provided by the treatment plants, if the separate analyses have not been done. This meant that the percentage of intact cells could not be properly analysed for the dataset from SVOA. While it was technically possible to obtain the %HNA from the files provided by Tekniska verken, this was not done due to the time constraints to this project.

As for process parameters, all three treatment plants provided head loss data for the individual filters. Tekniska verken had temperature data for the individual filters at Berggården, while SVOA gave the temperature for the two collection pipes leading from the SSF:s. Sydvatten did not provide any temperature data.

There was also a difference in how the timeline for the samplings looked. SVOA had data from two years (2024 and 2025), and their samplings stretched across all seasons of those years. Sydvatten had data from four different years, but sampling was only done during the summer. TCC, the FCM measurements that overlapped for all drinking water treatment plants, was plotted against the sampling date (Figure 3.36). The dates are shown without regard of year, to visualize the seasonal changes in the total cell counts.

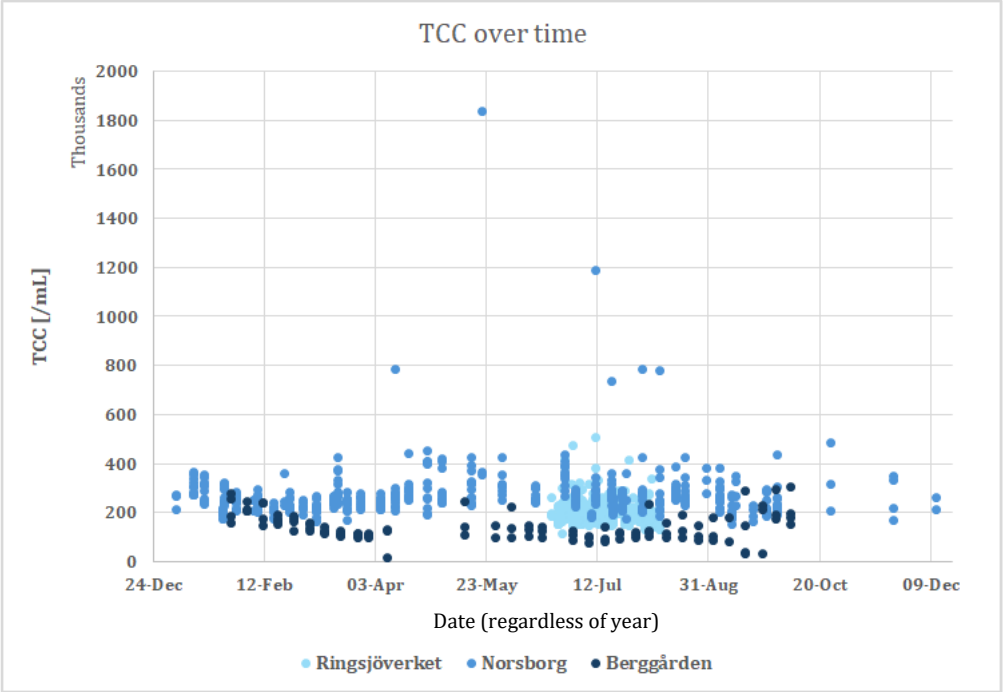


Figure 3.36: All TCC measurements, grouped by which DWTP they were taken at. The x-axis shows the date at which the sample was taken, but regardless of year. A sample taken June 1st, 2023, will be placed on the same date as a sample taken June 1st, 2025.

All three treatment plants have quite similar counts, but the seasonal changes seem to differ between the two where sampling stretched over longer time periods. Whereas Berggården's TCC:s decrease as sampling moves into April and then start increasing around August, Norsborg's instead increase over spring and summer.

When processing the data, new gates were applied to the fingerprints from Norsborg's DWTP and Ringsjöverket to ensure that the same gates were applied to all samples from the same DWTP. When this was done to the FCS-files from Berggården, the new total cell counts were two to three times higher than the total cell counts provided by Tekniska verken. This was not investigated further due to time constraints, and the already provided cell counts were used in the analyses described previously. This is also the reason no HNA content was available from Berggården, since this was not part of the data provided by Tekniska verken.

4. Discussion

In this study, data from three different drinking water treatment plants was reviewed. These DWTPs were: Norsborg's DWTP (Stockholm vatten och avfall, SVOA), Ringsjöverket (Sydvatten) and Berggården (Tekniska verken). Several different parameters were investigated. Since the datasets were pre-existing, the analyses that could be performed were dependent on which data had been collected by the respective water producer. Since the goal was to investigate the possibility of implementing FCM as a quality monitoring tool, it was of interest to compare the flow cytometric results to already established standard parameters, like 3-day HPC and coliform content. Few samples with more than 1 coliform/100mL were provided by both SVOA and Tekniska verken, so the HPC was chosen as the main parameter for these two. Sydvatten did not have any HPC data for individual filters, but many samples with high coliform content, and therefore coliforms/100mL was chosen here.

4.1 HPC, coliform content and FCM

4.1.1.3-day HPC at Norsborg's DWTP and Berggården

At Norsborg's DWTP (SVOA), HPC seemed to correlate with the total cell counts, where higher cell counts indicated higher HPCs (Figure 3.3, Figure 3.5). This correlation was not observed in the data from Berggården (Tekniska verken) (Figure ZC). When applying machine learning to the fingerprints in the datasets, the data from SVOA gave a quite high accuracy of 83% when using logistic regression, the model with the highest accuracy for this dataset. The feature identified as most important in this case was pixel 77 (Figure 3.11, Appendix 4), a part of the fingerprint in the HNA fraction when looking at the green fluorescence, but with lower red fluorescence than the bulk of the HNA population. The bacteria showing up in this part of the fingerprint therefore is either of a variety that can form colonies under the conditions of the 3-day HPC or is co-occurring with species that can.

This correlation between FCM results and HPC was not seen in the data from Berggården, neither in the exploratory part of the study nor when machine learning was applied. The logistic regression model performed worse than the random forest model, with a 55% accuracy and a 60% accuracy respectively. This difference in performance of the machine learning models for SVOA and Tekniska verken indicates that the correlation between flow cytometric data and heterotrophic plate counts is site-specific. This finding is partially reflected in the literature, where TCC and HPC has been compared. A linear relationship between TCC and HPC cannot be found in the literature, and when any correlation has been observed, it has been highly site-specific [34], [35].

Diversity in FCM fingerprints has been shown to correlate with the microbiological quality of drinking water [36]. Perhaps the bacteria showing up in the HPC (or bacteria that co-occur with these) at Norsborg's DWTP are just found in a part of the fluorescence fingerprint with fewer other events, enabling easier pattern detection. It could be that the effluent from the SSFs at Norsborg's DWTP is more uniform than

that of those at Berggården. SVOA has one covered SSF which was not sampled for HPC during the period studied, while half of the studied filters from Tekniska verken are covered. The difference in fingerprints from covered filters and uncovered filters was not investigated in this study but could possibly be a confounding factor when looking for patterns in relation to HPC.

The water entering Berggården is much lower in humic compounds than that entering Norsborg's DWTP [31]. Perhaps the higher access to organic matter at Norsborg's DWTP enable more bacterial growth, and so when the HPC increases, there are more species overall, and can therefore be more easily detected in the fingerprints, than in the low nutrition water at Berggården. These two waterworks have different processes leading up to the slow sand filters, with Berggården lacking flocculation and sedimentation before rapid sand filtration. It could be that there are bacterial communities that would be removed in a flocculation and sedimentation step that are confounding in the fingerprints, and lead to lower accuracy in the machine learning models for Berggården.

4.1.2 Coliform bacteria at Ringsjöverket

Because of lack of positive samples provided by SVOA and Tekniska verken, this relationship was only investigated for the dataset from Sydvatten. No relationship between FCM data and coliform content was observed in the exploratory data analysis, where TCC, %HNA and %ICC were compared to coliform content/100mL (Figure 3.17). This has been observed in previous studies, where FCM data has shown no strong correlations with coliform content [36], [37].

No strong correlation between the fluorescence fingerprints and coliform content was found in the machine learning, where the logistic regression model was 56% accurate and the random forest was 66% accurate. The best performing model, the random forest, listed pixel 102 as the most important feature when classifying the samples as having more or less than 8 coliforms/100mL (Figure 3.25). This is a pixel in the LNA region of the histogram, at the edge of the bulk of the population.

A previous study has suggested that analysing the fluorescence fingerprints could give information about indicator organisms such as coliform bacteria in water sampled at different points in the drinking water treatment process [38]. There, no gates were used, which is quite similar to how the data was analysed for this study, where the gates for the machine learning were set in a way so that the whole of the fingerprints would be within the grid. The divergence in findings could be due to differences in the treatment of water, where slow sand filters could introduce bacteria that cause the fingerprints to be more confounding than those of non-biological treatment steps.

The goal of using machine learning in this case was not to find exactly where in the fingerprint coliform bacteria can be seen, but if there were any parts of the fingerprint that change when coliforms can be found in a sample. This could mean either seeing

the coliform bacteria, or other bacteria that co-occur with them. In this case, it could be that these communities were too varied or diffuse for the models used in this study. Collecting more data to increase the sample size could possibly help resolve this.

As seen in the study by Chan et.al. (2018), the function of slow sand filters fluctuate throughout their maturation process and scraping cycles [22]. It is not possible to know without sequencing data or other much more specific cytometric methods, but perhaps the bacteria released in the effluent could therefore be different between filters and between sampling times for the same filter. This could cause fingerprints to differ in ways unrelated to coliform content and therefore confound the classification of samples. For future studies, it could therefore be helpful to sample the same filter throughout a longer period to remove the inconsistencies introduced by studying several filters simultaneously. It could also be beneficial to work out a more rigorous system of sampling times around scrapings, to avoid these events causing unnecessary confusion.

Ringsjöverket has a flocculation and sedimentation step in their process. While this could have been a reason for the relatively better performance of the machine learning models for Norsborg's DWTP, this does not seem to be the case for Ringsjöverket. The category "coliform bacteria" is more specific than "bacteria culturable on YEA" and if the flocculation and sedimentation help lessen the diversity in the bacterial communities, this does not seem to have been enough to lessen confusion for the machine learning models.

4.1.1 FCM in relation to the Swedish standards for drinking water

While there is no standard for what levels of coliform or culturable bacteria are allowed in the effluent of slow sand filters, it is still worth considering it. If FCM can be correlated to either coliform bacteria or 3-day HPC, it could mean up to three days faster results for these parameters, and a possibility of implementation on drinking water entering the distribution system. From what was seen in this study, more data and analysis is needed to get to a point where one can definitively state if FCM could be useful in determining drinking water quality. As previously stated, previous studies have shown little to no link between TCC and HPC, but little is known about the significance of fingerprint analysis for predicting HPC. Fingerprint diversity has been found to be useful in predicting the presence of indicator organisms like coliform bacteria, which could only be done with an accuracy of 66% during this project. More research is needed to determine if the predictive power of fingerprints is simply site-specific, or if the biological nature of SSF introduces confounding factors, making classifications difficult.

4.2 Logistic regressions and random forests

4.2.1 Logistic regression or random forest?

For both the Sydvatten and Tekniska verken data the random forest models were most accurate. This was not the case with the data from SVOA, for which logistic regression was more accurate. In the data analysis leading up to the machine learning, SVOAs data showed by far the largest separation between positive and negative samples, both in the PCA and LDA (Figure 3.8, Figure 3.9, Figure 3.20, Figure 3.21, Figure 3.30, Figure 3.31). While there was not full separation in the PCA, the negative samples clustered when the positives did not. Linear regression is a model well-suited for discreet, two-category classifications. It could be that this separation between the two categories led to logistic regression being more suited for this dataset than for the others.

4.2.2 Overfitting

There is also the issue of overfitting. Overfitting means that the model picks up on patterns in the noise of the training data and is therefore not particularly useable outside of the training setting. It makes for an unstable and unreliable model. Logistic regression is prone to overfitting when the number of input features is large and the sample size small [39]. In the case of Norsborg's DWTP, there were only 67 datapoints with HPC data attached to the FCM data. With 100 input features for each sample, it is very possible that the model could be overfit. This means that the accuracy seen, 83%, is not at all representative of how the model would perform in a real-life scenario, with data that has not been part of the training data.

Random forest is a model that reduces the risk of overfitting by utilizing several decision trees [30]. This could be the reason that the random forest performed worse for the data from SVOA, that the logistic regression overfits to the training data, and the random forest at least partially removes this effect. For future studies, it could be interesting to narrow down the number of features used for the logistic regression by identifying the most important ones and then running a model using only a few "most important features".

4.2.3 Normalisation of cell counts in the grid

The histograms in this study were not normalized prior to analysis. This meant that the total cell counts in each pixel created by the grid were the input features. Normalizing the cell counts by dividing each pixel's event count by the cell count in the pixel with the largest recorded cell count. This would make community changes clearer, allowing for more precise identification of small, abnormal changes in the fingerprint. The pixels identified as important in the models used in this study were often on the edges of the populations, where few events were recorded overall. Normalizing the data might make it so that small changes could be more easily detected, with less confusion caused by very high cell counts.

4.2.4 Applying one DWTP's model to another DWTP's data

For future studies, it could be of interest to investigate if a model trained on data from one DWTP's data could be applied to another DWTP's samples. This was difficult in this study, as data was collected independently by the water producers. This meant that three different flow cytometers and were used, as well as different software to compile the data. This meant that the proportions of the fingerprints were different, and a method to match them would have to be developed. If this is not done, the models used could get confused by mismatches caused by improper alignment of the fingerprints.

4.3 Head loss and FCM

4.3.1 TCC in relation to head loss

It was observed during this study that the total cell count seemed to correlate with the head loss in the individual filters at Norsborg's DWTP and Ringsjöverket (Figure 3.4, 3.K). It has previously been noted by Chan et.al. (2018) that the TCC increased the day after scraping a filter [22]. Scraping means removal of the top-most layer of the sand filter, the *schmutzdecke,* and is done when the head loss in the filter has gotten too large. The head loss is therefore at its lowest right after scraping. The relationship between head loss and TCC was not observed in the data from Tekniska verken (Figure3.29). This could be an effect of the sampling schedules at the different treatment plants. It is not known whether Sydvatten and SVOA sampled their filters near the time of scrapings. At Berggården, however, there were at least a week where no samples were taken after a filter had been scraped. It is possible that the scrapings were more important to the rise in the TCCs than the head loss itself, and this is the reason the relationship was not seen at Berggården.

4.3.2 %HNA and %ICC in relation to head loss

Looking at the %ICC and %HNA in relation to the head loss, no relationship was seen (Figure 3.19). What this suggests is that the population in the effluent has not changed much from the higher head loss samples to the lower head loss samples. This points towards the rise in TCC being on account of biofilm in the filter loosening and other microorganisms living in the filter, and not from breakthrough of bacteria not usually seen in the effluent. To further investigate a possible community change in the SSF effluent after scrapings, the fingerprints could be more closely studied. It could then be investigated if abnormal fingerprints correlate with the timing of scrapings.

4.4 Seasons and temperatures

For both Norsborg's DWTP and Berggården data was collected over several seasons. In Figure 3.36 the total cell counts have been plotted over time and shows how the TCCs fluctuate over a year. In the summer TCCs recorded at Berggården decrease, while those at Norsborg's DWTP largely seem to stay the same or increase a bit. There are many factors that could affect this, not least of which is the source water. The water in Mälaren is much more nutrient rich than that of Vättern, which could provide niches for opportunistic bacteria [40]. It could be that this leads to a larger increase of bacteria coming to Norsborg's DWTP than the increase of bacteria coming to Berggården. Slow sand filters function better at higher temperatures than at lower, so it could be this effect removes the "extra" bacteria seen in the summer and more at Berggården, but not at Norsborg's DWTP, due to difference in increases.

It was seen in the data from Berggården that the filters performed very similarly in colder temperatures and very differently in warmer (Figure 3.28). More data is needed, but it could be that differences in the filter communities become more apparent at higher temperatures. More in-depth analysis of the data would be needed, and it could be of interest to investigate the difference between covered and uncovered filters, using for example PERMANOVA or similar analyses.

5. Conclusions

This project's goals were to evaluate possible relationships between flow cytometric data and other analysis results in the quality assuring work at drinking water treatment plants, as well as investigating the implementation of machine learning for fingerprint analysis. While investigating the individual waterworks was relatively straightforward, comparing the results proved more difficult. The reason for this was largely on account of the different sampling practices of the DWTPs. While the same techniques were used, the datasets did not entirely overlap, and different instruments were used when collecting the FCM data. This meant that the same parameters could not be investigates at all three sites, and perhaps most importantly, the fingerprints from the DWTPS were stored differently.

Not being able to compare the fingerprints meant that the machine learning models could not be tested using fingerprints from a site they were not trained on. Because of this, it is not known if a drinking water producer would have to create their own dataset before implementing a method like those described in this report. This large initial investment might make such methods prohibitively expensive, since large volumes of data are needed to train models. This is especially so when there is no certainty about the accuracy of the models once trained.

In this study, a connection between HPC and TCC was found at Norsborg's DWTP, but not at Berggården. Both machine learning models were more accurate for the data from Norsborg's DWTP than for the data from Berggården. This speaks to a highly site-specific correlation; something reflected in the literature on the matter. This also makes implementation of a model predicting HPC from fluorescence fingerprints less feasible, as it is not certain a connection exists at all at a given site. More information is needed to investigate what factors contribute to the predictability of HPC, with the goal of finding ways of evaluating the feasibility of implementation at a studied DWTP.

No connection was seen between coliform bacteria and TCC, %ICC or %HNA. The correlation between the fingerprints and the coliform content also seemed relatively weak, given the accuracy of 66% from the random forest model and the poor separation of positive and negative samples in the PCA and LDA. It would therefore not be recommended that this sort of monitoring be implemented on any larger scale outside of research interests.

Overall, more research is needed to determine whether using FCM as quality monitoring system is a possibility at all DWTP, at only a few select sites with favourable conditions or too unreliable all together. There seems to be circumstances where it is applicable, as seen at Norsborg's DWTP in this project, and if this can be

replicated elsewhere, it would be a valuable tool for water producers wanting a fast and high through-put monitoring method. It could alert producers ahead of time, instead of only knowing about analysis results after the water has left the treatment plants and reached the distribution system. These methods show promise, but do not have enough evidence to substantiate a recommendation of more widespread implementation.

While this is true, it is important to note that flow cytometry should not be completely written off in the drinking water quality monitoring process. Machine learning applied to flow cytometric fingerprints have shown promise in other areas and could be of value here as well. Where the random forest and logistic regression models were found lacking, it could be useful to consider using an unsupervised learning model, like the isolation forest, to let the flow cytometric data stand on its own. Flow cytometry could be a powerful tool in and of itself, and it could be limiting to view it as a one-to-one replacement of the methods mentioned in the standards of today.

References

[1] UN-Water, ‘Summary Progress Update 2021: SDG 6 -- water and sanitation for all’. 2021.

[2] D. van der Kooij *et al.*, *Microbial Growth in Drinking Water Supplies*. IWA Publishing, 2013.

[3] C. Schleich *et al.*, ‘Mapping Dynamics of Bacterial Communities in a Full-Scale Drinking Water Distribution System Using Flow Cytometry’, *Water*, vol. 11, no. 10, p. 2137, Oct. 2019, doi: 10.3390/w11102137.

[4] H. L. and W. W. E., *Slow sand filtration*. Geneva: World Health Organization WHO, 1974.

[5] LIVSMEDELSVERKET, ‘Mikrobiologiska säkerhetsbarriärer’, *Kontrollwiki*. Livsmedelsverket.se, Dec. 18, 2024. Accessed: Sep. 08, 2025. [Online]. Available: https://kontrollwiki.livsmedelsverket.se/artikel/339/mikrobiologiska-sakerhetsbarriarer

[6] LIVSMEDELSVERKET, ‘Bedömning och rapportering av resultat’, *Kontrollwiki*. Livsmedelsverket.se, Jan. 30, 2025. Accessed: Sep. 09, 2025. [Online]. Available: https://kontrollwiki.livsmedelsverket.se/artikel/388/bedomning-och-rapportering-av-resultat

[7] LIVSMEDELSVERKET, ‘Indikatororganismer’, *Kontrollwiki*. Livsmedelsverket.se, Apr. 27, 2023. Accessed: Sep. 09, 2025. [Online]. Available: https://kontrollwiki.livsmedelsverket.se/artikel/150/indikatororganismer

[8] LIVSMEDELSVERKET, *Livsmedelsverkets föreskrifter om dricksvatten; LIVSFS 2022:12*, ISSN 1651-3533, Dec. 15, 2022.

[9] M. L. Davis, *Water and Wastewater Engineering - Design Principles and Practice*. 2010.

[10] J. J. Borrego and M. J. Figueras, ‘Microbiological quality of natural waters’, *Microbiología*, vol. 13, Dec. 1997.

[11] A. Rompré, P. Servais, J. Baudart, M.-R. de-Roubin, and P. Laurent, ‘Detection and enumeration of coliforms in drinking water: current methods and emerging approaches’, *Journal of Microbiological Methods*, vol. 49, no. 1, pp. 31–54, Mar. 2002, doi: 10.1016/S0167-7012(01)00351-7.

[12] ‘Colilert-18 - Total Coliform, E. coli, and Fecal Coliform Water Test Media for Laboratories - IDEXX US’. Accessed: Feb. 04, 2026. [Online]. Available: https://www.idexx.com/en/water/water-products-services/colilert-18/

[13] *Water quality - Enumeration of Escherichia coli and coliform bacteria - Part 2: Most probable method (ISO 9308-2:2012)*.

[14] W. L. Chao, ‘Evaluation of Colilert-18 for the detection of coliforms and Escherichia coli in tropical fresh water’, *Lett. Appl. Microbiol.*, vol. 42, no. 2, pp. 115–120, Feb. 2006, doi: 10.1111/j.1472-765X.2005.01814.x.

[15] *Water quality - Enumeration of culturable micro-organisms - colony count by inoculation in a nutrient agar culture medium (SS-EN ISO 6222)*, SS-EN ISO 6222, Oct. 22, 1999.

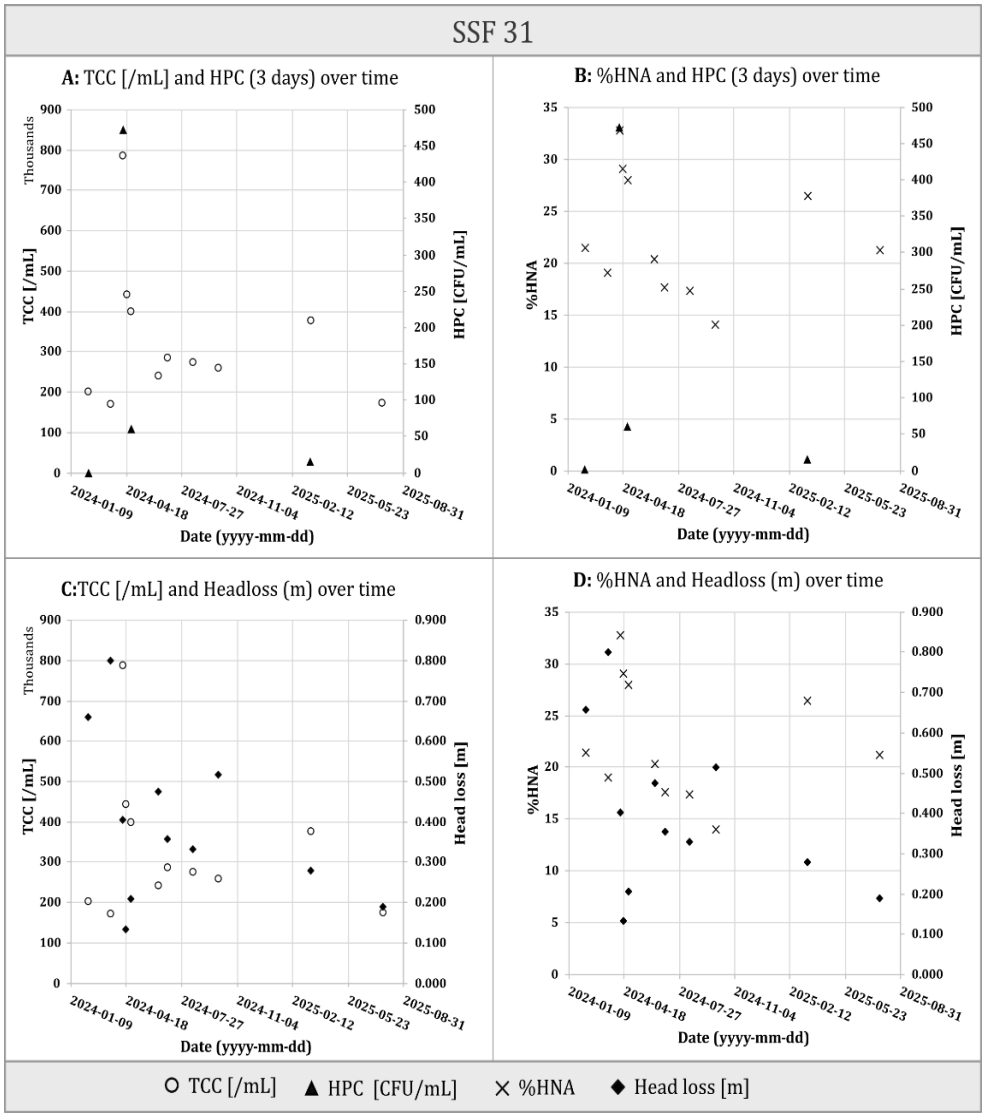
[16] J. Bartram, J. A. Cotruvo, M. Exner, C. Fricker, and A. Glasmacher, *Heterotrophic Plate Counts and Drinking-water Safety*. IWA Publishing, 2003.

[17] ‘Tolka resultatet av din dricksvattenanalys’. Accessed: Dec. 02, 2025. [Online]. Available: https://www.livsmedelsverket.se/livsmedel-och-innehall/dricksvatten/egen-brunn2/vattenprov-och-analys-av-ditt-dricksvatten/tolka-ditt-vattenanalysresultat/
[18] LIVSMEDELSVERKET, ‘Riktvärden’. Accessed: Dec. 02, 2025. [Online]. Available: https://www.livsmedelsverket.se/livsmedel-och-innehall/dricksvatten/egen-brunn2/vattenprov-och-analys-av-ditt-dricksvatten/lista-over-riktvarden/
[19] A. Adan, G. Alizada, Y. Kiraz, Y. Baran, and A. Nalbant, ‘Flow cytometry: basic principles and applications’, *Critical Reviews in Biotechnology*, vol. 37, no. 2, pp. 163–176, Feb. 2017, doi: 10.3109/07388551.2015.1128876.
[20] K. M. McKinnon, ‘Flow Cytometry: An Overview’, *Current Protocols in Immunology*, vol. 120, no. 1, p. 5.1.1-5.1.11, Jan. 2018, doi: 10.1002/cpim.40.
[21] M. Berney, M. Vital, I. Hülshoff, H.-U. Weilenmann, T. Egli, and F. Hammes, ‘Rapid, cultivation-independent assessment of microbial viability in drinking water’, *Water Research*, vol. 42, no. 14, pp. 4010–4018, Aug. 2008, doi: 10.1016/j.watres.2008.07.017.
[22] S. Chan, K. Pullerits, J. Riechelmann, K. M. Persson, P. Rådström, and C. J. Paul, ‘Monitoring biofilm function in new and matured full-scale slow sand filters using flow cytometric histogram image comparison (CHIC)’, *Water Research*, vol. 138, pp. 27–36, Jul. 2018, doi: 10.1016/j.watres.2018.03.032.
[23] F. Hammes, C. Berger, O. Köster, and T. Egli, ‘Assessing biological stability of drinking water without disinfectant residuals in a full-scale water supply system’, *Journal of Water Supply: Research and Technology-Aqua*, vol. 59, no. 1, pp. 31–40, Feb. 2010, doi: 10.2166/aqua.2010.052.
[24] M. K. Ramseier, U. von Gunten, P. Freihofer, and F. Hammes, ‘Kinetics of membrane damage to high (HNA) and low (LNA) nucleic acid bacterial clusters in drinking water by ozone, chlorine, chlorine dioxide, monochloramine, ferrate(VI), and permanganate’, *Water Research*, vol. 45, no. 3, pp. 1490–1500, Jan. 2011, doi: 10.1016/j.watres.2010.11.016.
[25] S. Barbesti, S. Citterio, M. Labra, M. D. Baroni, M. G. Neri, and S. Sgorbati, ‘Two and three-color fluorescence flow cytometric analysis of immunoidentified viable bacteria’, *Cytometry*, vol. 40, no. 3, pp. 214–218, Jul. 2000, doi: 10.1002/1097-0320(20000701)40:3%3C214::AID-CYTO6%3E3.0.CO;2-M.
[26] G. Grégori *et al.*, ‘Resolution of Viable and Membrane-Compromised Bacteria in Freshwater and Marine Waters Based on Analytical Flow Cytometry and Nucleic Acid Double Staining’, *Appl Environ Microbiol*, vol. 67, no. 10, pp. 4662–4670, Oct. 2001, doi: 10.1128/AEM.67.10.4662-4670.2001.
[27] A. Nescerecka, F. Hammes, and T. Juhna, ‘A pipeline for developing and testing staining protocols for flow cytometry, demonstrated with SYBR Green I and propidium iodide viability staining’, *Journal of Microbiological Methods*, vol. 131, pp. 172–180, Dec. 2016, doi: 10.1016/j.mimet.2016.10.022.
[28] ‘What is Machine Learning? | IBM’. Accessed: Feb. 03, 2026. [Online]. Available: https://www.ibm.com/think/topics/machine-learning
[29] ‘What Is Logistic Regression? | IBM’. Accessed: Feb. 03, 2026. [Online]. Available: https://www.ibm.com/think/topics/logistic-regression

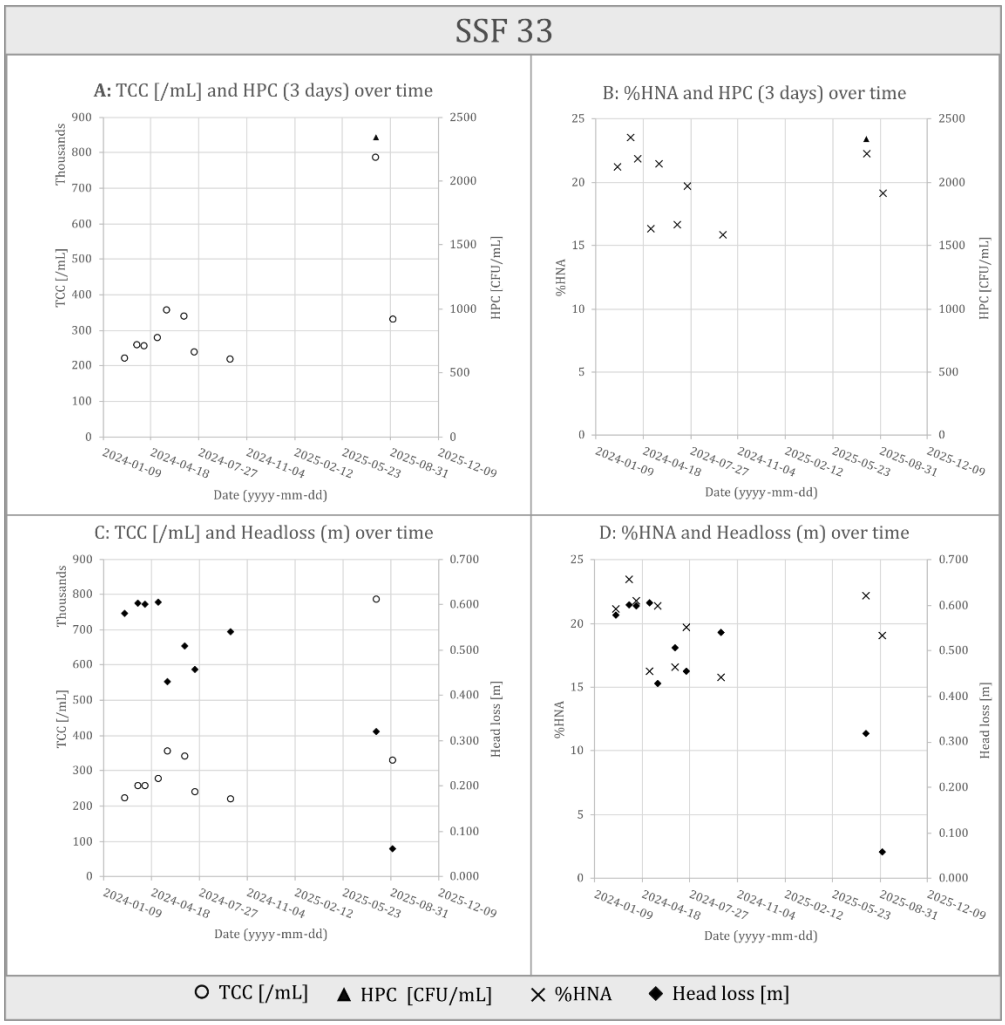
[30] ‘What Is Random Forest? | IBM’. Accessed: Feb. 03, 2026. [Online]. Available: https://www.ibm.com/think/topics/random-forest
[31] ‘VISS-Vatteninformationssystem Sverige’. Accessed: Feb. 02, 2026. [Online]. Available: http://viss.lansstyrelsen.se
[32] I. K. Erb, N. Gador, M. Jinbäck, E. Lindberg, and C. J. Paul, ‘A data-driven early warning system for Escherichia coli in water based on microbial community analysis using flow cytometry 2D histograms’, *Water Research X*, vol. 29, Dec. 2025, doi: https://doi.org/10.1016/j.wroa.2025.100404.
[33] ‘LogisticRegression’, scikit-learn. Accessed: Jan. 24, 2026. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[34] ‘Understanding the Use of Flow Cytometry for Monitoring of Drinking Water’, Drinking Water Inspectorate. Accessed: Feb. 01, 2026. [Online]. Available: https://www.dwi.gov.uk/research/completed-research/monitoring-microbiological/understanding-the-use-of-flow-cytometry-for-monitoring-of-drinking-water/
[35] S. Van Nevel *et al.*, ‘Flow cytometric bacterial cell counts challenge conventional heterotrophic plate counts for routine microbiological drinking water monitoring’, *Water Research*, vol. 113, pp. 191–206, Apr. 2017, doi: 10.1016/j.watres.2017.01.065.
[36] L. Claveau, N. Hudson, P. Jarvis, P. Jeffrey, and F. Hassard, ‘Microbial water quality investigation through flow cytometry fingerprinting: from source to tap’, *Sustain. Microbiol.*, vol. 1, no. 1, p. qvae003, Jan. 2024, doi: 10.1093/sumbio/qvae003.
[37] R. Cheswick *et al.*, ‘Comparing flow cytometry with culture-based methods for microbial monitoring and as a diagnostic tool for assessing drinking water treatment processes’, *Environment International*, vol. 130, p. 104893, Sep. 2019, doi: 10.1016/j.envint.2019.06.003.
[38] L. Claveau, N. Hudson, P. Jeffrey, and F. Hassard, ‘To gate or not to gate: Revisiting drinking water microbial assessment through flow cytometry fingerprinting’, *Science of The Total Environment*, vol. 912, p. 169138, Feb. 2024, doi: 10.1016/j.scitotenv.2023.169138.
[39] J. C. Stoltzfus, ‘Logistic Regression: A Brief Primer’, *Academic Emergency Medicine*, vol. 18, no. 10, pp. 1099–1104, 2011, doi: 10.1111/j.1553-2712.2011.01185.x.
[40] B. Brindefalk *et al.*, ‘Bacterial composition in Swedish raw drinking water reveals three major interacting ubiquitous metacommunities’, *Microbiologyopen*, vol. 11, no. 5, p. e1320, Sep. 2022, doi: 10.1002/mbo3.1320.

APPENDIX

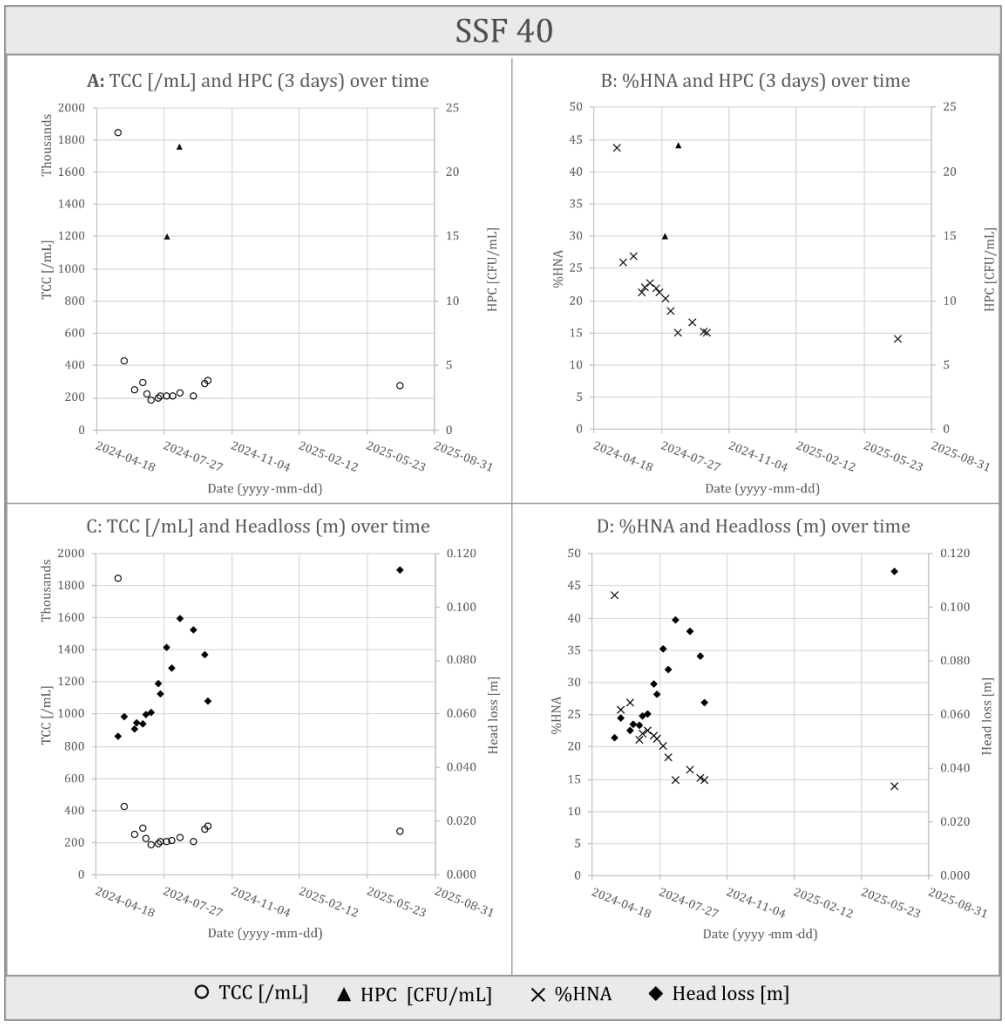
Appendix 1: In-depth plots for filters at Norsborg's DWTP



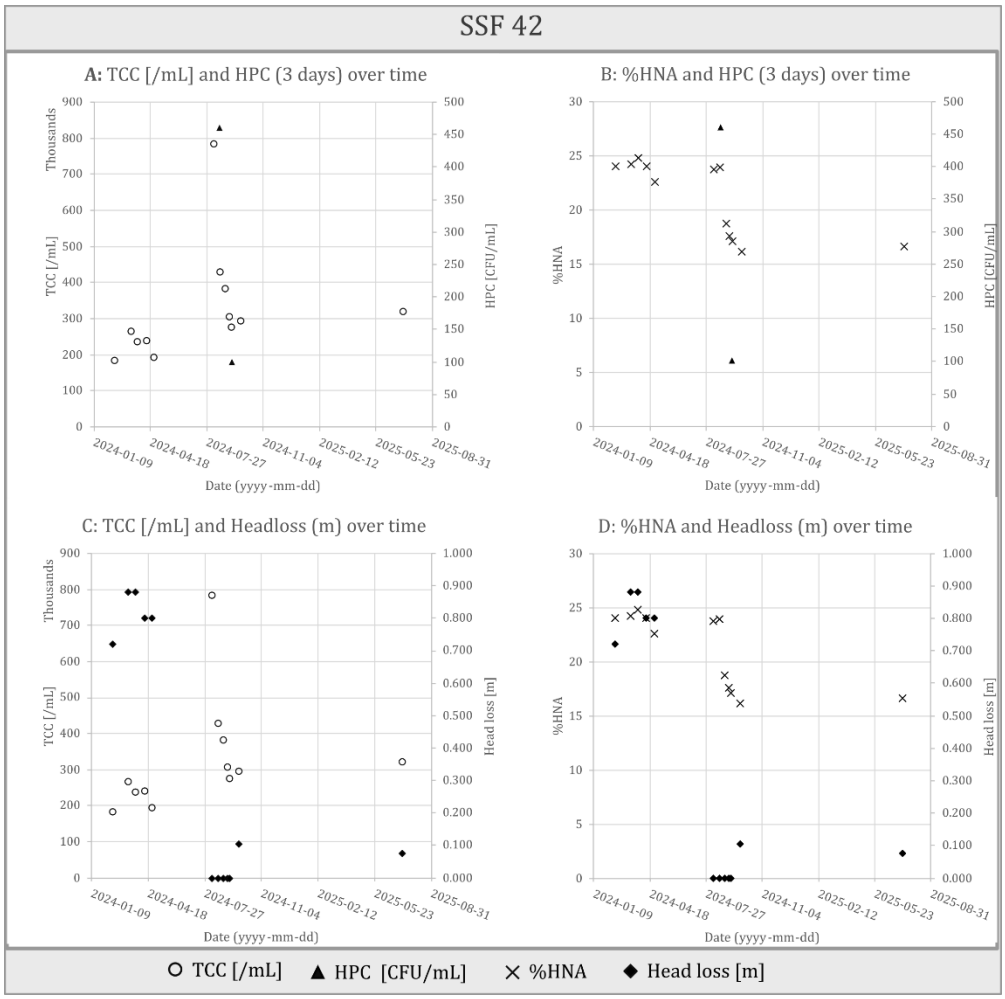
A1.1: Plots showing TCC, HPC, %HNA and head loss for filter 31 at Norsborg's DWTP.



A1.2: Plots showing TCC, HPC, %HNA and head loss for filter 33 at Norsborg's DWTP.

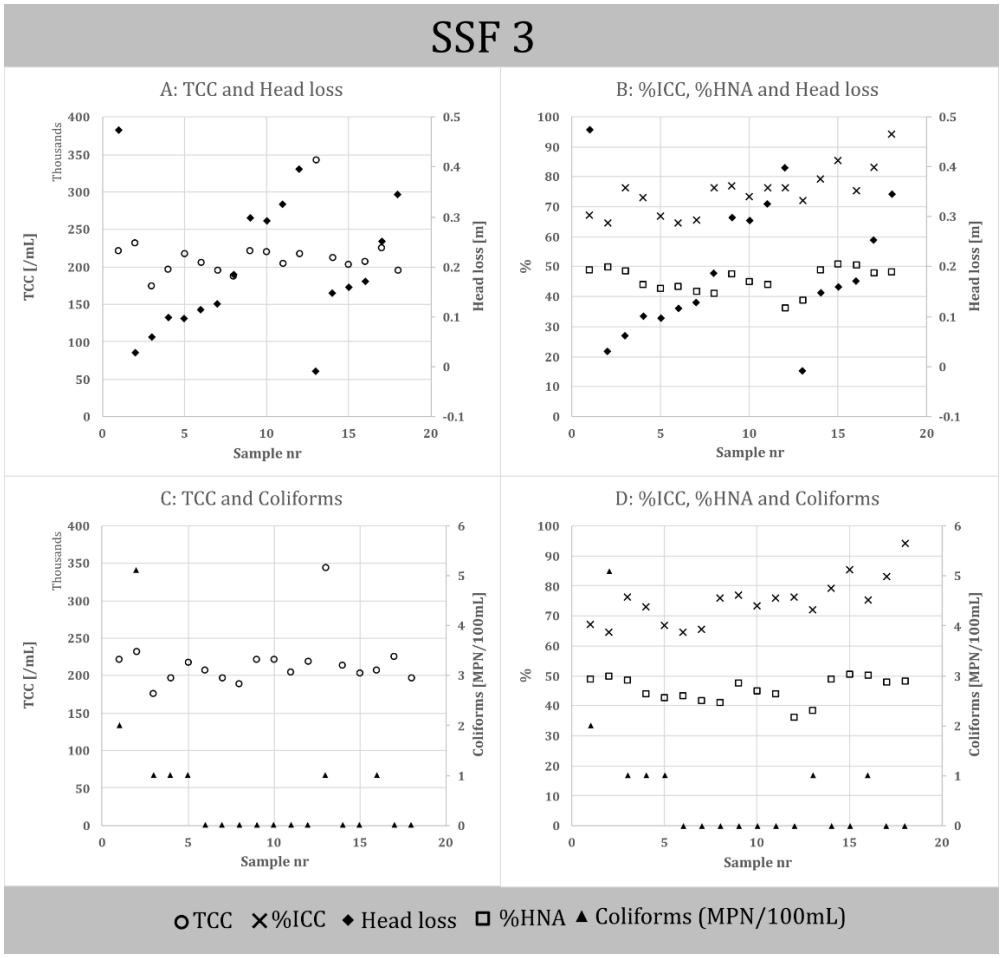


A1.3: Plots showing TCC, HPC, %HNA and head loss for filter 40 at Norsborg's DWTP.

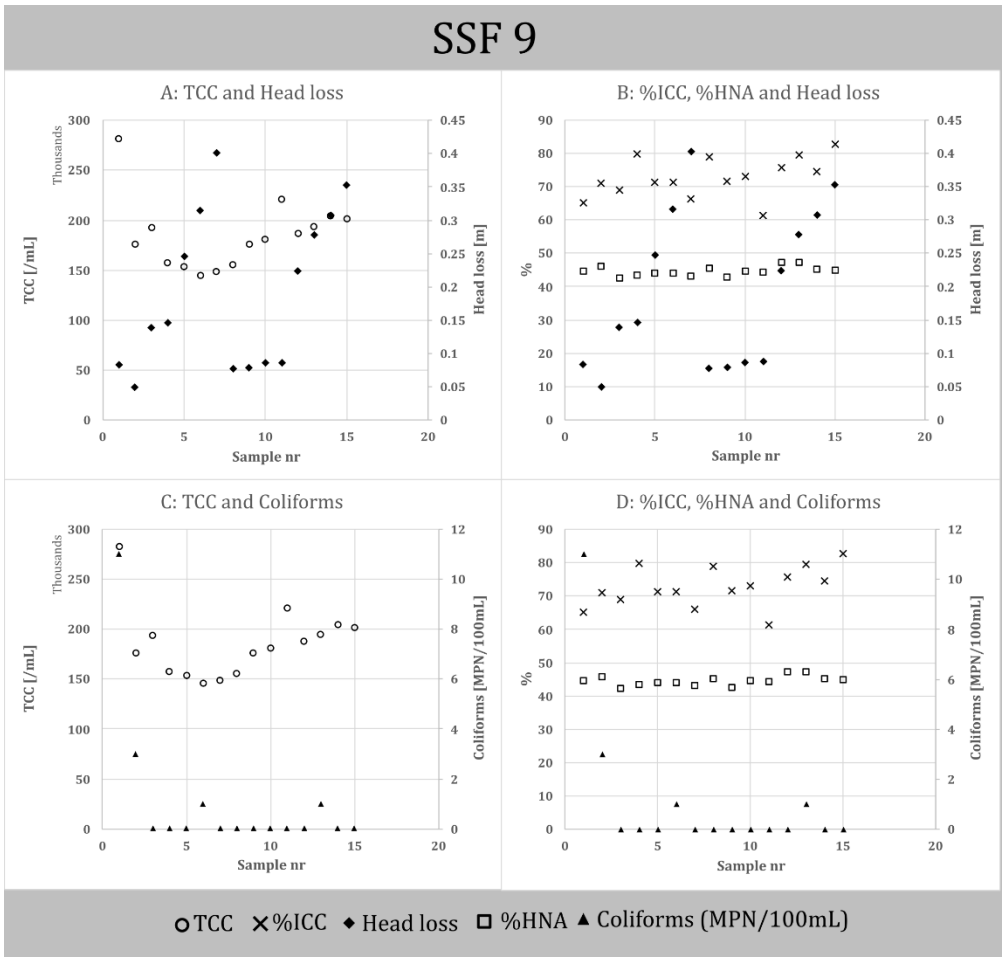


A1.4: Plots showing TCC, HPC, %HNA and head loss for filter 42 at Norsborg's DWTP.

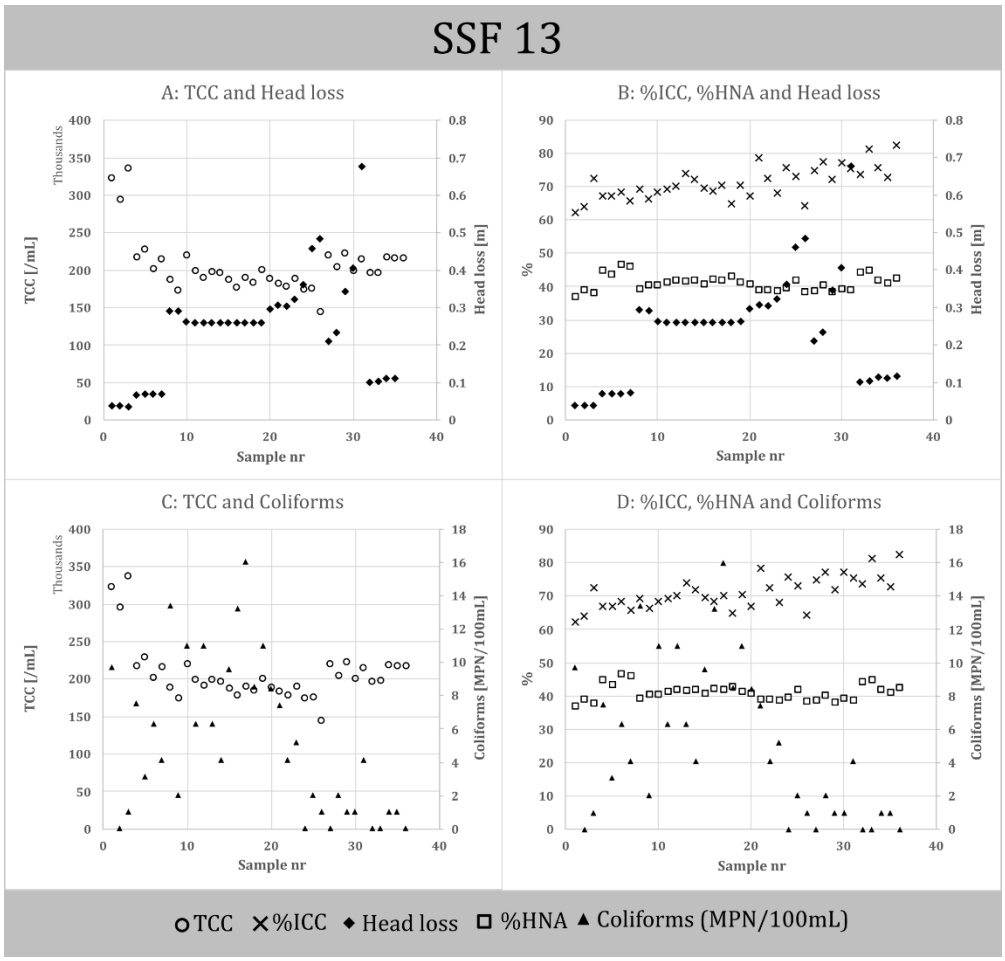
Appendix 2: In-depth plots for filters at Ringsjöverket



A2.1: Plots showing TCC, %ICC, head loss, %HNA and coliforms for filter 3 at Ringsjöverket.

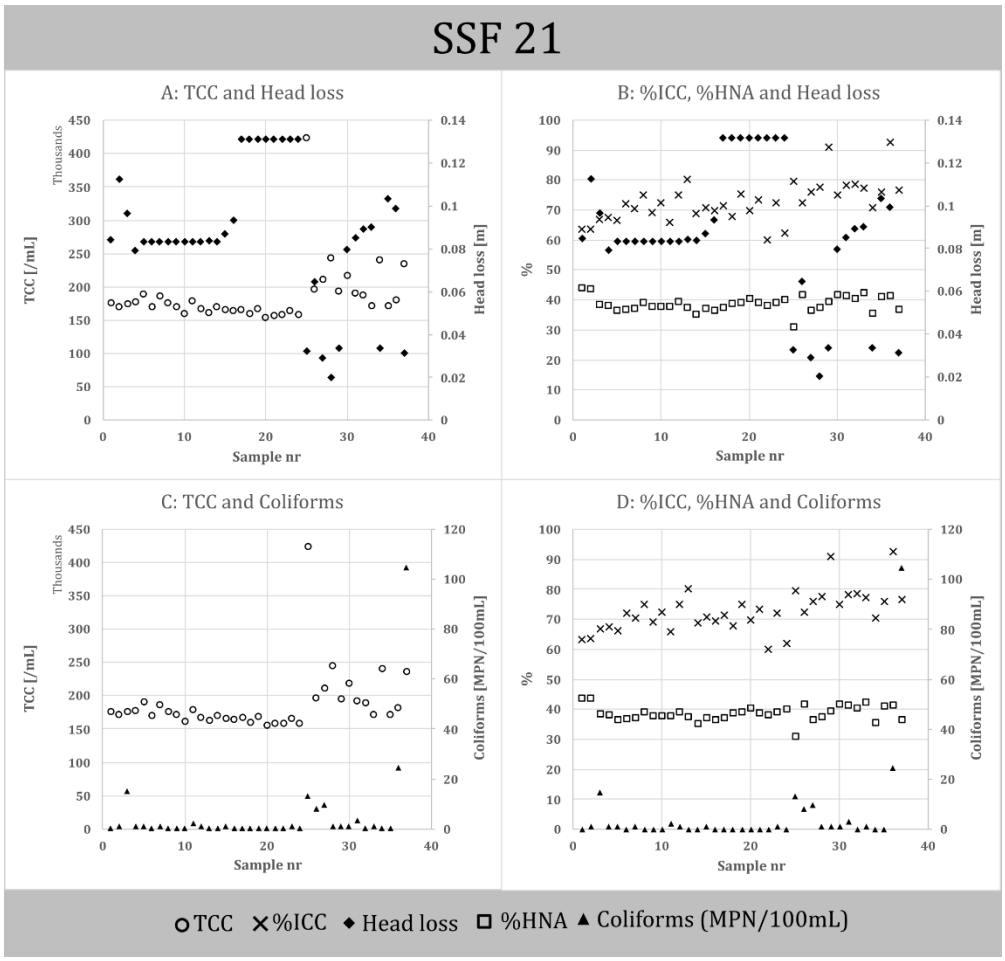


A2.2: Plots showing TCC, %ICC, head loss, %HNA and coliforms for filter 9 at Ringsjöverket.

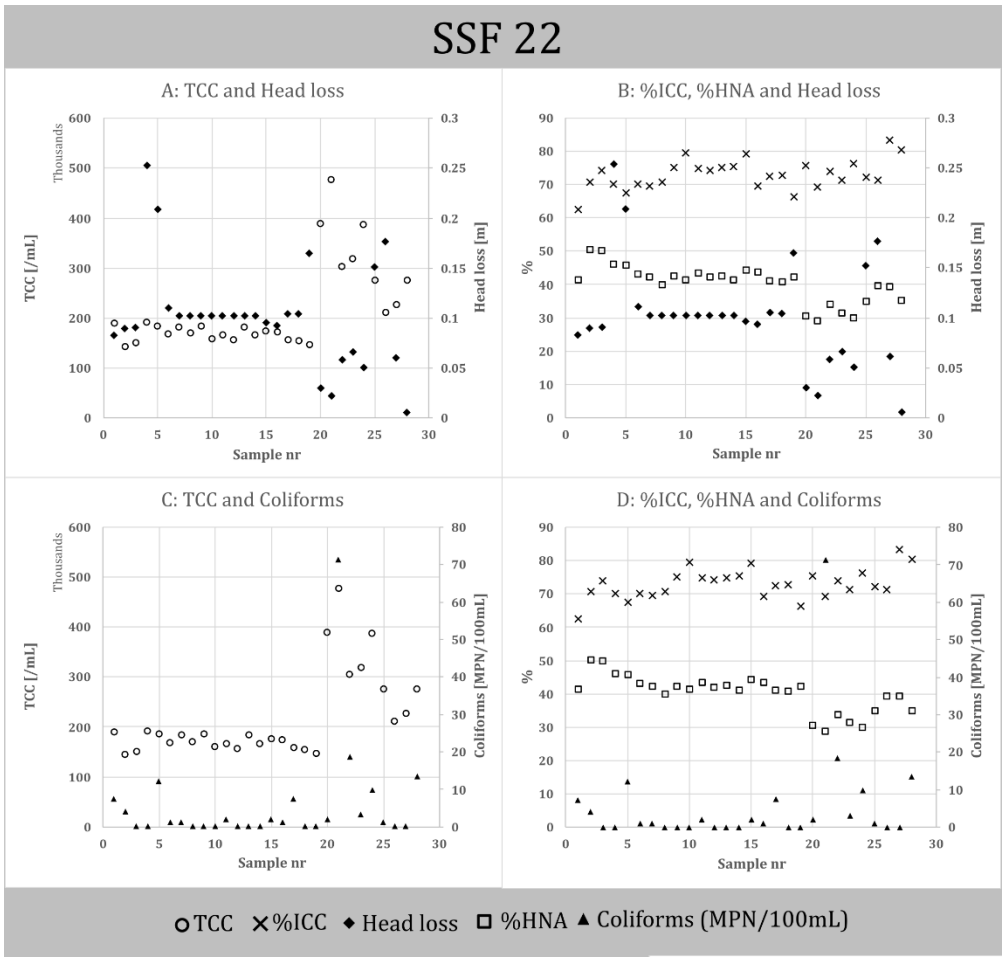


A2.3: Plots showing TCC, %ICC, head loss, %HNA and coliforms for filter 13 at Ringsjöverket.

A2.4: Plots showing TCC, %ICC, head loss, %HNA and coliforms for filter 20 at Ringsjöverket.

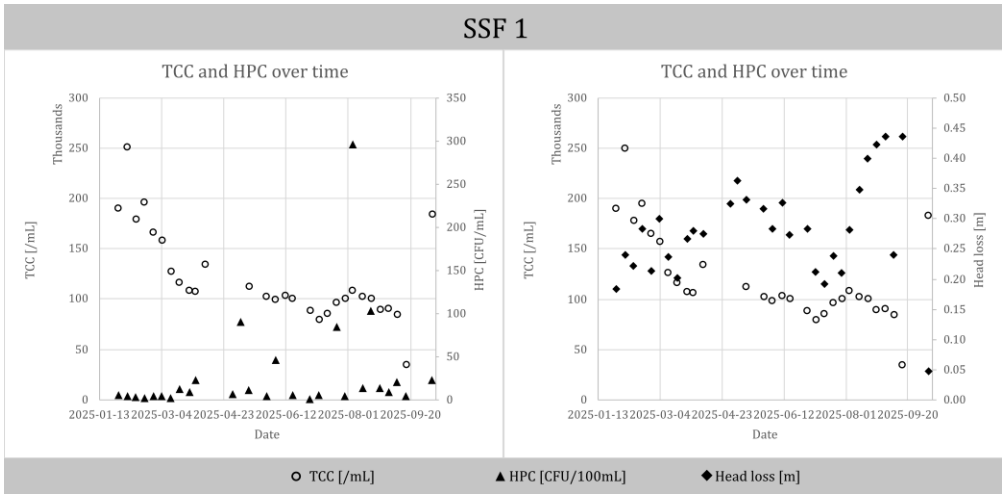


A2.5: Plots showing TCC, %ICC, head loss, %HNA and coliforms for filter 21 at Ringsjöverket.

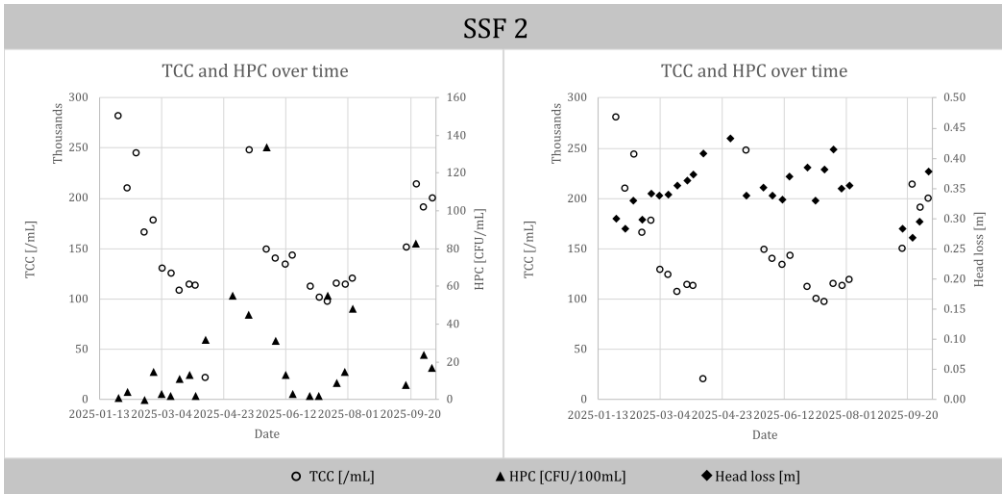


A2.6: Plots showing TCC, %ICC, head loss, %HNA and coliforms for filter 22 at Ringsjöverket.

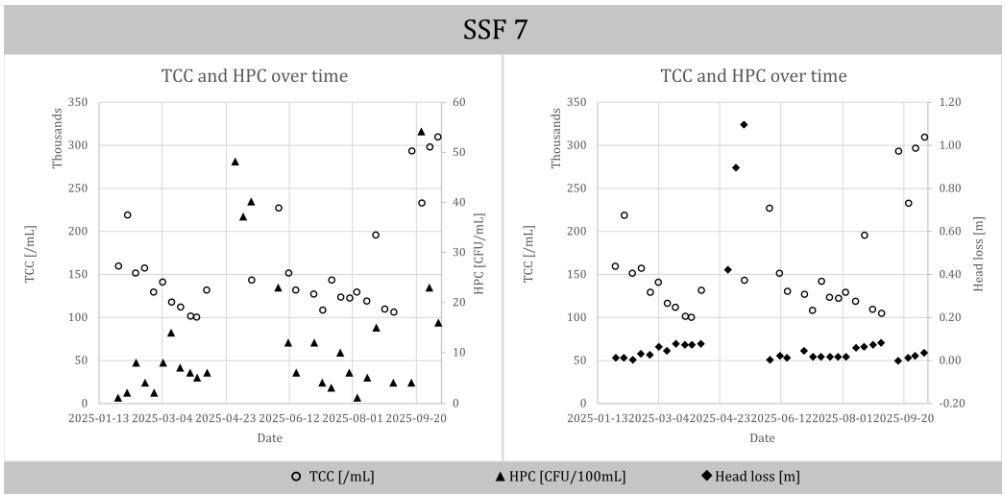
Appendix 3: In-depth plots for filters at Berggården



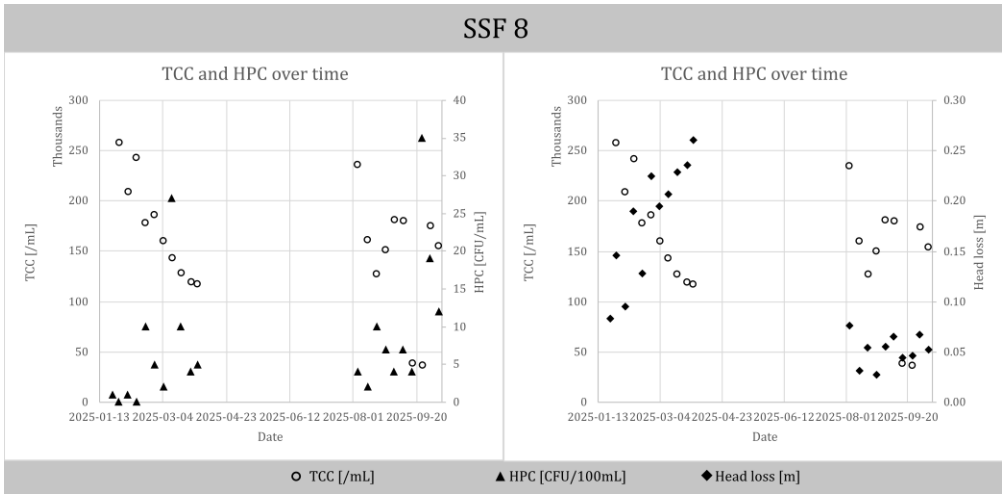
A3.1: Plots showing TCC, HPC and head loss for filter 1 at Berggpården.



A3.2: Plots showing TCC, HPC and head loss for filter 2 at Berggpården.

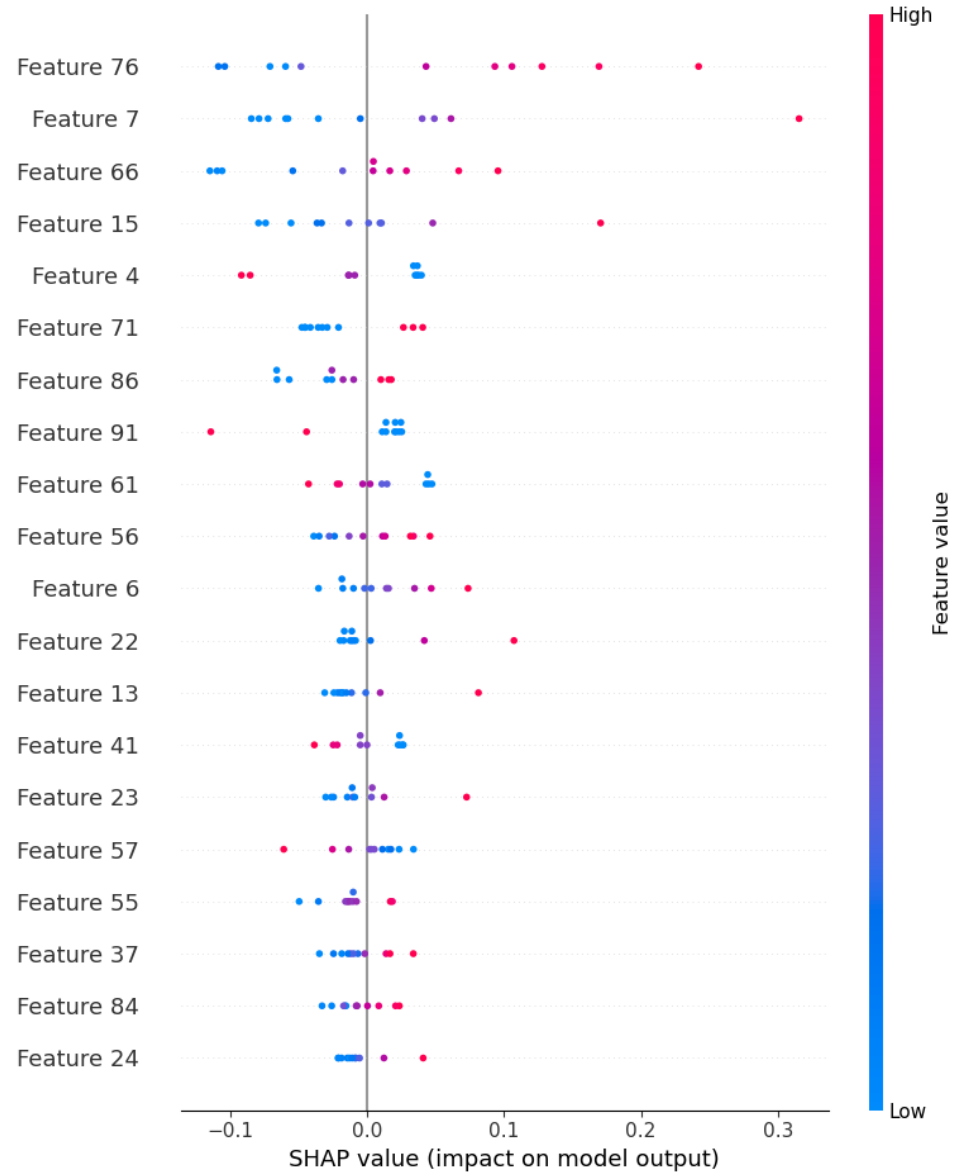


A3.3: Plots showing TCC, HPC and head loss for filter 7 at Berggpården.

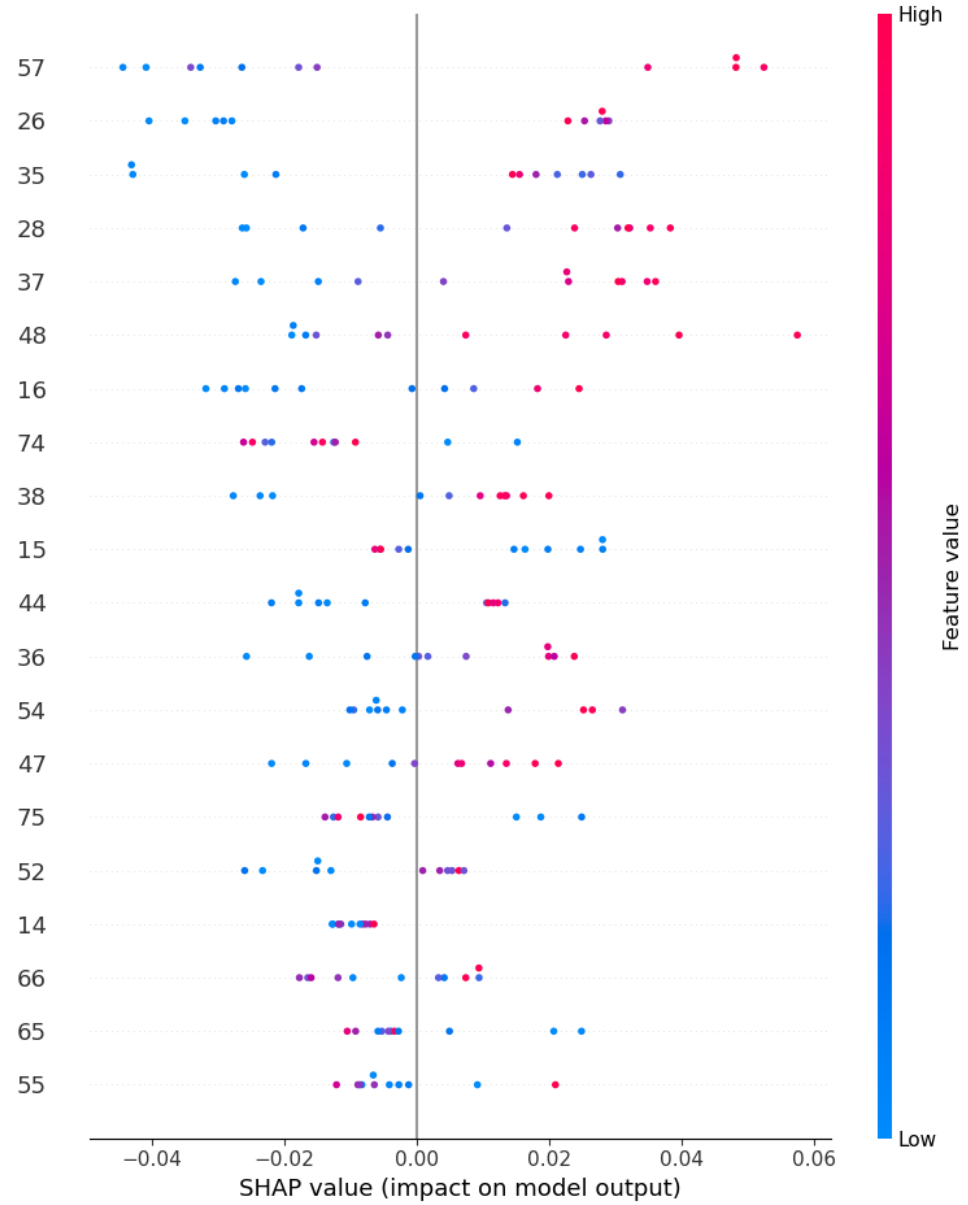


A3.4: Plots showing TCC, HPC and head loss for filter 8 at Berggpården.

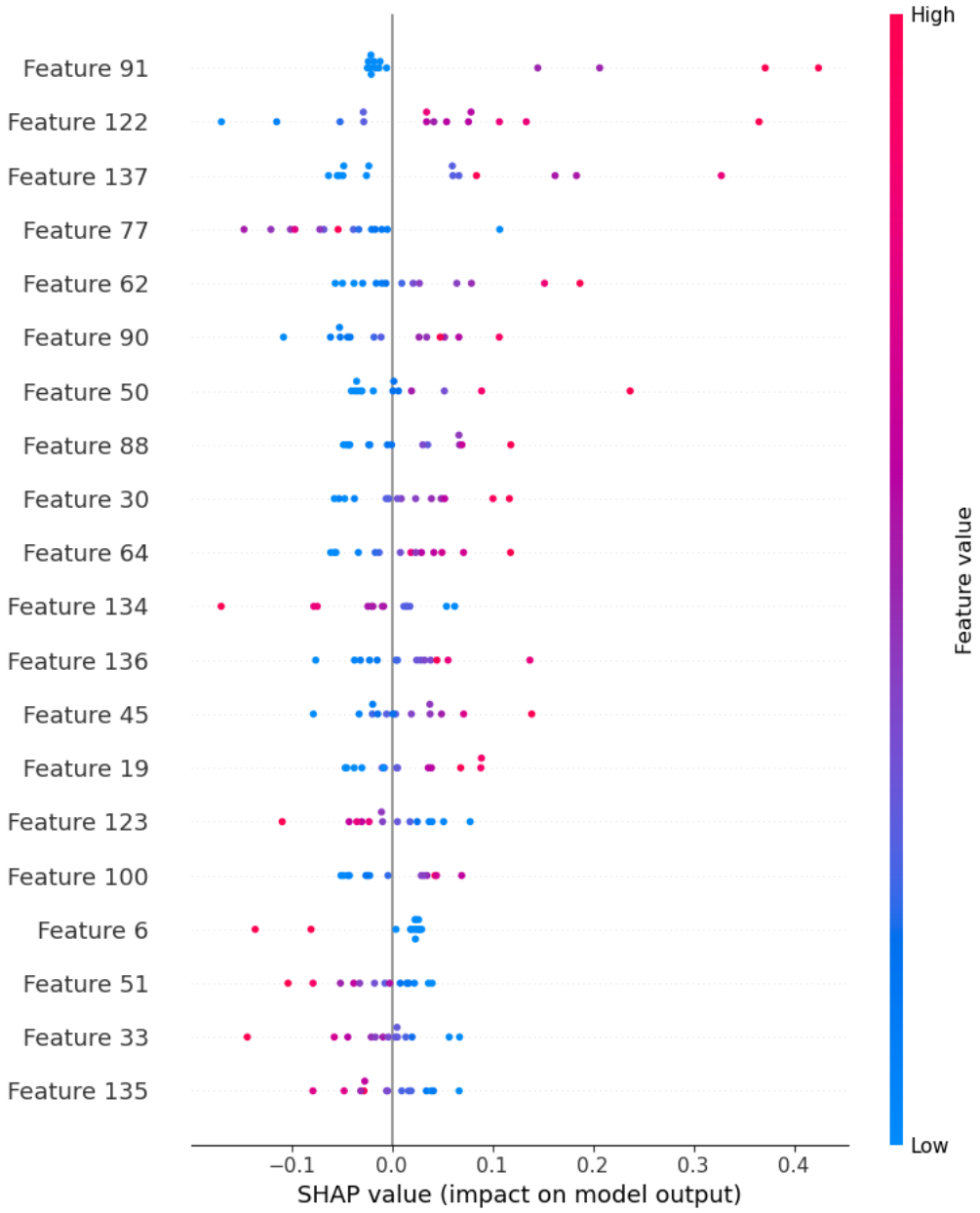
Appendix 4: SHAP plots



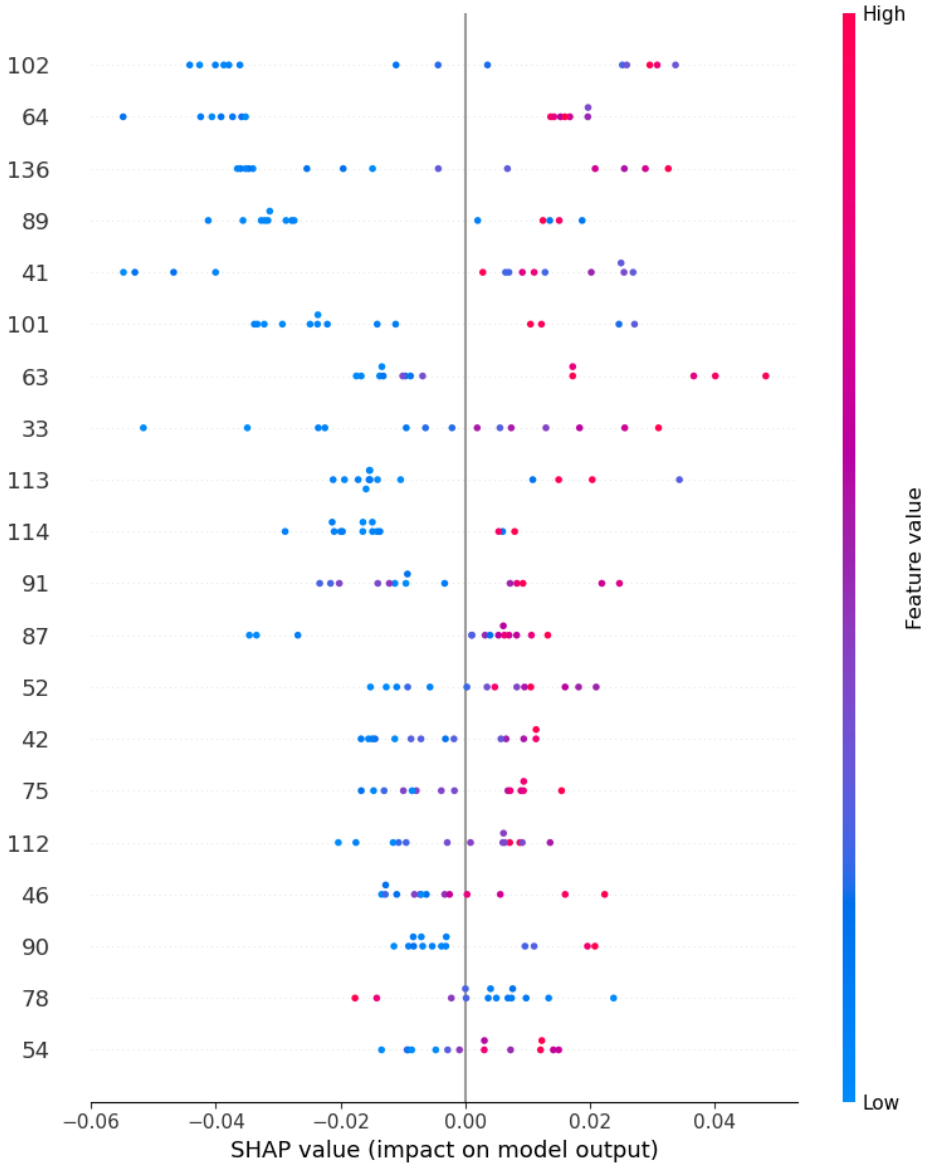
A4.1: SHAP for SVOA logistic regression.



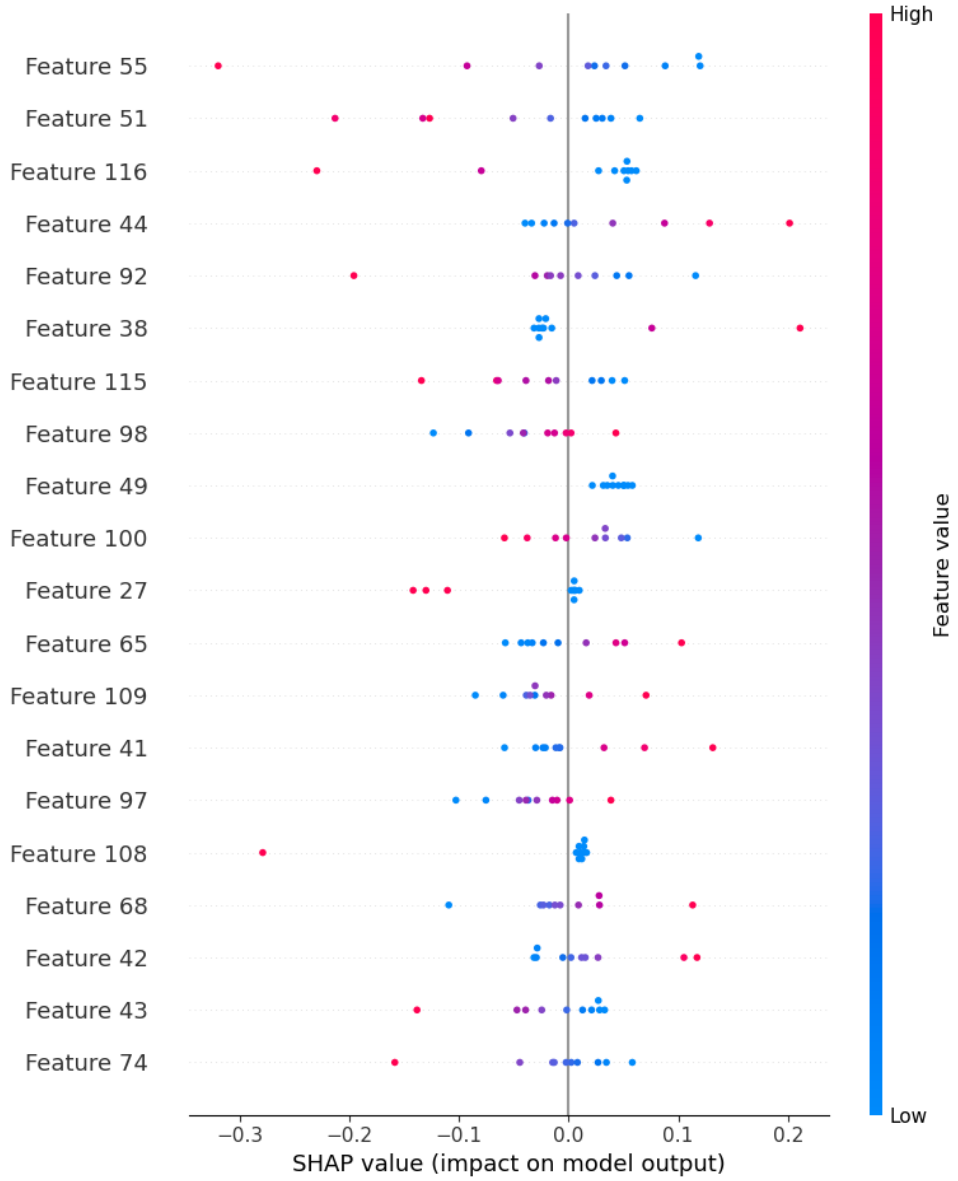
A4.2: SHAP for SVOA random forest.



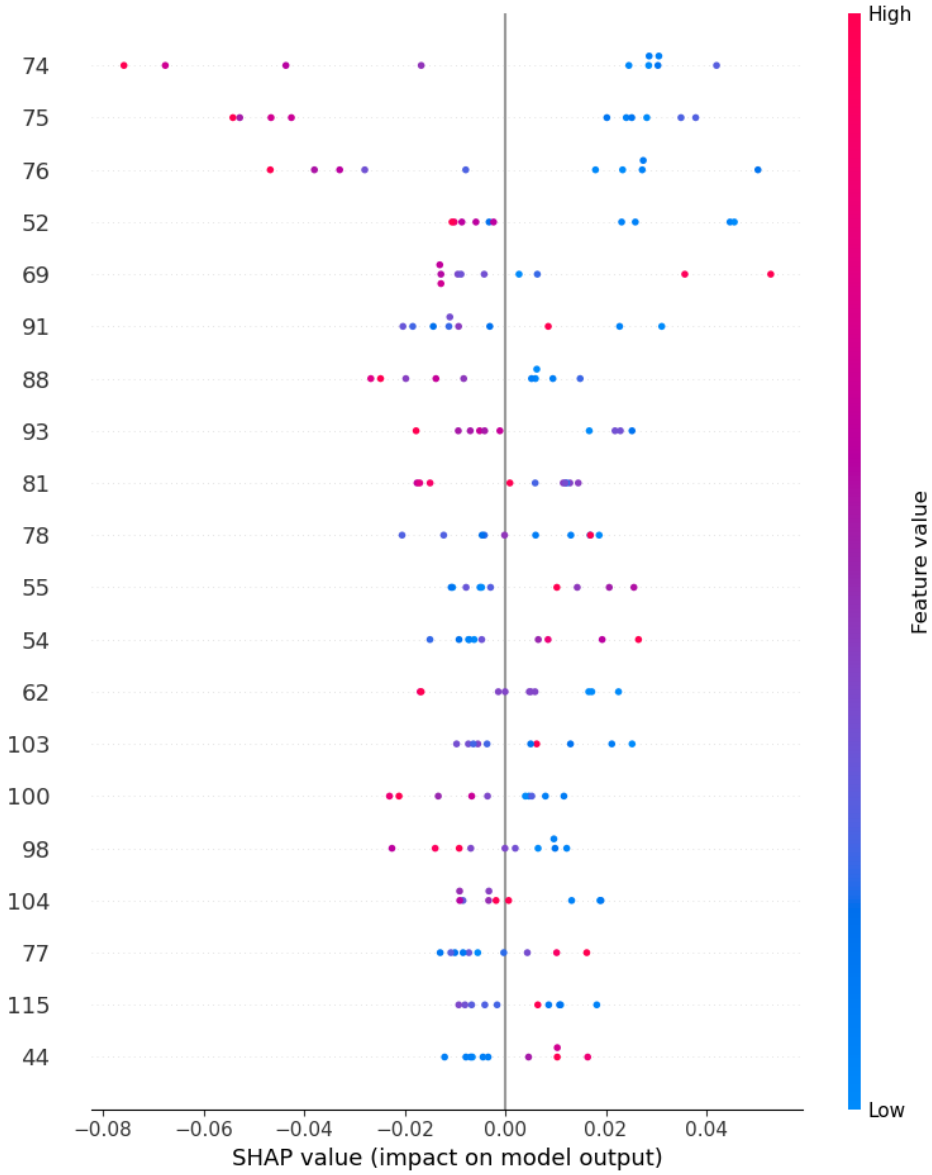
A4.3: SHAP for Sydvatten logistic regression.



A4.4: SHAP for Sydvatten random forest.



A4.5: SHAP for Tekniska verken logistic regression.



A4.6: SHAP for Tekniska verken random forest.