Predicting Sedimentation Rates in the Estuarine Botlek Harbour using Regression Machine Learning Algorithms

by

D.H.J. van Wijngaarden
4720199
October 2024

to obtain the degree of Master of Science
at Delft University of Technology

Graduation Committee:

Chair of Committee:	Prof. Dr. Ir. M. van Koningsveld, TU Delft
Committee member:	Dr. A. Kirichek, TU Delft
Committee member:	Dr. F. Baart, TU Delft
Committee member:	Dr. Ing. J.A.E. Antolínez, TU Delft
Supervisor:	E.B.J. Hupkes, Port of Rotterdam
Supervisor:	Ir. J.E. Vettorato, Port of Rotterdam

Preface

This report covers my master thesis on 'Predicting Sedimentation Rates in the Estuarine Botlek Harbour Using Regression Machine Learning Algorithms'. The thesis investigated the implementation of Machine Learning algorithms to predict sedimentation rates in the Botlek, a harbour in the Port of Rotterdam. This thesis project was carried out to complete the Master of Hydraulic Engineering at the Faculty of Civil Engineering and Geosciences at the Delft University of Technology. The project was commissioned by the Port of Rotterdam and the Ports & Waterways Department of the Civil Engineering faculty.

I want to express my gratitude to my graduation committee. Starting with Mark van Koningsveld as the chair and Alex Kirichek for providing their knowledge and experience on the hydraulic and operational processes in the Port of Rotterdam. I would also like to thank Fedor Baart and José Antolínez for their support and guidance in the Machine Learning field. The composition of the committee allowed me to explore a whole new side of engineering, which I truly enjoyed. Additionally, I would like to thank my Port of Rotterdam supervisors. Edwin Hupkes, thank you for always taking the time to answer quick questions and providing the information and context needed to conduct my research. Jeroen Vettorato, thank you for your expertise in data science and the clarifying sparring sessions on my modelling approach. Lastly, thank you to the dredging desk and Andre van Hasselt for walking me through the maintenance operations and providing me with the data.

Dirk van Wijngaarden
Rotterdam, October 2024

Abstract

Maintaining navigational channels and infrastructure in large commercial ports is costly and complex. Sedimentation poses a significant challenge in maintenance by reducing the depth of the channels and basins, thus threatening accessibility and navigational safety. Traditionally, port authorities rely on bathymetry surveys to guide their dredging operations, but this approach limits their ability to anticipate sediment accumulation and prevent operational obstructions or navigational hazards. This research aims to develop Machine Learning (ML) methods to predict Sedimentation Rates (SR) by analysing patterns in hydrological and meteorological (hydro-meteo) conditions and dredging data. By developing the SR prediction models, this research can contribute to efficient maintenance operations and improved navigational safety.

The study focuses on the Botlek, a harbour in the tide-dominated Port of Rotterdam (POR) in the Rhine-Meuse estuary. The Botlek is situated at the transitional border of fresh and saline water. The near-bed currents resulting from the density gradients between the riverine and saline water are the dominant factor that transport suspended sediment into the Botlek. Three data types are integrated to capture the dynamic interplay of saline and riverine factors within the harbour. These types are Multibeam bathymetry surveys, hydro-meteo variables (such as salinity, river discharge, and tidal variation), and dredging logs. The surveys provide the net sediment accumulation, which is assumed to result from a specific period of hydro-meteo conditions and dredging. The net accumulation values serve as label for the training samples.

The algorithms evaluated in this research are Linear Regression (LR), Random Forest Regression (RFR), and Support Vector Regression (SVR). The feature importance scores from the RFR and the accuracy on small datasets of the SVR were decisive in this selection. The algorithms require one-dimensional arrays as input. Therefore, the hydro-meteo time series are transposed into lagged features to allow the algorithms to recognise the time dependency of the data. Moreover, the samples must be of a uniform length. As the Botlek dredging areas with the highest SR are surveyed at an interval of 30 days and the less dynamic regions every 60 days, a selection in considered areas must be made. All 30-day areas and three 60-day areas are chosen. The three 60-day areas are included to prevent the sample set from being too small. The downside of this approach is that the SR values of the 60-day interval do not result from a period of 30 days, potentially leading to inconsistent results.

All algorithms are tested and refined over multiple development phases to determine their predictive accuracy across different data configurations. The development phases consist of different configurations of features, sample types, and dependent variables. The sample set also contains samples with a negative SR, indicating erosion. The port authorities state that large erosion values in the Botlek are unrealistic. Therefore, part of the different development phases is including or excluding the erosion samples from the dataset. The total number of samples used for model development is 181 with erosion and 142 without. To reduce the risk of overfitting due to the high dimensionality relative to the number of samples, the hydro-meteo time series are aggregated into daily, weekly, and monthly means.

First, a baseline performance is established by only selecting the dredging data as input. The unit of the dependent variable SR is set to m^3 and m^3/day to investigate if normalising SR results in a better performance. In the following phases, the hydro-meteo variables are included in the feature set in their daily, weekly, and monthly forms to create three sample types. Including multiple time scales allows for an investigation of the trade-off between preserving the temporal resolution and managing the size of the feature space. Moreover, the runs are performed over multiple random states to analyse the performance over different dataset splits. The last development phase will select the best-performing configurations of the preceding phases and tune the models to achieve optimal performance.

The POR states that the ML models do not have to be able to predict sediment volumes with a small confidence interval to be practical for the port authorities. Instead, the models should be able to accurately predict trends in sediment accumulation and provide actionable insights to anticipate operational obstructions or navigational hazards. The Asset Management Department mentions that $\pm 30\%$ is an acceptable error margin for the predicted SR as long as there are no outliers.

The model runs show that the ML algorithms can reasonably predict SR. The baseline configuration with only dredging data and SR in m^3/day results in a mean R^2 of 0.65 for LR and SVR and a mean $RMSE$ of 405.76 and 407.95, respectively. The RFR models are outperformed in this phase. The next phase (all samples and SR in m^3) shows a significant decrease in performance when adding the hydro-meteo variables to the feature set. Additionally, reducing the number of hydro-meteo features in this phase does not improve performance, as the runs with daily, weekly, and monthly sample types all showed varying performances. The results from normalising SR in the next phase are more promising but still have not reached the baseline performance. The models, particularly the SVR ones, capture the relations better, and decreasing the number of features allows for better performance as the monthly sample types provide the best mean performances for LR, RFR, and SVR. An essential takeaway is that all models consistently predict the negative SR values as positive, confirming that the erosion samples are unreliable.

The last development phase summarises the intermediate conclusions into the optimal configuration: no erosion samples and SR in m^3/day. With these configurations, the SVR models trained on the monthly samples outperform the baseline with a mean R^2 of 0.69 and a $RMSE$ of 388.78. Analysing the predictions shows that the SVR can generalise well, as the residuals of extreme SR values are of the same order as smaller SR, except for some outliers. The RFR and LR outperform their predecessors in the other phases but do not surpass the baseline performance. Consequently, the RFR feature importance scores from this phase are the most reliable. The dredged volume between surveys, salinity, discharge in the Nieuwe Maas, and the tidal variation have the highest scores. The scores agree with the previously described sediment transport process in the Botlek. These variables are selected for a reduced feature set that does not improve the absolute mean SVR performance but significantly increases the consistency over the different random states.

Across all phases, SVR outperforms LR and RFR, except for the baseline performance where SVR and LR are equal. Moreover, including the hydro-meteo variables along with the dredging data into the feature space improves the performance in the case of RFR and SVR, whereas LR struggles with the higher dimensionality. The monthly mean samples deliver the best performance, showing that the reduced number of features outweighs the added information of the daily and weekly samples.

The predictive accuracy of the models varied over the dredging areas, with the 30-day areas as the best performing. Separating the 30 and 60-day interval areas into two smaller but more specific datasets is the next step in achieving more reliable predictions. The initial assumption that this would lead to too small datasets is unjust, as the model performance stagnates well before all training samples are used. Furthermore, the hydro-meteo data is not area-specific, which forces an interpolation of the data over the Botlek, leading to inaccuracies while capturing local sedimentation dynamics.

Integrating hydro-meteo, dredging, and survey data to predict SR shows promising results, as a significant share of the predictions fall within the acceptable error margin of $\pm30\%$. Despite some limitations, the findings demonstrate that ML models can support sediment management in complex estuarine environments. However, additional development steps are needed to improve the reliability of the predictive models. Furthermore, research is needed to implement the ML models into maintenance operations. Developing a data pipeline will be crucial for port authorities to use the predicted SR values effectively and optimise dredging operations in the future.

Contents

Acronyms

2WERKH 2e Werkhaven

AI Articial Intellegence

AM Asset Management

ANN Artificial Neural Network

BOTCGO Botlek Centrale Geul Oost

BOTCGW Botlek Centrale Geul West

BOTM Botlek Mond

EDA Exploratory Data Analysis

ETM Estuarine Turbidity Maximum

GBR Gradient Boost Regression

hydro-meteo hydrological and meteorological

IL Identity Line

LR Linear Regression

LSTM Long Short-Term Memory

ML Machine Learning

NGD Nautical Guaranteed Depth

OSR Operationeel Stromingsmodel Rotterdam/Operational Flow Model Rotterdam

PCA Principal Component Analysis

POR Port of Rotterdam

RFR Random Forest Regression

RWS Rijkswaterstaat

SL Supervised Learning

SPM Suspended Particulate Matter

SR Sedimentation Rates

SVM Support Vector Machine

SVR Support Vector Regression

UL Unsupervised Learning

XGB eXtreme Gradient Boosting

1 Introduction

1.1. Research context

Ports serve as a transfer hub for global trade, facilitating the movement of goods between land and sea. With the growing demand for transportation, ports are increasingly competing with each other. An aspect of achieving a better competitive position is providing the highest quality infrastructure, which is why port authorities worldwide continuously explore maintenance strategies that can optimise their operations (Sepehri et al., 2024). Part of infrastructure maintenance is monitoring and preserving the depth of the navigational channels and basins affected by sedimentation. Sedimentation reduces the channel depths, threatening accessibility and navigational safety (Deltares, 2024). To counter this, maintenance dredging can be employed as part of the sediment management strategy.

Understanding the patterns and causes of sedimentation in a port helps the authorities make better-informed decisions, allowing them to anticipate sediment accumulation. However, sediment dynamics in coastal environments, such as estuaries, are complex and often not fully understood. The complexity in estuaries can be ascribed to dynamics caused by river flow, tidal movement, wind and wave forces, and the mixing of fresh and saltwater (Feng et al., 2023). Tidal movement forces saline water upstream, where it interacts with the fresh water. This interaction results in a fresh-saltwater interface that creates pressure gradients due to the density differences. These gradients drive a vertical flow circulation called the estuarine circulation (Bosboom and Stive, 2023). Ports in estuaries face these challenging conditions while trying to increase their system knowledge to ensure the safety of visiting vessels and optimise their operational efficiency.

The largest port in Europe is the tide-dominated Port of Rotterdam (POR) (Neumann et al., 2024), located in the Rhine-Meuse estuary. In this port, 12 to 15 million cubic meters of sediment are dredged annually (Kirichek et al., 2018). While necessary, dredging is costly and comes with a significant environmental impact. In fact, dredging is responsible for 46% of the 90 kilotons of CO2 emitted to maintain the POR infrastructure. Significantly more than the 19% for steel maintenance, the second largest emission (Port of Rotterdam, 2023). The POR authorities aim to be a front-runner in the energy transition to achieve the international climate goal of limiting global warming. Improving sediment management efficiency might reduce maintenance costs and contribute to this goal due to its large share of CO2 emissions.

Sediment management in the POR relies on bathymetry surveys of the channels. Surveying vessels equipped with a Multibeam Sonar provide a detailed and reliable image of the state of the waterbed, allowing the port authorities to make targeted decisions on dredging operations (Kirichek et al., 2018). However, a rapid indication of the state of the waterbed to plan a maintenance operation is missing. This delay in information occurs because a surveying vessel can only be employed when there is an available window in the operational schedule of the port. Without real-time data or insight into the future state of the waterbed, sedimentation may not be addressed until it has disrupted the port operations. Having (near) real-time data or sedimentation forecasts would enable more precise planning and avoid unnecessary surveying or dredging trips, enabling the port to prioritise critical areas and improve maintenance efficiency.

To address the need for accurate and timely data, POR has started the Innovative Sediment Management (PRISMA) program. This program aims to integrate data and research with experience to increase the knowledge of the hydraulic systems in the port (TKI Deltatechnologie, 2023). By collaborating with research institutes like Deltares and Delft University of Technology, PRISMA enhances the understanding of sediment dynamics and dredging operations to improve overall maintenance efficiency. The most recent phase, PRISMA III, consists of four work packages, one being *'Data science for more efficient dredging trips.'* This package analyses a large dredging trips dataset to investigate trends and correlations that can be used to improve operations in terms of timing and location. In addition to the trips dataset, the POR possesses years of information on the environmental conditions in the port region. These hydrological and meteorological (hydro-meteo) variables are either measured or predicted by the Operationeel Stromingsmodel Rotterdam/Operational Flow Model Rotterdam (OSR) (Svasek Hydraulics, 2024) and offer insight into the conditions that shape the estuary.

The PRISMA effort is part of a worldwide movement towards intelligent port management. As Sepehri et al. (2024) highlights, there has been an increase in research on sediment and smart port management with a total of 128 articles published since 2001, of which 9% is aimed at predictive maintenance improvement. This trend reflects the need for innovative methods to enhance decision-making processes and optimise operations. One interpretation of predictive maintenance would be a model that can predict Sedimentation Rates (SR) to allow ports to anticipate sediment buildup, avoid disruptions, and efficiently deploy their resources. Consequently, the development of such a model would contribute to answering the overall need for innovation.

1.2. Research problem

Sedimentation processes in ports are determined by the hydrological (Y. Guo, 2022) and meteorological (Gonzalez Rodriguez et al., 2023) conditions, along with human interference like dredging operations. Dredging removes sediment from a system, and the hydro-meteo variables drive the dynamics and processes in the channels and basins. However, understanding and predicting how these factors influence the sediment dynamics in a complex estuarine environment is difficult for port authorities. Configuring these factors into models that can accurately predict SR would allow ports to anticipate sediment buildup and improve their maintenance operations. This approach would be based on the assumption that the hydro-meteo conditions and the dredging operations within a certain time frame result in a predictable SR. The challenge of this approach is finding a method that can determine the complex relationships between the variables and the SR and process these into practical insights for port maintenance. Machine Learning (ML) could present a promising solution to this challenge. ML is a form of Articial Intellegence (AI) that can iteratively learn from data to discover patterns and relations without the need for explicit programming (França et al., 2021).

This research investigates how the benefits of ML can be leveraged to develop models that learn the relations between the mentioned hydro-meteo and dredging data and the sedimentation in an estuarine harbour. Several ML algorithms and configurations of the different data sources will be evaluated to determine which algorithms can accurately predict SR to provide port authorities with reliable insights into sedimentation processes.

1.3. Research gap

Given the size of the POR, it is wise to limit the research area. There are many areas in the 12,500 acres of the POR (Port of Rotterdam, 2022) for which SR could be predicted. Of these areas, the Botlek harbour is of particular interest. The harbour is situated at the transitional zone of saline and fresh water, thus experiencing the effects of the previously mentioned fresh-saltwater interface (de Nijs, 2012). Moreover, a study by Bruijn (2018) showed that the SR in the Botlek are among the highest in the entire port. Developing a method to increase the dredging efficiency in this area could, therefore, have a significant positive effect on operations, system knowledge, and navigational safety and reduce the environmental impact of dredging. The POR divides its assets into dredging areas to maintain the channels and basins. The Botlek dredging areas are shown in Figure 1.1.

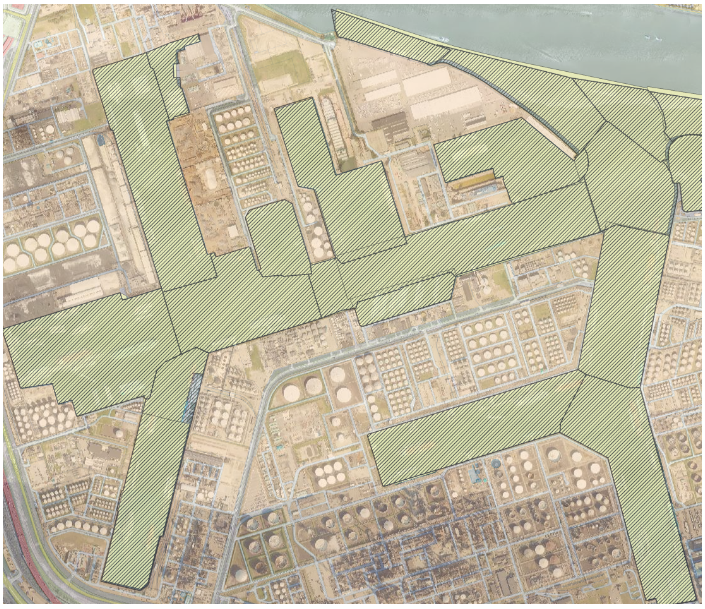


Figure 1.1: Dredging compartments of the Botlek Harbour (Port of Rotterdam, 2024)

Aside from the PRISMA program, other efforts have investigated the hydrodynamic processes for the POR or a specific area of the POR. These studies provide valuable knowledge and highlight critical sediment transport and accumulation factors. In recent years, sedimentation and the dynamics in the Botlek have been studied as a dissertation by de Nijs (2012) and as a thesis project by El Hamdi (2012) and Tempel (2019). These studies aimed to clarify part of the dynamics within the harbour. The transport and sedimentation processes in the Botlek are described by de Nijs (2012) using 3D hydrostatic models and tracer analyses. The results showed that the previously mentioned density gradients are the dominant factor in transporting suspended sediment in the Botlek. El Hamdi (2012) focused on how siltation in the Botlek can be reduced, while Tempel (2019) described the functioning of sediment traps in the Botlek.

In terms of applying ML, Goldstein et al. (2019) reviewed over 60 papers that applied ML for predicting coastal morphodynamics, sediment transport, and coastal morphology. They stated that ML, unlike traditional statistical methods, does not necessarily require assumptions about the relationships within the data. Instead, the algorithms can automatically search for patterns, making them effective in handling complex and high-dimensional datasets and, in many cases, successful when implemented for coastal research. Rajaee and Jafari (2020) showed that implementing AI for modelling sediment concentrations in rivers has made much progress in the past decades, and some methods can replace conventional but time-consuming mathematical techniques. A more recent example is a research project by Latif et al. (2023) on the Johor River in Malaysia, where the capabilities of different ML techniques were successfully compared to predict suspended sediment load. Another recent effort is the quantitative forecasting of bed sediment loads in river engineering by Fuladipanah et al. (2024). They used river characteristics such as flow discharge and depth to predict sediment loads in multiple rivers. However promising and successful, previous research projects primarily focused on sediment prediction in upstream sections of rivers or coastal regions. Therefore, the dynamic interactions within estuaries are often left out of the scope.

From the promising ML studies, some have been applied to estuarine environments. A conceptual prediction of harbour sedimentation using AI was made on a basin in Ezbet Elborg in Egypt in 2021. The researchers used several methods to find a relation between varying breakwater parameters and sedimentation volumes in the basin under the influence of waves and tidal movements (Elnabwy et al., 2022). Another study applied ML methods to predict satellite-captured morphological variation in the Da Dien estuary in Vietnam, specifically the variation in throat width of the river mouth (Pham et al., 2019). Both these projects had a location at the shoreline as a focus area. In these areas, the water density predominately was close to that of seawater, therefore neglecting the gradients that follow from the mixing in an estuary. A research project from 2023 assumed these gradients as a vital factor in their scope. The study stated that the vertical mixing of the water could lead to complex turbulent flows. Therefore, salinity was one of the input parameters for their turbidity prediction model (Kim et al., 2023). The research area was the Guem River Estuary in South Korea, a semi-closed estuary because of the sea dike in the Guem River and not an open estuary like the Port of Rotterdam. This means that the mixing of fresh and saline water is no longer a natural process because it depends on artificial discharges from the dike.

The mentioned projects all have elements relevant to estuaries like the Rhine-Meuse estuary but, logically, are not an exact fit. There is no global relationship for estuaries; for each estuary, a specific relationship must be developed (Hinwood and McLean, 2018). Furthermore, studies that include hydro-meteo variables, dredging data, or both into the input space are limited. From the studies that do, most do not consider the natural mixing of saline and riverine water. By developing models that integrate these factors to predict SR in the Botlek, this thesis addresses this gap in research and increases the system knowledge of the Rhine-Meuse estuary.

1.4. Research aim

This research seeks to assist the POR in improving monitoring and maintenance efficiency and contribute to the field of predictive sediment modelling by achieving two objectives. The first is the development of ML models that can predict SR in the Botlek area based on the hydro-meteo variables. The second is to accurately describe and analyse the findings and limitations encountered during the development of these models to provide recommendations and tools for port authorities and future researchers on improving sediment management practices, including potential changes in monitoring and surveying approaches.

The research will address the following research question:

To what extent can machine learning methods be utilised in predicting sedimentation rates in an estuarine harbour, considering the dynamic interplay of marine and riverine influences?

The research follows the sub-questions described below. This order is followed to structurally describe the process and requirements needed to develop a predictive model for SR in an estuary:

- *What data is available and relevant for predicting sedimentation rates using machine learning in estuarine harbours?*
- *Which machine learning algorithms are most suitable for predicting sedimentation rates?*
- *How can the selected machine learning algorithms and features be configured to predict sedimentation rates?*
- *How do the selected machine learning algorithms perform across different configurations?*
- *What factors could enhance the performance of machine learning models in predicting sedimentation rates?*
- *What practical insights and implementations can be gained from the model results regarding sediment behaviour and maintenance efficiency in an estuarine harbour?*

1.5. Thesis outline

The literature review in Chapter 2 addresses the general principles behind ML, relevant studies regarding implementing ML for sediment behaviour prediction, and discusses research regarding the hydro and morphological processes in the Botlek.

Chapter 2 provides the necessary foundation for Chapter 3 'Methodology'. This chapter starts with a case study, describing the Botlek area and the maintenance strategy of the port authorities. Once the case is established, Section 3.2 'Available data' and Section 3.3 'Data analysis' will answer the first research question and provide an overview of all data used for this study.

The data characteristics, combined with the findings of Section 2.2, provide a basis to select the most suitable ML algorithms for predicting SR in this research. Section 3.4 'Algorithm Selection' covers the motivation behind this selection.

After selecting the ML algorithms, the method to capture the trends and correlations between the hydro-meteo variables and the SR is constructed. Section 3.5 'Data Processing' describes how the data is engineered to fit the ML models. Section 3.6 'Model engineering' states how the process behind the hyperparameter tuning and feature engineering. Lastly, Section 3.6.3 'Modelling phases' summarises how the many possible data and ML configurations are divided into model development phases. These three sections answer the third sub-question.

Chapter 4 'Results' will cover the results of the model development phase and answer the fourth sub-question. Section **??** shows the effect of the iterative development steps and displays how well the models performed. This section will serve as input for the last sub-questions in Chapter 5 'Discussion'. Section 5.1 'Result analysis' in this chapter contextualises the results and analyses the underlying meaning and practical implications. Next, in Section 5.2 'Limitations', the limitations and restrictions of the models are discussed. These two sections form the basis of answering the fifth sub-question in Section 5.3 'Model improvement factors'. The discussion chapter ends with the practical implications and usefulness of the developed models in Section 5.4 'Practical recommendations and implementations'.

After covering all sub-questions, the main question is answered in Chapter 6 'Conclusion'. The report is finalised by including the recommendations for further development and implementation of the results in Chapter 7 'Recommendations'.

2 Literature review

The field of sediment behaviour has been extensively studied due to the significant impact that erosion and sedimentation have on waterways and coastal areas. Traditionally, the behaviour is modelled and predicted by numerical models based on experience from the authorities or organisations responsible for maintaining the waterways. Efforts to capture sediment behaviour with alternative methods have increased simultaneously with the rising popularity of AI. ML, a powerful aspect of AI, is particularly well-suited for capturing the complex and nonlinear relations between environmental factors within rivers and estuaries.

This literature review discusses the relevant and state-of-the-art literature regarding the use of ML for sedimentation prediction. As the case area for this master thesis is the Botlek, an area of the POR which lies in an open estuary called the Rhine-Meuse Estuary, the focus of the chapter is reviewing research and modelling efforts related to open estuaries. A gap in the literature is expected due to the specific search objective. Therefore, research regarding ML modelling in rivers and coastlines will be an essential source of information. Another aspect is finding literature that discusses the influence of hydrological factors on sedimentation within an estuary, including the effects of salinity. This is because the Botlek lies in an area affected by the fresh-saltwater interface.

The relevant literature has been searched for on Google Scholar, ScienceDirect, Taylor & Francis, and Springer.

2.1. Machine Learning

2.1.1. General principles

The general principles of ML are discussed in this section. These principles are necessary to describe the preceding literature effectively. MLis a form of AI that uses statistical methods to train models to make predictions or classifications based on relations within data. Hyperparameters control the process in which the model is trained (IBM, 2024b). The ML methods that are relevant to this study can be divided into two main archetypes:

1. Supervised Learning: Supervised Learning (SL) uses labelled datasets to train algorithms to recognise patterns and make predictions. It is called supervised because the labelled datasets supervise algorithms into making accurate predictions or classifications (IBM, 2024b), and the labelling relies on human input. SL is divided into two categories (Google Cloud, 2024):

- Classification: These algorithms group data by predicting an output variable or label based on the input data.
- Regression: These algorithms predict an output value by detecting a relationship between variables in the input data.

The predictive abilities of the regression techniques are essential for this research. For this reason, this literature review will primarily focus on supervised models involving regression. This does not mean that other methods are not suitable for predicting SR.

2. Unsupervised Learning: Unsupervised Learning (UL) techniques are algorithms that learn from data without labels or defined predictions, which is very helpful in discovering underlying unknown relations. However, the accuracy of the results from UL techniques is difficult to determine since the output is not based on labelled data (Rajoub, 2020). This makes the output variables difficult to interpret. Therefore, UL is not necessarily applicable to this research, as the goal is to get an interpretable prediction based on historical data on environmental factors.

Data normalisation

Data normalisation is an essential part of model development when the input variables have varying scales because it transforms the range of the variables to a standard scale. It enhances the performance and improves the accuracy of ML models by preventing the domination of larger-scale variables during the learning process. Two common techniques are Min-Max and Z-score normalisation (D. Singh and B. Singh, 2020).

Ensemble learning

Ensembled Learning combines several weak learning models into a strong learning one. There are three main techniques for this (Mohammed and Kora, 2023):

- Stacking: Combines various estimators to reduce their biases. The estimator predictions are stacked and used as input for the final estimator.
- Bagging: Selects random data samples after which the weak learners are trained independently but simultaneously. The average of the predictions gives the final, more accurate result.
- Boosting: Similar to bagging but trains the models sequentially. Each model tries to improve the error of its predecessor.

2.1.2. Model evaluation

Supervised ML methods rely on labelled data to ‘learn’. How well a model will learn depends on the quality and quantity of that data. It is common practice to split a dataset in the development to be able to evaluate the performance of the model. The data is often split into three sets: training, validation, and testing (Murphy, 2022). The first is used to fit the model. The second set provides an unbiased evaluation of the fit of the training while tuning the hyperparameters. Lastly, the test set is used to unbiasedly evaluate the final fit of the model (Brownlee, 2020b). This section describes the theory behind model fitting on these datasets.

Regression

Regression is used to determine the relation between input variables and a dependent outcome variable and is often applied to make predictions within ML (Kadam et al., 2020). The simplest form of regression assumes a linear relationship between a dependent variable and independent variable(s). In many cases, including this thesis, the relationship between the variables is nonlinear, causing a possible bad fit. Many types of nonlinear regression algorithms can overcome this problem. An example is shown below. Here Y is the dependent variable; β_n the slopes; β_0 the intercept; X_n the independent variables; and ϵ the random error (Chang and Hsu, 2006):

$$Polynomial: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_p^n + \epsilon \tag{2.1}$$

Section 2.2 will discuss more complex and advanced regression methods applied in previous research.

Performance metrics

The performance evaluation after training is done by using performance metrics. For the previously mentioned regression models, the output is a continuous variable. Therefore, the metrics evaluate the difference between the prediction and the so-called ground truth. The most common metrics are displayed below with y_j the ground truth;$\check{y}_j$ the prediction; N the number of predictions; and SE the squared error (Belyadi and Haghighat, 2021):

$$Mean\ Squared\ Error\ (MSE) = \frac{1}{N}\sum_{j=1}^{N}(y_j - \check{y}_j)^2 \tag{2.2}$$

$$Mean\ Absolute\ Error\ (MAE) = \frac{1}{N}\sum_{j=1}^{N}|y_j - \check{y}_j| \tag{2.3}$$

$$Root\ Mean\ Squared\ Error\ (RMSE) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \check{y}_j)^2} \tag{2.4}$$

$$R^2\ Coefficient\ of\ Determination = 1 - \frac{SE(\text{line})}{SE(\check{Y})} \tag{2.5}$$

Loss and cost function

Loss and cost functions are the same functions as seen above, but there is a difference in application. The functions are used during training, whereas performance metrics are used afterwards. The loss function refers to the error of a single prediction, and the cost function refers to the average of the loss functions of the entire training set (Nadeem, 2022). The goal of training a model is to minimise these functions to achieve better predictions.

Optimisation algorithms

The optimisation problem is finding the set of inputs to a selected cost function that results in minimum offset. It is one of the big challenges in many ML algorithms. Optimisation algorithms are differential and non-differential (Brownlee, 2021). The most common algorithms in regression problems are differential, the most important being the Gradient Descent Algorithm. When a problem has a multivariable function, the gradient is a vector consisting of all the partial derivatives. This would look like (Phillips, 2023):

$$w_{j+1} = w_j + \alpha \nabla f(w_j) \tag{2.6}$$

Here w_{j+1} indicates the updated weight. The last two terms represent the vector of partial derivatives of the cost function $f(w_j)$, and α is the learning rate. The learning rate is the rate at which an algorithm updates the values of an estimate.

2.2. Machine Learning for sediment prediction

This section will discuss how ML methods are implemented for sediment prediction. There is a distinction in literature because of the difference in input parameters between estuaries and rivers. Most rivers in the relevant literature are unaffected by tides, changes in salinity or waves, which causes the rivers to have different conditions. There is a smaller availability of ML research related to sedimentation in estuaries than in rivers. However, the application on estuaries is the focus of this thesis and will therefore be leading in selecting which methods are relevant. The introduction briefly mentioned three research projects that considered estuaries. There is a critical distinguishment between the output of the models in these projects. The Da Dien estuary project predicted whether the river mouth extended or narrowed, a binary classification (Pham et al., 2019). In the Geum River Estuary, settling and resuspension characteristics in response to tidal modulations were predicted (Kim et al., 2023). Lastly, in the project in Egypt, Elnabwy et al. (2022) modelled sedimentation quantities based on changes in the harbour layout. The desired outputs of all three projects deviate from what is aimed for in this thesis, which is SR. However, the methodologies and models used in these projects are still relevant as inspiration for setting up the models for the POR.

2.2.1. Artificial Neural Network

An Artificial Neural Network (ANN) is designed after the neurons in a human brain. The networks contain neurons called nodes. These are sorted into input, hidden, and output layers. The nodes are connected to the layers, and every node has an assigned weight (Dongare et al., 2012). The weights control the signal between nodes and determine how much influence the input will have on the output. An additional bias ensures the activation of the neurons. Figure 2.1 shows how the Da Dien researcher structured their ANN and visualises how the weights connect the layers.

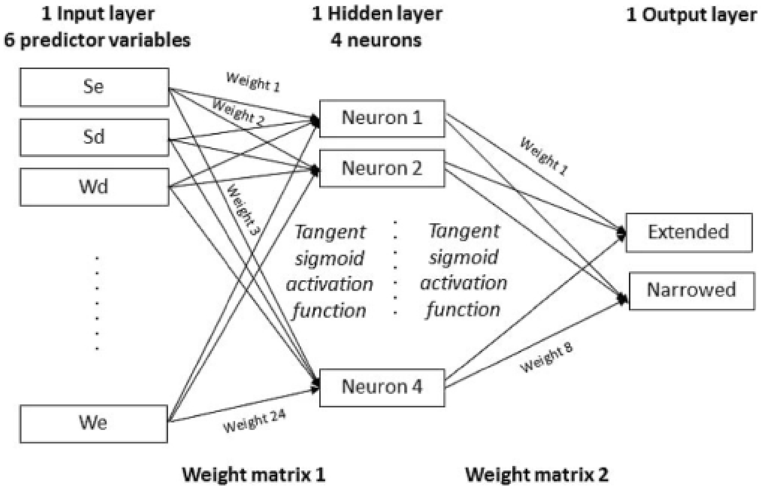


Figure 2.1: ANN for morphological classification (Pham et al., 2019)

Figure 2.1 shows additional important takeaways. The first is the binary classification in the output layer instead of a continuous variable. Secondly, the activation functions are visible in the centre of the figure. Activation functions are crucial in ANN as they introduce non-linearity, thus enabling a network to learn complex relations. The functions transform the input signal of a node into an output signal passed to the next layer (Sharma et al., 2020). For binary classification, the sigmoid function is often used as activation as this function turns the output into a probability between zero and one. Here, P is the probability of class 1, often called the positive class. 1-P is the probability of being class 0, the negative class. The class is classified as positive when P > 0.5 (Matveichev, 2023).

Figure 2.2 shows how Elnabwy et al. (2022) constructed the ANN structure when using regression to predict a singular continuous variable instead of a classification. The model has two hidden layers, and the input layer consists of harbour layout and environmental parameters.

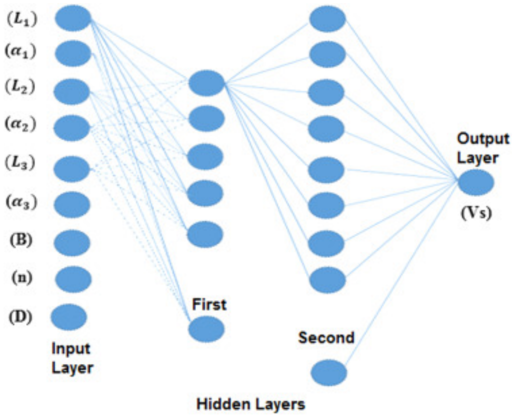


Figure 2.2: ANN configuration for a continuous dependent variable (Elnabwy et al., 2022)

Elnabwy et al. conveniently represent the output of Figure 2.2 as eq. 2.7, with Y the output; f the activation function; w_i the weight; x_i the input; and b the bias:

$$Y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.7}$$

Pham et al. (2019) converted their wind, swell, and tidal data into independent energy variables like swell or wind energy, therefore reducing the number of input variables as seen in the input layer in Figure 2.1. This reduction is often favourable when training an ANN. In the Da Dien case, classifying morphological changes of the river mouth was the aim, which explains the conversion to input parameters that heavily impact the bathymetry of the coastline. Kim et al. (2023) intended to predict turbidity distributions, which required the same strategy but with different input parameters. For example, temperature and salinity were expected to strongly influence the turbidity due to the pressure gradients that induce vertical mixing in the water column. They combined the two factors into a dimensionless input parameter called σ_t that represents a density, thus including both but decreasing the complexity of the model. Combining several factors into one parameter is an essential takeaway of these studies as it can be relevant in setting up the ML model for the POR. Lastly, Pham et al. (2019) and Kim et al. (2023) used Min-Max to normalise the input parameters, whereas Elnabwy et al. (2022) used Z-score standardisation.

Ren et al. (2023) used a Principal Component Analysis (PCA) to further develop a Long Short-Term Memory (LSTM) to predict surface suspended sediment concentrations after a typhoon in the Yangtze estuary. An LSTM is a more complex neural network that excels in determining long-term dependencies and relations in data. The PCA method is especially relevant as PCA can identify the main features in datasets. In this case, the datasets consisted of twelve meteorological and hydrological features similar to the ones available for the Botlek. However, PCA reduces interpretability because the principal components are uncorrelated combinations of the original independent variables. Using PCA, unfortunately, means that information is lost to reduce dimensionality.

2.2.2. Gradient Boost Regression and eXtreme Gradient Boosting

Boosting can be applied for classification and regression. Pham et al. (2019) use two types for their classification: AdaBoost and LogitBoost. AdaBoost assigns a higher weight to misclassified samples in each iteration, allowing the weaker learners to focus on those samples. LogitBoost adapts to AdaBoost and uses logistic models to predict a classification. It is generally not used for regression problems. Instead of classification, Sharafati et al. (2020) apply AdaBoost Regression (ABR), Gradient Boost Regression (GBR), and Random Forest Regression (RFR) to predict suspended sediment loads in the Mississippi River. They found that the ensemble ML models could make suspended sediment load predictions for the Mississippi. All models achieved an R^2 score close to 1.0. The RFR model slightly outperformed with a MAE of 1575.734 against 2195.211 for GBR and 2000.133 for ABR. RFR is described in a separate section.

GBR assumes that the best possible model in the sequence minimises the prediction error when combined with previous models. It is called gradient boosting because the target of the next model is based on the gradient of the prediction error from its predecessor. GBR can be described as the formula below where $\psi(y, F(x))$ represents the loss function, F(x) the prediction function, and δ the gradient of the residual (Sharafati et al., 2020):

$$F_0(x) = \arg\min_{\delta} \sum_{i=1}^{k} \psi(y_i, \delta) \tag{2.8}$$

GBR uses decision trees in its boosting process. Decision trees iteratively ask questions to reach a prediction and can be used for classification and regression but are prone to overfitting. GBR eliminates this issue by assembling multiple trees (Gaurav, 2022). This method is visualised in Figure 2.3.

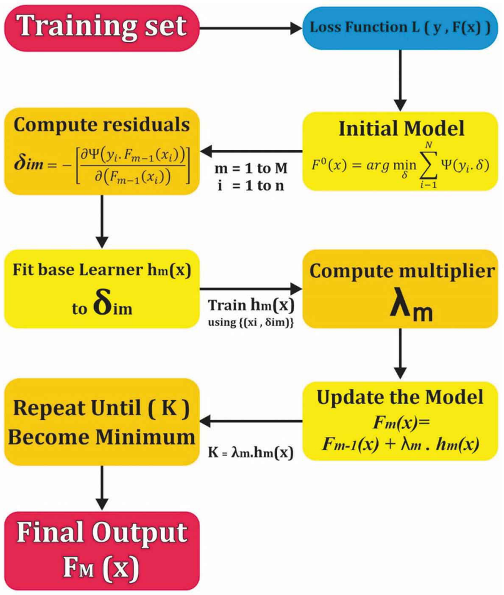


Figure 2.3: Conceptual Gradient Boosting Regression (Sharafati et al., 2020)

Another implementation of GBR is eXtreme Gradient Boosting (XGB). XGB uses ridge and lasso regularisation to penalise overfitting, whereas GBR only minimises the loss function. Additionally, XGB uses parallelisation during training, which results in a faster process (Belyadi and Haghighat, 2021). A study by Piraei et al. (2023) applied XGB to predict total sediment loads and compared its performance to other ML methods like ANN, ABR, and RFR. The comparison was based on six performance metrics and showed that the XGB model slightly outperformed the other techniques. XGB had an R^2 of 0.95, where ANN and RFR scored 0.87 and 0.89. The XGB algorithm is a viable option for predicting SR in the Botlek because of its high accuracy and ability to capture nonlinear relations. A limitation of XGB is the complexity of the algorithm due to the high number of hyperparameters. Moreover, XGB is computationally expensive (Dhumne, 2023).

2.2.3. Random Forest Regression

The RFR algorithm uses decision trees like the boosting algorithms described in the previous section. The difference lies in the architecture, as RFR is a bagging model that generates predictions in parallel. RFR constructs each tree with a different bootstrap data sample, after which the average of all predictions is used as the final prediction (Al-Mukhtar, 2019). Some studies on sediment concentrations or river loads show better RFR performance than several other ML methods. An example is the Mississippi River study, which showed a slight superiority over AdaBoost and GBR. Al-Mukhtar (2019) used RFR, ANN, and Support Vector Machine (SVM) to model suspended sediment in the Tigris River and demonstrated that the RFR model had superior prediction capability, with R^2 and $RMSE$ values of 0.8 and 130.71, respectively, compared to 0.67 and 194.02 for SVM, and 0.68 and 178.3 for ANN. However, the application of RFR in both these studies deviates from this thesis because the input parameters are only discharge values and suspended sediment concentrations at different points in time.Walsh et al. (2017) used RFR in a more applicable context. They used an RF approach to predict the spatial distribution of sediment pollution in an estuary. The researchers stated that modelling in estuarine systems requires capturing transport and fate dynamics, thus including sediment composition and bathymetry. Their RF model made predictions that generally agreed with the independent data and published measurements, achieving an R^2 of 0.63. Mitchell et al. (2021) comes closest to the goal of this research. They used RFR to spatially predict sedimentation accumulation rates in the Baltic Sea while using hydrological and spatial parameters like mean current speed, the orbital velocity at the seabed, and Suspended Particulate Matter (SPM). Their approach resulted in an R^2 of 0.419. On a coarser global scale, Restreppo et al. (2020) achieved a score of 0.89 to predict Oceanic sediment accumulation rates using a k-nearest neighbour (k-NN) algorithm. K-NN uses parametrically nearest observed data to predict a probable value for an area without data.

RFR can handle multivariate time series as input, which requires data transformation. In the case of this research, this means transposing the time series of the conditions in the Botlek into a single row instead of a feature matrix. While doing so, lagged variables are created. These lagged variables allow the RFR to recognise time dependencies among the entries in the sample row. The dependent variable is coupled to the transposed row in a target column. Figure 2.4 shows how Mussumeci and Codeço Coelho (2020) transposed a multivariate meteorological feature matrix into a vector to forecast weekly dengue incidences up to four weeks in multiple cities.

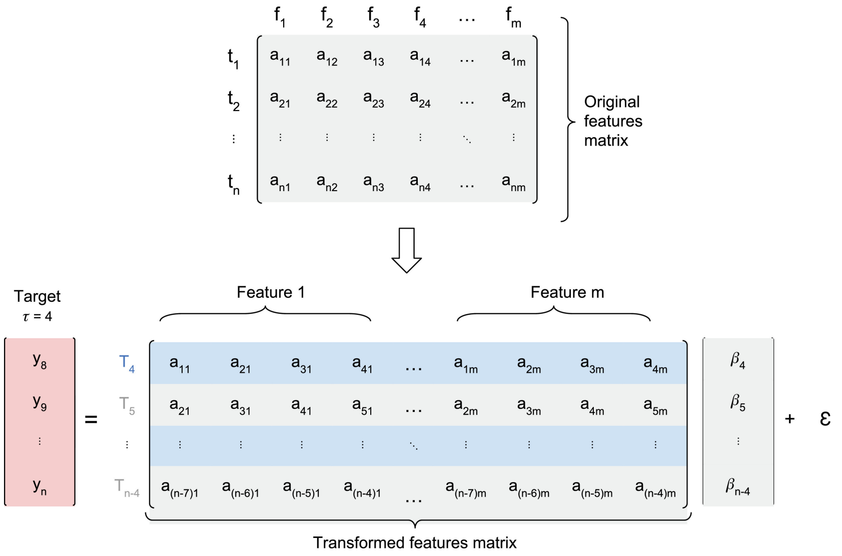

f1 f2 f3 f4 … fm
t1 a11 a12 a13 a14 … a1m
t2 a21 a22 a23 a24 … a2m
tn an1 an2 an3 an4 … anm
Original features matrix
Target
τ = 4
Feature 1
Feature m
y8
y9
yn
=
T4 a11 a21 a31 a41 … a1m a2m a3m a4m
T5 a21 a31 a41 a51 … a2m a3m a4m a5m
Tn-4 a(n-7)1 a(n-6)1 a(n-5)1 a(n-4)1 … a(n-7)m a(n-6)m a(n-5)m a(n-4)m
β4
β5
βn-4
+ ε
Transformed features matrix

Figure 2.4: Transformed data features for multivariate forecasting (Mussumeci and Codeço Coelho, 2020)

Their application of RFR is an excellent framework for this research. Mainly because their input parameters are similar to those of this thesis; for example, temperature and humidity are identical to the features in the Botlek scenario, and the use of multiple locations is comparable to the use of multiple dredging areas. Mussumeci and Codeço Coelho (2020) showed great forecasting potential and highlighted the ability of RFR to deal with the non-stationary and nonlinear nature of their problem. However, the RFR was outperformed by LSTM. The advantage RFR has is that it is computationally less expensive. This is favourable when weekly predictions for multiple locations are required, like in the Botlek.

RFR is highly accurate and provides information on feature importance. The latter is convenient for assessing which environmental parameters are essential in predicting dredging volumes. This and the promising results in the mentioned studies make RFR a viable option for this thesis. Disadvantages of RFR are the high sensitivity to noisy input data and the computational intensity that increases with complexity (Hengl et al., 2018).

2.2.4. Support Vector Regression

Support Vector Regression (SVR) is a variation of SVM and is a technique that finds a model with a margin around the predicted variables. This margin allows for a balance between fitting and overfitting. It is called SVR because the points nearest to the regression line define the error margin, and those points are called support vectors (Awad and Khanna, 2015). Elnabwy et al. (2022) use SVR with three different kernel functions: linear, polynomial, and radial basis. The predictions they made with SVR were close to those of ANN, but ANN still outperformed SVR. The regression problem is expressed below where eq. 2.9 represents the regression problem and eq. 2.10 the regularised risk function used to calculate internal parameters. C represents a weighing parameter, and ξ_i represents the distance between the boundary and real values. The SVR aims to minimize the squared weight, ξ_i and ξ_i^*:

$$y = f(x) = \sum_{i=1}^{n} w * K(X_i, X) + b \tag{2.9}$$

$$Minimise = R(C) : \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.10}$$

Doroudi et al. (2021) reviewed thirteen studies that applied SVR to predict suspended sediment loads with discharge and suspended sediment as input variables. They stated that SVR can be a fitting choice for modelling suspended sediment loads in rivers as it can solve common problems in hydrological datasets such as high dimensionality or small sample sizes. The characteristic of fitting small sample sizes could be relevant for modelling the sediment in the Botlek as parameters such as dredged volumes and bed level are measured monthly or weekly. The best performing SVR models showed a R^2 score around 0.94. A downside of SVR is that it is highly sensitive to hyperparameter changes, making tuning a challenging task (Yan et al., 2019).

SVR can be combined with PCA to reduce the dependencies between the features. Noori et al. (2022) implemented PCA on SVR models to model the total sediment load in rivers in g/L, achieving a $RMSE$ of 0.87. The PCA method assumes a linear correlation between the input parameters, so they use polynomial, sigmoid, and radial basis type functions to transform the input parameters to a nonlinear feature space. It was found that the PCA-based SVR model with the radial basis function outperformed the other kernel functions.

2.3. Environmental factors and sediment behaviour in the Botlek

The hydrodynamics in an estuary are complex and influence the sediment in many ways. The interface of riverine and marine systems results in an always-changing environment that causes resuspension, transport, or settling of sediment. Identifying and evaluating the critical factors within this environment is essential to selecting the correct features for developing the ML models.

The dissertation of de Nijs (2012) used the Botlek harbour as the area of interest to describe the transport and sedimentation processes of SPM in a stratified tidally energetic estuary. His work explains the internal flow structures and the transport of salt and sediment in the Botlek and is highly relevant to this research.

2.3.1. Tidal forcing and sediment transport

Tidal characteristics

Bosboom and Stive (2023) state that M2 (semi-diurnal, 12.42-hour period) is the tidal-current constituent that is the dominant condition in most tidal basins in the Netherlands. Moreover, the Rhine-Meuse estuary can be classified as a meso-tidal estuary, which means a tidal range between 2 and 4 meters (Van den Berg et al., 2007). Figure 2.5 shows the tidal variation at the radar station in the Geulhaven, a harbour inside the Botlek. The semi-diurnal characteristics and the beating of neap and spring tide are visible. The vertical dashed lines approximate the spring and neap periods. The range at spring tide can be up to 2.5 meters, showing the meso-tidal character. For neap, this is closer to 1 meter.

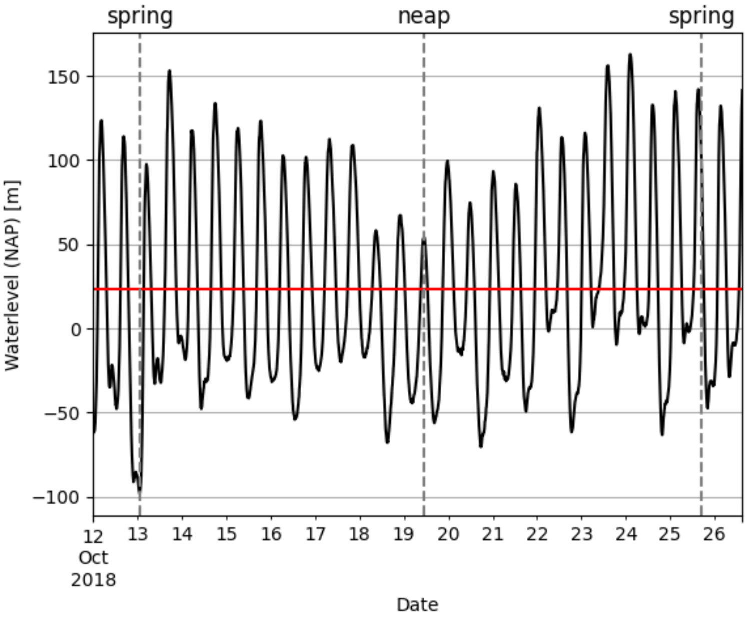


Figure 2.5: Spring and neap tide variation in the Botlek (Port of Rotterdam, 2024b)

Forming of the Estuarine Turbidity Maximum

Tidal forcing and river flow are the main parameters responsible for the dynamics in an estuary and, thus, for the POR. Tidal forcing results in saltwater intrusion, which ranges according to the height of the tidal cycles. The movement of suspended sediment and water in estuaries is controlled by the density gradients between salt and fresh water, which cause circulation. The variations in density are caused by the relative mixing rates that are a ratio of river and tidal flow (Allen et al., 1980). Regarding this, de Nijs (2012) stated that the tidal advection in the Rotterdam Waterway controls the phase at which turbid and saline water reaches the Botlek. Measurements showed that the shear during tidal cycles is not high enough to induce interfacial mixing, resulting in a stable advection of the salt wedge up to the mouth of the Botlek. Consequently, the stratification at the pycnocline (the density boundary between the fresh and saline water) (Britannica, 2024) limits how high the bed-generated turbulence can spread in the water column. The damping effect of the pycnocline traps SPM at the tip of the salt wedge, thereby forming a SPM balance beneath the pycnocline which is called the Estuarine Turbidity Maximum (ETM) (Geyer, 1993).

Sediment transport in the Botlek

The advection of the ETM determines the availability of SPM that can be exchanged with the harbours in the Rotterdam Waterway, a process that is determined by the saltwater intrusion length (Nijs et al., 2010). The harbour basins in the Rotterdam Waterway located near the salt intrusion limit are filled and emptied throughout tidal cycles. The SPM transported in the ETM can be exchanged with the harbour basis as a result of these exchange flows (Geraeds, 2020). In his dissertation, de Nijs (2012) concludes that the near-bed density currents, induced by the density gradients, are the dominant factor in the transport of SPM relatively far into the Botlek. The lack of vertical mixing due to the stratified water column makes it so that the SPM can settle, resulting in a trapping efficiency of nearly 100%.

2.3.2. Sediment properties and origin

The deposited sediment in the Botlek is primarily of fluvial origin (de Nijs, 2012). Differential advection during ebb causes fluvial SPM to be transported over the salt wedge. The SPM in the fresher upper part of the water column is able to settle because the stratification damps turbulence during the tidal cycle. 3D hydrostatic numeric modelling predicted that the fluvial SPM in the wedge is advected into the harbours along the waterway and settles there. Tracer analyses and sediment budgets substantiated this. The fact that fluvial SPM is the primary sediment source could already indicate that the discharge in the Rotterdam waterways is a critical parameter for this research.

Dredging material from the Botlek is predominately silt (Port of Rotterdam, 2022). There is a distinction between the transport of fine and coarse sediment. The hydrodynamic conditions at the point of consideration primarily determine the transport of coarse sediment. For fine sediment, the flow conditions upstream and in the past also influence the transport. This causes an essential difference in the tidal dynamics of coarse and fine sediment. For example, the timescale for sand is an order of magnitude smaller than the tidal period, which is why an instantaneous response can be assumed. In contrast, silt has a similar timescale (Bosboom and Stive, 2023). Therefore, the direction and stream velocity of the Botlek tide could be important. These variables vary over the basin during the tidal cycles, resulting in an inhomogeneous suspended sediment response.

Another effect of salinity on sediment properties is flocculation. Mhasshah et al. (2017) showed that the interaction between suspended sediment concentration and salinity controls flocculation size and settling velocity. Faster settling occurred at higher concentrations when salinity was low. This situation was reversed when the salinity was higher, but the concentration was lower. However, Eisma et al. (1991) concluded that flocculation is unimportant as observations showed that sediment is already aggregated in freshwater regions of the POR.

3 Methodology

This chapter describes the methodology for answering the main research question. Chapter 3 consists of six sections that all contribute to answering the first three sub-questions of this thesis:

- *What data is available and relevant for predicting SR using ML in estuarine harbours?*
- *Which ML algorithms are most suitable for predicting SR?*
- *How should the selected ML algorithms and features be configured to predict SR?*

This chapter starts with a case study, describing the study area and the maintenance strategy of the POR. Next, in Sections 3.2 and 3.3, the data available for this research is outlined and analysed with an Exploratory Data Analysis (EDA). The EDA increases the understanding of the data by providing insight into outliers and missing values. These two sections cover the first sub-question.

Section 3.4 will select the most suitable ML methods for predicting SR in the Botlek, thus answering the second sub-question. It does so by evaluating the ML algorithms covered in the literature review on criteria that are based on the limitations of this study. Together with the findings from the case study and the EDA, a well-considered choice in ML algorithms can be made.

Following selecting the appropriate algorithms and cleaning the available data, Section 3.5 and 3.6 explain the method behind configuring these into models that can predict SR. Section 3.5 focuses on overcoming any remaining problems with the data and the process of engineering the data into samples to train the ML models. Section 3.6 outlines the iterative steps of developing the selected ML models and the hyperparameter tuning process. The model development will be divided into modelling phases that build upon each other and implement the steps from Section 3.6 to reach the optimal configurations to predict SR. These phases are explained in Section 3.6.3. With these three sections finished, the third sub-question is covered.

The materials and methods acquired and developed in this chapter will provide the foundation to produce reliable results. Chapter 4 will display these results, answering the fourth sub-question. The discussion in Chapter 5 covers the result analysis, study evaluation, and the last sub-questions. Chapter 6 covers the general findings and the conclusion to the main question. Finally, in Chapter 7 the recommendations on further research and model improvements are presented. To summarise the complete workflow of this study, Figure 3.1 provides an overview of intermediate steps on the next page. The methodology is finished after the 'Algorithm Selection & Model Engineering' sections.

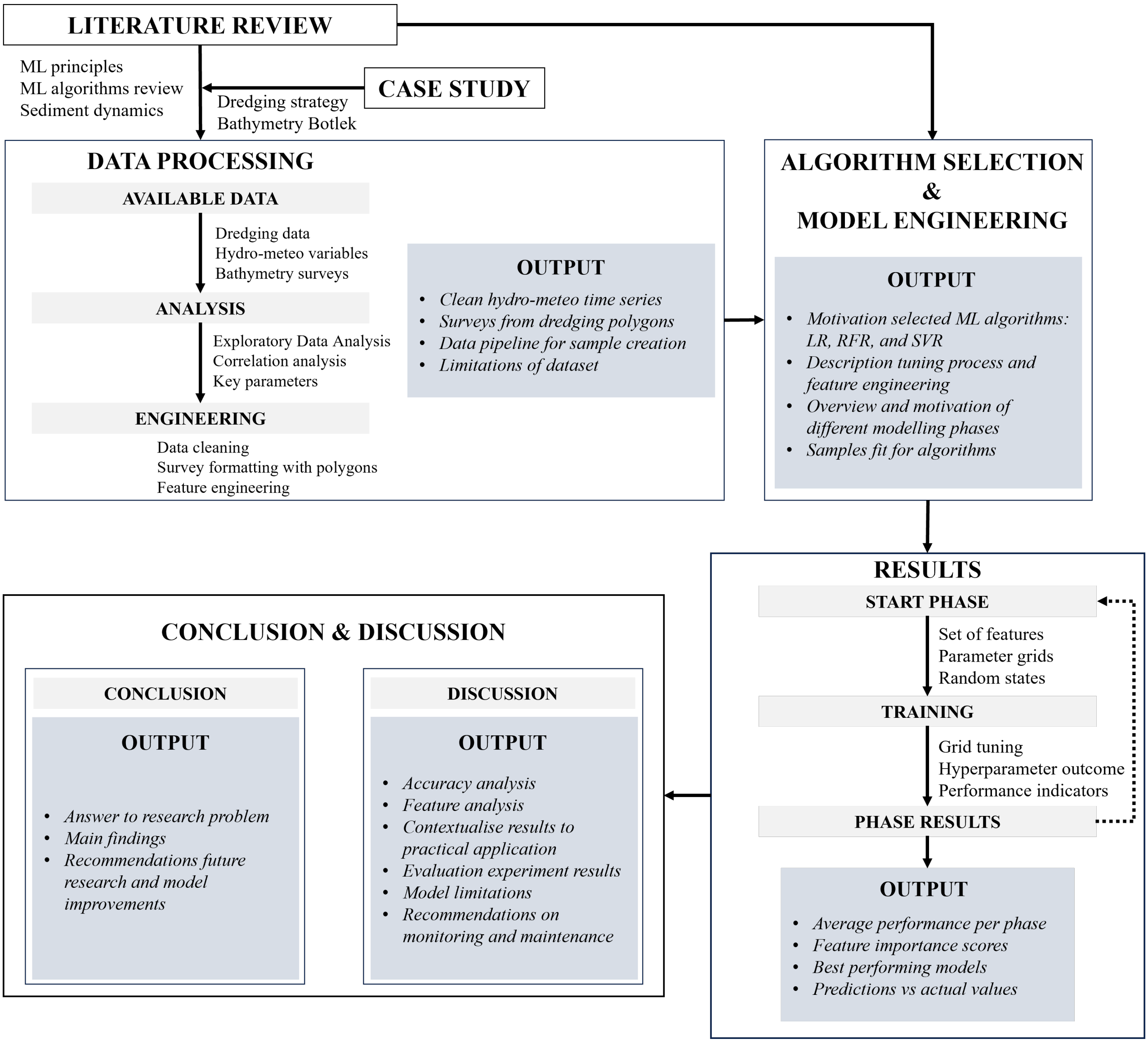


Figure 3.1: Flowchart research methodology

3.1. Case study

The area of interest of this thesis is the Botlek. The Botlek Harbour is a basin of the POR that is primarily dedicated to the chemical industry. The entrance of the Botlek lies in the transition zone from saline to fresh water, approximately 18 km upstream from where the Nieuwe Waterweg flows into the North Sea at Hoek van Holland. Section 2.3 showed that the basin is under a strong influence of the salt wedge induced by tidal advection and that this is the dominant factor in the high SR in the Botlek. The resulting conditions require frequent survey and dredging operations to ensure the navigational safety of the vessels. The SR here are so high that the Botlek accounts for most maintenance dredging costs (SOURCE). A detailed overview of the dynamics within the Botlek has already been covered in Section 2.3.

Understanding and predicting sediment behaviour under environmental conditions is valuable for port authorities as this can increase maintenance efficiency. It can prevent unnecessary surveying or dredging operations by providing missing information in decision-making. Moreover, possessing a predictive model can help anticipate sedimentation after extreme weather events. This research aims to develop these models to contribute to more efficient sediment management in the Botlek and, eventually, the rest of the POR.

3.1.1. Surveying and dredging in the Botlek

The information regarding the maintenance strategy was mainly acquired by interviews with Edwin Hupkes (project manager at Port Development and manager of the PRISMA program) and Andre van Hassent (Asset Management (AM)).

The port authority divided the riverbed into surveying areas to manage them as seperate assets. Figure 3.2 shows the seven areas relevant to this study delimited by the dotted lines in PortMaps, a software available to POR employees. The coloured regions within the surveying areas are dredging areas. The cutoff in the Botlek mond+zwaaikom is because the maintenance responsibility transfers to Rijkswaterstaat (RWS).

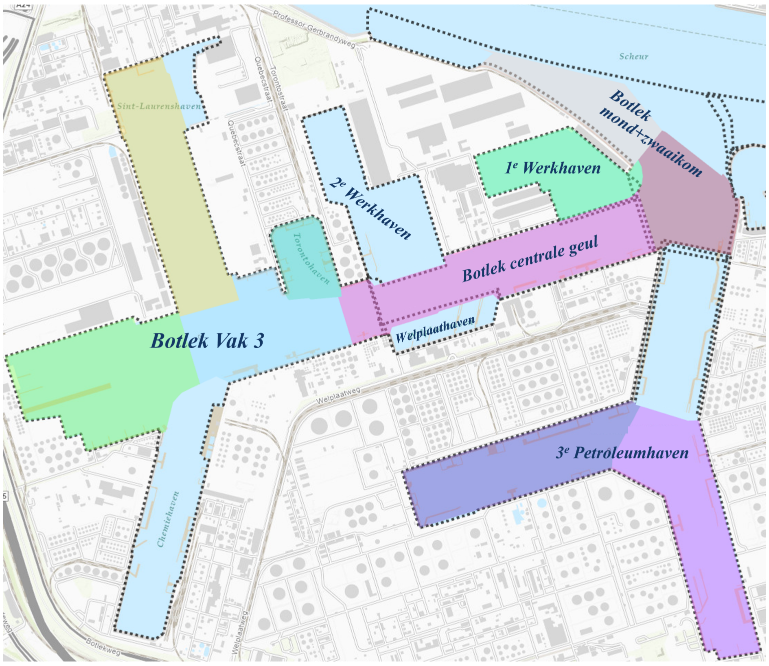


Figure 3.2: The surveying areas of the Botlek harbour (Port of Rotterdam, 2024c). The areas are separated by the dashed lines.

The three areas that experience the highest sedimentation are the 3e Petroleumhaven, Botlek centrale geul, and Botlek mond+zwaaikom. Consequently, the authorities survey these areas most frequently. The surveying interval for each area is relevant for developing the models. Variance in the interval leads to an inconsistent time series length for every pair of sequential surveys. Ideally, this length is identical as the ML algorithms from Section 2.2 require samples of the same shape. Uneven spacing, therefore, leads to unnecessary removal or padding of data, which results in a noisy dataset. Unfortunately, maintaining a constant interval is impossible as authorities rely on factors such as the traffic inside their port, equipment availability, occupied berths, and environmental conditions. Even so, the average survey intervals for the primary areas are close to exactly 31 days. The areas and their respective intervals are displayed in Table 3.1. The codes explained in the caption are essential in the modelling phase as these serve as sample labels.

Table 3.1: Overview of dredging locations and number of surveys per location. The 'Function place' columns provide the code used to refer to the surveying areas. 'Number' is the number of surveys recorded since January 1st, 2018 till April 31, 2024

Name location	Function place	Dredging location	Number	Average survey interval
Botlek Centrale Geul	H-L-N-BT-004-PLV-009	H-L-N-BT-004-BGV-ABG	69	30 days
Botlek mond	H-L-N-BT-004-PLV-014	H-L-N-BT-004-BGV-ABH	31	31 days
		H-L-N-BT-004-BGV-ABK		
Botlek Vak 3	H-L-N-BT-004-PLV-017	H-L-N-BT-139-BGV-AOZ	33	64 days
		H-L-N-BT-004-BGV-ACM		
		H-L-N-BT-128-BGV-AFO		
		H-L-N-BT-004-BGV-AAW		
		H-L-N-BT-128-BGV-ABF		
3e Petroleumhaven	H-L-N-BT-096-PLV-003	H-L-N-BT-096-BGV-AAM	69	31 days
		H-L-N-BT-096-BGV-AAN		
		H-L-N-BT-096-BGV-AAO		
Welplaathaven	H-L-N-BT-145-PLV-027	H-L-N-BT-145-BGV-AJE	10	213 days
		H-L-N-BT-145-BGV-AJO		
1e Werkhaven	H-L-N-BT-167-PLV-001	H-L-N-BT-167-BGV-AAF	44	50 days
2e Werkhaven	H-L-N-BT-168-PLV-002	H-L-N-BT-168-BGV-AAL	34	64 days

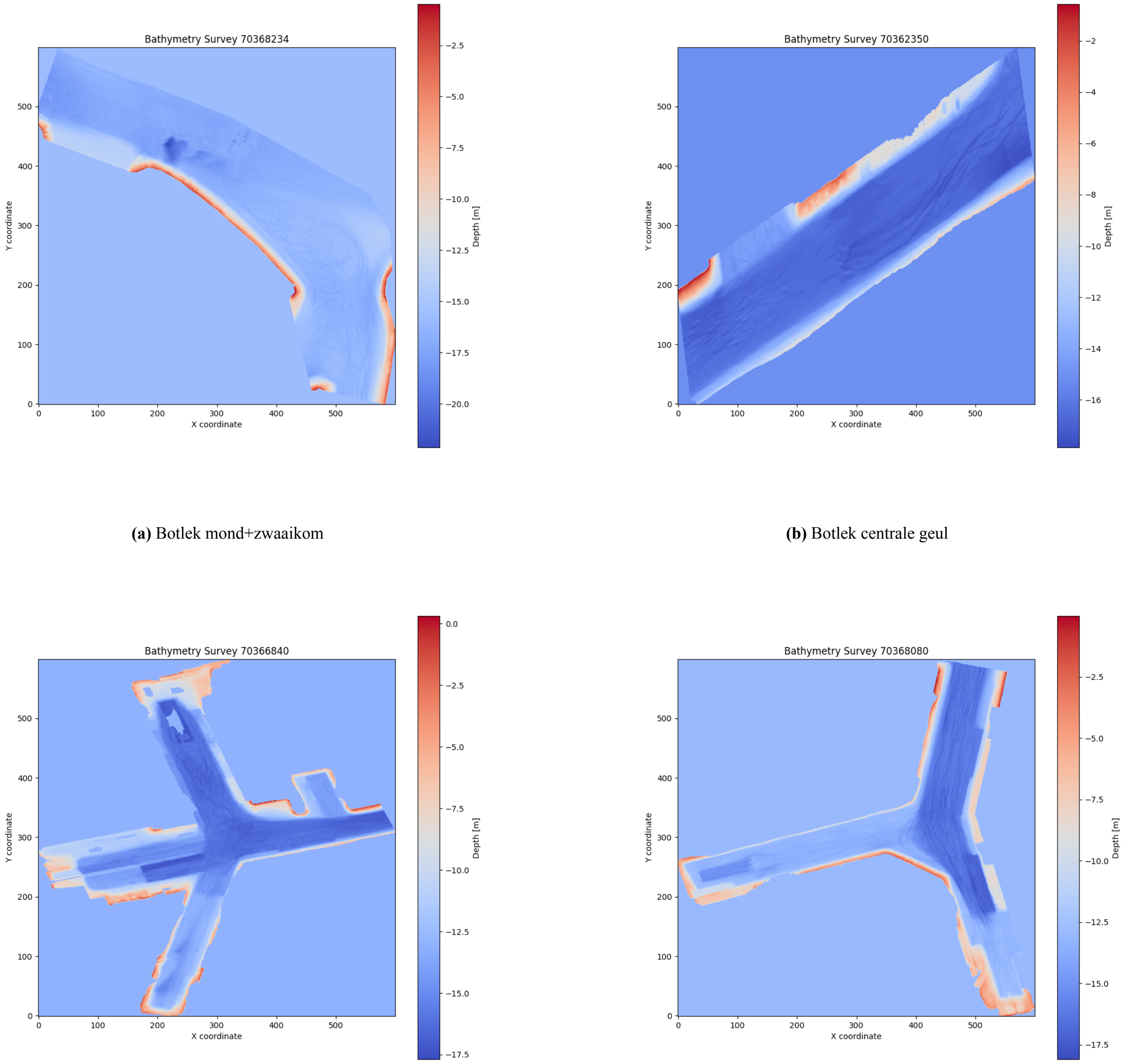
Maintenance dredging operations do not follow the same pattern as survey operations. The first reason for this is that some surveying areas consist of multiple dredging areas (see Table 3.1). A dredging operation is sometimes limited to a single area, dependent on the level of urgency of the surrounding areas. For example, the two purple areas in the 3e Petroleumhaven are less prone to sedimentation than the upper blue area. It is then likely that only the blue area is dredged even though the survey covers all areas. The second reason is that the surveys not always indicate a reason to dredge. If AM analyses the bathymetry survey and notices that the channels are still below the NGD, then dredging is not necessary. When the NGD is breached, AM creates a dredging order for that area that can take several days and sometimes longer before it is completed. Again this depends on factors like equipment availability and traffic. These inconsistencies and dependencies make maintenance dredging a continuous process that does not follow the exact timing of the surveys.

Dredging area analysis

Knowledge on the individual dredging areas can provide initial insights into SR in the different areas of the Botlek. The dredging and surveying data is analysed and processed in Sections 3.2, 3.3, and 3.5. Therefore, a review of the SR and conditions of the individual areas is performed once the data is cleaned and processed. The areas are covered in Section 3.5.3. Figure 3.4 in the next section shows all individual dredging areas and their respective location code, along with the measurement stations in the Botlek.

3.1.2. Bathymetry

The port must adapt its infrastructure to the continuously growing vessel size to keep its competitive position. The most significant alteration to the study area was the deepening of the Nieuwe Waterweg and Botlek from 2018 to 2019. The new depth of -15.9 meters NAP assured the navigational safety of Aframax and the, at the time, new Panamax vessels. The deepening is essential for this research because it changed the conditions in the Botlek. The implications of these changes are discussed in Section 3.2. The -15.9 meters NAP is currently still maintained in the channels. Figure 3.3 displays the bathymetries of the entrance (3.3a), the central channel (3.3b) up the center of Botlek Vak 3 (3.3c), and part of 3e Petroleumhaven (3.3d). The Multibeam surveys were taken in April 2024. The navigational channels of approximately -15.9 meters NAP are distinguishable by their dark blue colour. The east side of the central channel harbours a sediment trap with a maximum depth of -19 meters NAP. The trap is visible in the figure 3.3b.



(a) Botlek mond+zwaaikom

(b) Botlek centrale geul

(c) Botlek Vak 3

(d) 3e Petroleumhaven

Figure 3.3: Multibeam bathymetry surveys of the Botlek. The colour bar shows the depth in meters.

3.2. Available data

The literature review mentioned the algorithms commonly applied for sediment prediction and their corresponding characteristics. The data provided by the POR and other sources is essential for developing these ML models because the quantity and quality of the data determines which algorithms can be applied and what the model development approach must be. This section will provide a detailed overview of the data and serve as input for Section 3.4 and 3.6 as these will expand more on the algorithm selection and model development. Along with the EDA in Section 3.3, the first sub-question is covered and finally answered in a conclusive section:

What data is available and relevant for predicting SR using ML in estuarine harbours?

The data that is available can be divided into three groups:

- Geospatial Multibeam surveys of all areas in the Botlek harbour
- Dredging logs of all operations commissioned by the POR authorities
- 10-minute interval measurements or forecasts of all available hydro-meteo variables

This data was acquired from October 9, 2018, until May 2024 because October 9, 2018, is the first recorded day in the POR database for many variables. Moreover, 2018 marks the start of the deepening of the Nieuwe Waterweg and Botlek that was mentioned in the bathymetry section of 3.1.1. The asset management department highlighted that this infrastructural intervention impacted the sedimentation behaviour. Consequently, environmental data collected before October 2018 is not helpful for model training, as the ML models would attempt to capture trends and correlations that are representative of the present situation. However, October 2018 is not chosen as the starting period of the survey data collection to prevent unnecessary data loss. Instead, the surveys are acquired from the start of 2018 as these can provide insight into the changes in bathymetry after the deepening. The last deepening operation was completed in May 2019. When selecting suitable input data, the surveys before May 2019 will likely be discarded as unreliable labels.

3.2.1. Multibeam surveys

The surveys are conducted with a Multibeam sonar by the surveying vessels from the port authorities and converted into high-resolution data, allowing for a detailed visualisation (Figure 3.3). A survey is conducted within a surveying area and linked to its specific area code, denoted by the *Function place* column from Table 3.1. Additionally, each survey has a unique order number. The surveys are stored in the POR Spark Database and can be extracted by specifying a Spark SQL query. A list of all order numbers linked to a survey conducted in the Botlek was constructed. By specifying the order number and date of surveying, all relevant surveys stored since 2018 could be extracted using the information in the list. Table A.1 in Appendix A shows an example of this list.

A survey from the Multibeam can be classified as geospatial point data. Geospatial data combines location information (coordinates) with attribute information (depth) (IBM, 2024a). The POR uses the EPSG:28992 Amersfoort system for the location information, also known as the Rijksrectangularcoordinate (RD) system. This system projects xy-coordinates into meters and covers The Netherlands and the Dutch Exclusive Economic Zone of the North Sea (NSGI, 2024).

The ’Number’ column in Table 3.1 contains 290 surveys. Section 3.5.1 explains the method behind exporting and storing all these surveys.

3.2.2. Dredging data

The dredging data consists of the detailed on all maintenance dredging operations in the POR jurisdiction since 2007. A dredging operation is denoted with the year and week it took place, a dredging location code from Table 3.1, the volume that was dredged, and the time it took to complete the operation. Through the dredging area code, each region can be coupled to its respective surveying area.

3.2.3. Hydrological and meteorological variables

The hydro-meteo conditions are monitored throughout the port area. In cooperation with RWS, the POR manages the ’Weather, Tides and Water Depth’ portal, providing all current conditions needed to navigate the port safely (Port of Rotterdam, 2024a). The POR authorities have measurement stations in all harbours and channels, RWS covers the main navigational channels (Nieuwe Waterweg, Nieuwe Maas, Het Scheur), and the meteorological conditions are monitored by the Royal Netherlands Meteorological Institute (KNMI) (KNMI, 2024). Additionally, the OSR provides highly detailed forecasts of water level and flow dynamics. OSR refers to the numerical model developed by Svasek Hydraulics. The model is built upon the WAQUA (Water Movement and Water Quality modelling) in the SIMONA platform from RWS and is used to solve 2D shallow water equations (Svasek Hydraulics, 2024).

POR employees can access and download all historical data, including RWS data, through the Historic Data Store (HDS) HydroMeteo. The historical data can be exported as a CSV file containing the time series of the selected variables. Table 3.2 shows an overview of the available variables. All variables are measured every 10 minutes, except for precipitation, which is recorded hourly.

Table 3.2: Overview of available modelling variables in the Botlek. The 'Code' column provides the code names of the conditions. 'Measurements Locations' refers to the name codes of the measurement locations (see Table A.2). The 'Method' indicates whether the variable is measured or predicted by the OSR model. The 'Source' column either contains RWS, POR, or KNMI as entry

Variable	Unit	Code	Measurement Locations	Method	Source
Discharge	m^3/s	Q10	LOBITH	measured	RWS
Height of tide	m	H10	RP10	measured	RWS
Depth averaged tidal direction	deg	PTSDDA10	BOTCGO, BOTM, BOTNM	OSR	POR
Depth averaged tidal stream	m/s	PTSRDA10	BOTCGO, BOTM, BOTNM	OSR	POR
Precipitation	mm/hr	RH	RP10	measured	KNMI
Salinity	g/kg	PSAB10	RP10, BOTCGW, 2WERKH	OSR	POR
Water temperature	C°	WT10	HARK, RIJNH, HOEK, LEKH	measured	POR
Wind direction	deg	WD10	RP10	measured	POR
Wind velocity	m/s	WV10	RP10	measured	POR

The measurement and OSR locations are divided over the port to cover all relevant areas and are recognisable by a name code in the table above. A critical observation is that the number of variables measured at each location differs. For example, salinity is modelled by OSR at the Botlek Centrale Geul West (BOTCGW) but not at the Botlek Mond (BOTM). The actual names and area of these locations are provided by Table A.2, Figure A.1, and Figure A.2 in Appendix A.

For the data analysis in Section 3.3.1, all variables at their respective measurement locations have been exported from October 1, 2018 until April 29, 2024.

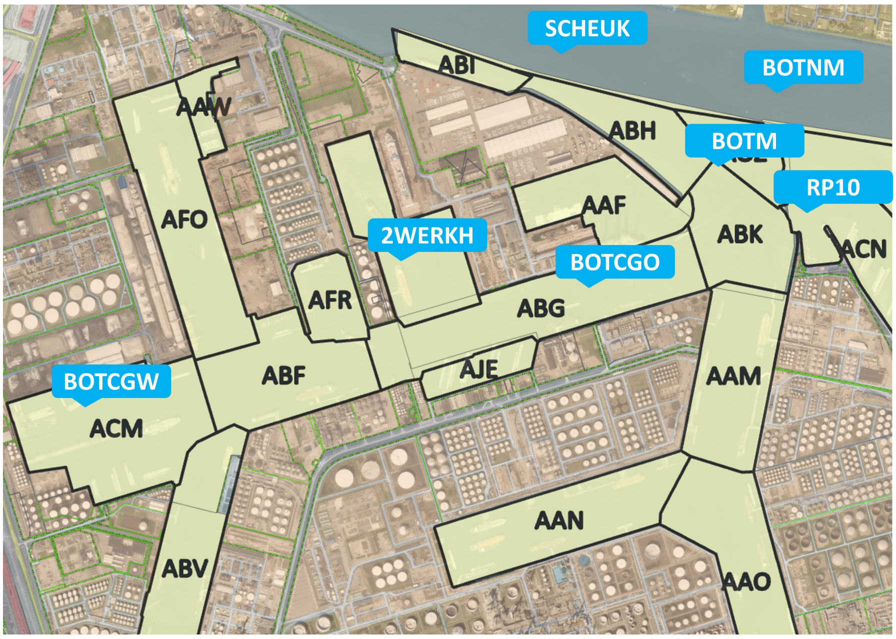


Figure 3.4: The dredging areas in the Botlek and their respective area codes. The area codes are the last three letters from 'Dredging locations' in Table 3.1. The blue codes are the measurement stations from Table 3.2.

3.3. Data analysis

This section covers the analysis and preprocessing steps of the data from Section 3.2. These steps are necessary to provide a reliable dataset for the model development. An EDA is performed to summarise the main characteristics of the datasets. This analysis will increase the understanding of underlying distributions of individual input and identify outliers and missing data. The EDA was performed in Python with the Pandas, matplotlib, and Seaborn packages. Appendix B contains the plots, distributions, and tables with characteristics of all variables and will often be referred to.

3.3.1. EDA and processing of hydro-meteo variables

An EDA increases the understanding of the data. It consists, among other things, of the following steps (Brownlee, 2020a): descriptive statistics summary, visualisation by line charts and histograms, correlation matrices, and bivariate plots. The pandas.describe() function provides an overview of the descriptive statistics of a dataset. The function returns the count, mean, standard deviation (STD), minimum, quantiles, and maximum of the set. The plots can be made with the matplotlib package and the correlation matrices and bivariate plots with Seaborn.

Each variable could need engineering to solve issues with missing values or outliers identified by the EDA. This section describes the EDA and preprocessing steps.

Missing values and Large p, Small n

Section 3.2 states that all data before May 2019 will likely not be used for modelling. Therefore, missing data and outliers before this date are not interpolated or filled with other methods.

During model training, keeping the number of input features (p) below the number of samples (n) is preferred. When p $>>$ n, the performance of the models can be poor because the training sample size is relatively small compared to the features vector size, resulting in possible over-fitting. This problem is called the 'Large p, Small n issue' (Huynh et al., 2020). This research aims to use the hydro-meteo time series as features while the surveys provide the labels for the samples. The total of 290 surveys is significantly smaller than p if the 10-minute measurements are chosen as features. For example, if the mean interval of 31 days between surveys is selected as time series length, the sequence would have a length of 4.464 features for a single variable. To prevent p $>>$ n, the time series will be resampled to daily, weekly, or monthly averages to reduce the number of features. Before resampling, stand-alone missing values are filled with appropriate methods. More extended periods with missing values will be discarded.

Discharge (Q10)

The discharge is measured at Lobith by RWS. According to the port authorities, the water takes approximately four days to travel from Lobith to the Botlek (Port of Rotterdam, 2024a). Therefore, the data is shifted by four days to be more accurate. Figure B.1 provides the first noticeable outlier. The minimum discharge in the set was recorded at -3604.98 m^3/s. A negative discharge is not possible in the Rhine. Historically, the discharge in the Rhine at Lobith has never been below 620 m^3/s (H20, 2022). The line chart in Figure B.1 shows multiple spikes towards zero or below, indicating more outliers. All values below the realistic minimum discharge of 620 $m3/s$ are set to NaN to remove the measurement errors. Figure B.2a shows the resulting line chart with many missing values before April 2019. All NaN-values after April 2019 are interpolated.

Height of tide (H10)

The tidal variation in Figure B.3a shows a fluctuation around a mean water level of + 0.19 meters NAP. The measurement station is located in the Geulhaven with a depth of -6 meters NAP (Port of Rotterdam, 2024c). The high and low water (HW/LW) of each tidal cycle are recorded separately, as seen in Figure B.3b). At these moments, H10 is supposed to denote a NaN-value, but this process is ineffective. The time index of HW and LW does not follow the 10-minute interval, and the maximum or minimum value often does not coincide with the actual water level. Therefore, filling the H10 with its respective LW or HW value is unreliable and handling all tidal cycles manually takes time. Instead, the missing values in H10 are interpolated with the forward fill method.

Precipitation (RH)

The Precipitation dataset provided by the KNMI does not contain missing values (KNMI, 2024). It is the only variable from Table 3.2 measured as an hourly cumulative in 0.1 mm/hr. When the hourly precipitation was below 0.05 mm/hr, the KNMI denoted a -1, as seen in FigureB.9a. For this study, these values were replaced with zeros, and the unit was converted to mm/hr.

Salinity (PSA10)

Salinity is modelled by OSR at the surface (PSAS), middle (PSAM), and bottom (PSAB) of the water column for three locations in the Botlek. Figure B.10 shows the difference between these layers, with the bottom layer containing the denser, more saline water. The mean values at BOTCGW and 2e Werkhaven (2WERKH) do not significantly deviate from each other. displays lower levels of salinity, which could be ascribed to the fact that the location is stationed in the shallower Geulhaven. The data contains minimal missing values and no prolonged periods. Therefore, the NaN-values can be interpolated without extra steps.

Water temperature (WT10)

The water temperature is measured at multiple locations across the port. Figure B.14 provided the statistics of all locations. The mean temperatures across the different locations do not deviate much more than 0.1 degrees Celsius, except for HOEK, located near the shoreline. Figures B.15 and B.16 show the line charts of the seasonal temperature fluctuations. The periods of missing values are visible in all charts except for the HARK location.

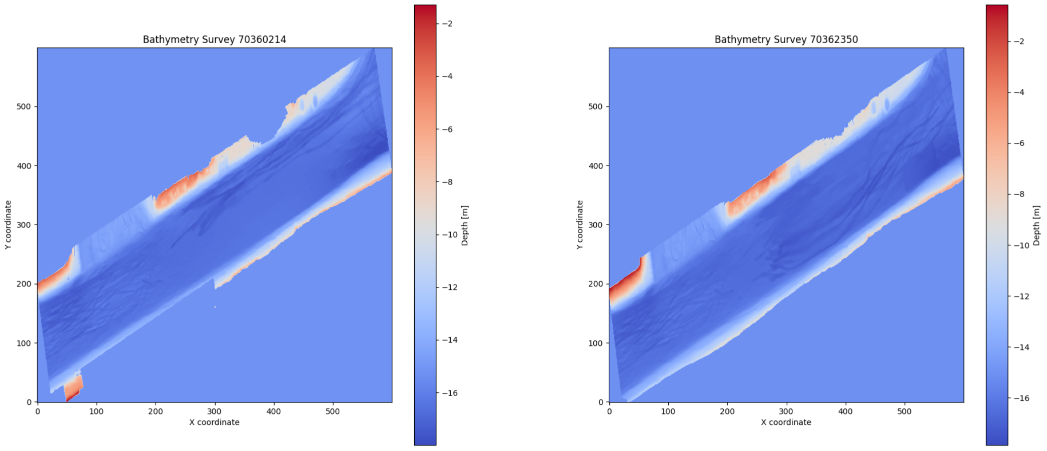
The Botlek is situated between LEKH and HOEK. Ideally, these two locations would be interpolated to estimate the temperature at the Botlek. This is not possible due to the significant periods of missing values at both locations. Of all locations, HARK is the most reliable source. The missing values in the HARK data still have to be interpolated.

Wind direction/velocity (WD10/WV10)

The wind is measured at RP10. Figure B.17a shows the distribution of WD10 and is as expected for the Netherlands. The dominant south-to-southwest wind direction (Janssen, 2024) is visible by the prominent peaks at 180 to 270 degrees. The WD10/WV10 data did not show significant periods of missing data.

3.3.2. Survey analysis

Section 3.2.1 describes the first steps in exporting the surveys towards usable data. An SQL query is needed to access the surveys. If only the order number is specified, all data points are exported. The minimum and maximum coordinates define the grid size, which is filled by the depth values, resulting in the bathymetries in Figure 3.5.



(a) January 1, 2024

(b) February 2, 2024

Figure 3.5: Example of misalignment in survey borders and coordinates at BOTCGO

An immediate observation is the difference between Figure 3.5a and 3.5b. The surveys are not conducted over the same area, as the edges of the channel are captured differently for both surveys. Consequently, the number of data points in both surveys is not identical. This can be seen in Figure B.18 where survey 3.5a consists of 227,134 points and survey 3.5b of 238,997. The mismatch means that rasters of the same area are not evenly filled. Therefore, performing a raster substraction to find the change in bed level will cause down or upward spikes, as shown in Figure 3.6. The colour map shows a bed level change ranging from +15 to -14 meters, implying that some sections either rise or fall almost the entire depth of the channel within a month. This is not realistic.

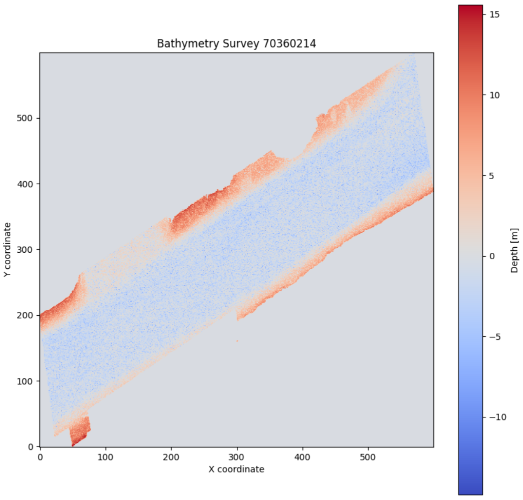


Figure 3.6: Change in bed level between survey 3.5a and 3.5b between January 1, 2024 and February 2, 2024.

The variation in bed level between two surveys can be determined once the surveys are mapped over the same grid. Section 3.5.1 will cover the steps that are taken to acquire reliable survey information.

3.3.3. Conclusion available data analysis

The sub-question that Sections 3.2 and 3.3 answer is:

What data is available and relevant for predicting SR using ML in estuarine harbours?

The data analysis indicates that there is sufficient data to train ML models on not only dredging data but also the hydro-meteo conditions. Multibeam surveys, dredging, and hydro-meteo data is available and all three data types are necessary for constructing reliable samples to train ML models to predict SR. The data has been exported from October 9, 2018, until May 2024. However, due to the deepening operations in the Botlek, all data before May 2019 will likely be discarded as unreliable as the deepening was finished in May. The surveys are essential in providing detailed changes in bathymetry, while the dredging logs add context to how much sediment leaves the system. Table 3.2 summarises the available number of hydro-meteo variables and the locations where these are measured. The multiple locations allow for (somewhat generalised) dredging area-specific hydro-meteo conditions.

The data still needs extensive engineering before it is usable. First, the survey grid problem from Figure 3.6 must be overcome to provide the sedimentation between two surveys. Secondly, the time series of the hydro-meteo variables and the dredging data need to be linked to the dates on which the surveys were conducted. Section 3.5 explains the detailed method behind this approach.

3.4. Algorithm selection

Section 2.2 discussed ANN, XGB, RFR, and SVR as applied methods within the field of sedimentation modelling. All methods showed promising results, but the time restrictions of this study limit the application of all algorithms. Therefore, a selection must be made. With this selection, the second sub-question is answered:

Which ML algorithms are most suitable for predicting SR?

The Goldstein et al. (2019) review paper on ML applications to coastal sediment transport stated that comparing ML predictors is often impossible because information on the final predictor is not provided, and the datasets differ. This is the case even when the same research question is investigated. Nonetheless, each ML algorithm has its own characteristics that can fit the goal of this study, and the reviewed literature shows proof of concepts or working models that can substantiate the choice for a particular algorithm. The simplest model that is selected to set a baseline performance is a Linear Regression (LR).

The additional ML algorithms are selected based on the goals and restrictions of this study:

1. Interpretability of the decision-making process and output of the algorithm. Interpretability is essential for formulating practical insights regarding sediment behaviour in the Botlek.
2. The limited number of surveys results in a small dataset. Therefore, the algorithm must be able to make predictions without overfitting on the small training set.
3. Ideally, the literature review shows that the algorithm can deal with the study restrictions while maintaining a high predictive accuracy on a dependent variable similar to or equal to SR.

Restrictions one and two are considered as the most essential points. The first reason is that the amount of data available for the Botlek is nearly unchangeable for this project. Surveys are added to the dataset at a monthly rate. Therefore, the set will not significantly increase in size throughout this thesis. Secondly, this effort to predict SR with ML algorithms is the first in the Botlek. Interpretability of the outcome and insight into the impact of hydro-meteo variables are essential in further developing the models and providing practical insights into the maintenance strategy of the POR.

Interpretability

Within ML, a distinction between black-box and white-box models can be made. ANN and SVR can be labelled as black-box as these contain complex distance functions and representation spaces that are hard to explain and understand in practical applications. Tree and pattern-based models like RFR and XGB are labelled as white-box, as these can provide a more interpretable model that is closer to human language (Loyola-González, 2019).

Sections 2.2.2 and 2.2.3 mentioned that XGB and RFR can provide a feature importance overview directly. The importance score shows the value each feature contributes to reducing the prediction error. Moreover, the RFR decision trees can be visualised, showing the decision-making process. The scores assist in improving the model performance as insignificant features can be identified and removed from the input set. This also increases the practical application of the models as the relation between the individual hydro-meteo variables and the SR is shown. Feature importance analysis for an ANN is possible but more complex. This would involve changing input features and recording changes in predictions or through backpropagation (Musolf et al., 2022). For SVR, Üstün et al. (2007) developed a method to visualise the information from the kernel matrix to interpret the optimised SVR model. The downside of these approaches is that both require additional modelling efforts, whereas RFR and XGB have this readily available in their packages.

Sample size and overfitting

Overfitting occurs when a model fits the training data too well, including the noise in the set. This results in poor performance on the test set as the model cannot generalise to unseen data. One of the main reasons for overfitting is a small training set (Ying, 2020).

Of the considered algorithms, SVR can handle smaller datasets well. N. Guo et al. (2020) applied and compared RFR, SVR, and GBR to predict energy consumption and mentioned that many researchers stated the superiority of SVR on smaller datasets. Moreover, SVR can tackle the standard problem of overfitting, especially for multivariate problems (Basak et al., 2007). Section 2.2.4 mentions that the tuning process is challenging and SVR is sensitive to hyperparameter changes. Prevention of overfitting is, therefore, not a given when implementing SVR.

RFR and XGB are both tree-based ensemble algorithms but differ in approach. The bagging technique used in RF decreases the risk of overfitting by taking random subsets of the input for every decision tree and averaging the predicted output, resulting in a lower variance (Belyadi and Haghighat, 2021). RF is often applied in medical research because it can handle low data availability in a high-dimensional feature space (Qi, 2012). Section 2.2.3 did highlight that RF requires high-quality data. Zhang et al. (2024) state that the sequential boosting approach of XGB and the regularisation term in the cost function can prevent overfitting. Despite these advantages, the tuning process is more complex than RF due to the many hyperparameters. Therefore, XGB is still susceptible to overfitting.

When considering performance on small datasets, ANN is not the ideal candidate as ANN often requires large quantities of data to prevent overfitting (Goodfellow et al., 2016). Nda et al. (2023) state that ANN is becoming increasingly popular in sediment prediction because of its ability to model complex relations, but highlight that overfitting poses a challenge without enough data.

Predictive accuracy

The accuracy of the proposed algorithms can be evaluated based on the results of the studies from the literature review. Table 3.3 summarises the performance metrics and dependent variables of all papers. Even though most of these studies did not consider SR as the dependent variable, they still offer great insight into the predictive performances of the different algorithms. The paper by Doroudi et al. (2021) covered thirteen studies that implemented SVR, which is why the ’R^2’ column consists of a range of scores. The ’Input variables’ column does not include the lagged or engineered features constructed from the unique input variables. How many additional features were used to develop the models is often unclear. The k-NN from the Restreppo et al. (2020) study is included because the dependent variable is similar to the one of this study.

Table 3.3: Summary of studies from Section 2.2. ’Algorithms’ contains the ML methods used in the study. The three performance indicators mention the scores in the same order as the abbreviations in the ’Algorithms’ column. ’Type’ either contains classification or regression to indicate the purpose of the model. ’Input variables’ shows the number of unique variables used for modelling. ’Dependent Variable’ shows the output variable, and ’Unit’ is the unit of that variable.

Source	**Algorithm(s)**	R^2	$RMSE$	MSE	**Type**	**Input variables**	**Dependent variable**	**Unit**
(Walsh et al., 2017)	RFR	0.63	-	0.49	Regression	7	TCS concentration	$\%TOC$
(Al-Mukhtar, 2019)	ANN, RFR, SVM	0.68, 0.8, 0.67	194, 130, 178	-	Regression	1	SSC	mg/L
(Pham et al., 2019)	ANN	-	0.47	-	Classification	6	Throat width	m
(Sharafati et al., 2020)	RFR	0.995	1575	-	Regression	3	SSL	ton/day
(Restreppo et al., 2020)	k-NN	0.89	-	-	Regression	-	SAR	$cm/year$
(Doroudi et al., 2021)	SVR	0.81-0.95	-	-	Regression	1-3	SSL	mg/L
(Mitchell et al., 2021)	RFR	0.419	-	0.12	Regression	7	SAR	$cm/year$
(Elnabwy et al., 2022)	ANN, SVR	0.99, 0.95	930, 2215	-	Regression	9	Sediment Volume	m^3
(Kim et al., 2023)	ANN	0.99	-	15.01	Regression	7	Turbidity	FTU
(Piraei et al., 2023)	ANN, RFR XGB	0.87, 0.89, 0.95	316, 299, 216	-	Regression	6	SSL	kg/s

All four algorithms show varying degrees of success when looking at their performance indicators and dependent variables. As stated before, comparing the performance of algorithms applied in different settings is complex but not redundant. For example, ANN reaches a near-perfect R^2 score of 0.99 for Elnabwy et al. (2022), outperforming SVR by 0.04. RFR slightly outperforms ANN in both reviewed studies from Table 3.3 that include the two algorithms. Additionally, RFR scores are near-perfect for Sharafati et al. (2020) in terms of R^2. Doroudi et al. (2021) showed that SVR can perform at a high level in thirteen studies considering SSL with a minimum score of 0.81, indicating a reliable performance level. Lastly, Piraei et al. (2023) state that XGB scores best on both R^2 and $RMSE$ within their framework.

3.4.1. Selected Algorithms

A summary of the advantages and disadvantages of the four covered algorithms and LR is constructed in Table 3.4 to support the algorithm selection process. Looking at the overview for ANN, it can be concluded that the need for large datasets and the lack of interpretability eliminates the algorithm as a candidate. These are the most essential criteria, and ANN scores low on both. The LR will be selected to produce a baseline performance. That leaves RFR, SVR, and XGB.

SVR is selected as one of two additional algorithms. The literature has shown that SVR is often superior for smaller datasets. This, together with the overall reliable level of performance in Table 3.3, makes SVR a suitable choice for predicting SR. Table 3.4 does show that SVR can be complex and uninterpretable. Therefore, taking SVR as a starting point is not practical. Without feature importance, it is challenging to eliminate insignificant hydro-meteo variables as input and formulate practical insights. Both RFR and XGB are strong contestants to complement SVR because of their ability to provide feature importance scores, but only one of these is chosen to be able to fit the time restriction. XGB is shown to be highly accurate but does require complex tuning to prevent overfitting. It is not ideal when both of the selected algorithms are complex to tune and developed. Therefore, RFR is more suitable to provide a baseline performance and identify the most significant variables. Once the RFR feature analysis has been performed, the new set can serve as input for the SVR or the LR models.

Table 3.4: Overview of advantages and disadvantages of the ML algorithms considered for modelling SR. The entries in ’Advantages’ and ’Disadvantages’ are either mentioned in Section 2.2 or Section 3.4. The selected models are in bold.

Algorithm	**Advantages**	**Disadvantages**
ANN	Captures complex relations well Flexible architecture High accuracy	Requires large datasets Risk to overfitting Difficult to interpret
LR	**Simple Implementation** **Low computational costs**	**Limited to linear relations**
RFR	**Provides feature importance scores** **Relatively straightforward tuning** **Bagging reduces risk to overfitting**	**Can be computationally expensive** **Sensitive to low quality data** **Prone to overfit noisy data**
SVR	**Effective on small datasets** **Less prone to overfitting** **Ability to handle multivariate problems**	**Sensitive to hyperparameters** **Complex tuning can cause underperformance** **Difficult to interpret**
XGB	High predictive accuracy Regularisation prevents overfitting Provides feature importance scores	Computationally expensive Complex hyperparameter tuning Susceptible to overfitting with poor tuning

3.5. Data processing

The previous sections covered the data cleaning and the motivation behind the selected ML algorithms. The output of these two sections must be combined into models that can accurately predict SR. By describing the method of modelling, the third sub-question is answered:

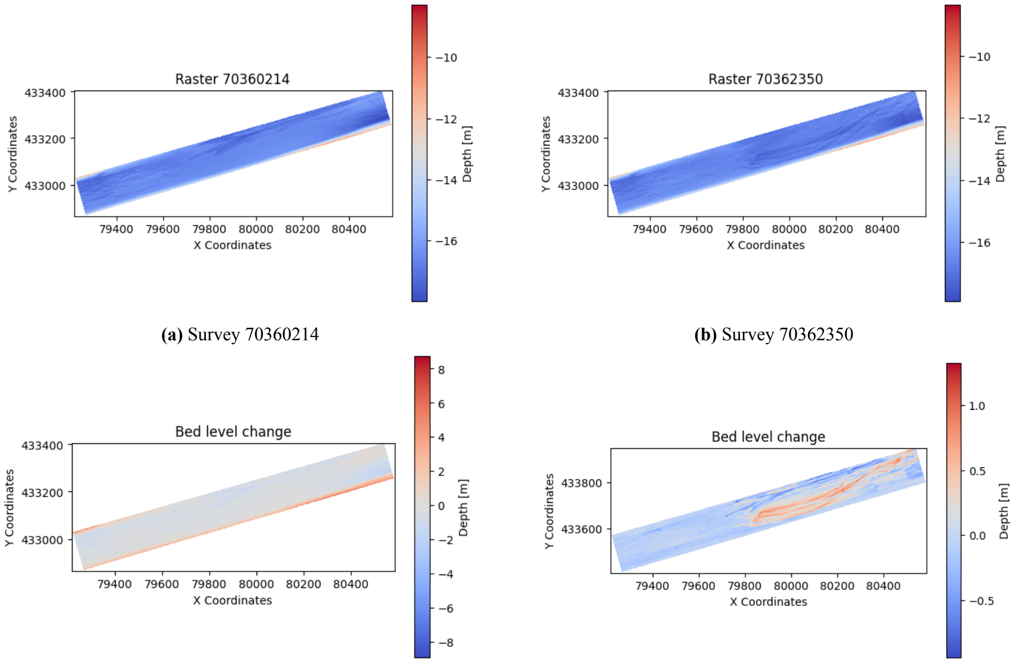
How can the selected ML algorithms and features be configured to predict SR?

Sections 3.5.1 through 3.5.3 describe how the surveys and hydro-meteo variables are engineered into samples that can serve as train and test data. Then, in Section 3.6, the possible feature engineering steps and hyperparameter tuning process are outlined. The entire ML training process is constructed of different phases that all consist of different configurations of samples, feature spaces, and dependent variables. The phase are described in Section 3.6.3. All findings are summarised into a conclusion to the third sub-question in Section 3.6.4.

3.5.1. Survey formatting

An accurate estimation of the bed level change is needed to provide information on the SR. The grid problems resulting in Figure 3.6 must be solved to achieve this. Removing the differences at the edges of the channels can be overcome by defining a polygon in the survey area. The polygon must capture the most critical aspects of the surveying area, such as the channels and the slope to the edges. It should be precise enough to ensure that nearly all surveys cover the full extent of the polygon. An inaccurate polygon still leads to deviations similar to the ones in Figure B.19. Combining a well-defined polygon with the WITHIN function of Spark SQL eventually leads to the desired result. The WITHIN function only extracts the survey data points within the defined area. The polygons used for this study are presented in Table B.1 and visualised in Figure B.20.

The polygons solve the inconsistent edge problem, but the coordinate mismatch from Figure B.18 remains. This can be overcome by interpolating over a grid. The grids are formed by the numpy.meshgrid() function. Meshgrid creates a rectangular grid with Cartesian indexing, the index being the xy-coordinates (ESPG:2289) that are defined for each data point within a survey. The data points do not cover all grid points within the mesh grid. Therefore, the scipy.griddata() function is used to interpolate between the grid points to fill missing values linearly. The results are visually identical survey areas as seen in Figure 3.7a and 3.7b. In reality, the grids are not identical due to the difference in the minimum and maximum xy-coordinates of the two surveys. The spikes in bed level change in Figure 3.7c prove this. Therefore, the last step must define a predetermined grid for each dredging area polygon. This grid is defined by the xy-coordinates of the first survey exported for that area, after which the following surveys can be interpolated as well. Now, when the bed level change is calculated, there are no spikes, resulting in a realistic image as seen in Figure 3.7d.



(a) Survey 70360214

(b) Survey 70362350

(c) Bed level change when using two coordinate grids

(d) Bed level change when using same coordinate grid

Figure 3.7: Two sequential surveys rasterised over same polygon. Figure 3.7c shows the result of subtracting the surveys if the rasters are not interpolated over the same coordinate grid. Figure 3.7d shows the result of subtracting after interpolating over the same coordinate grid.

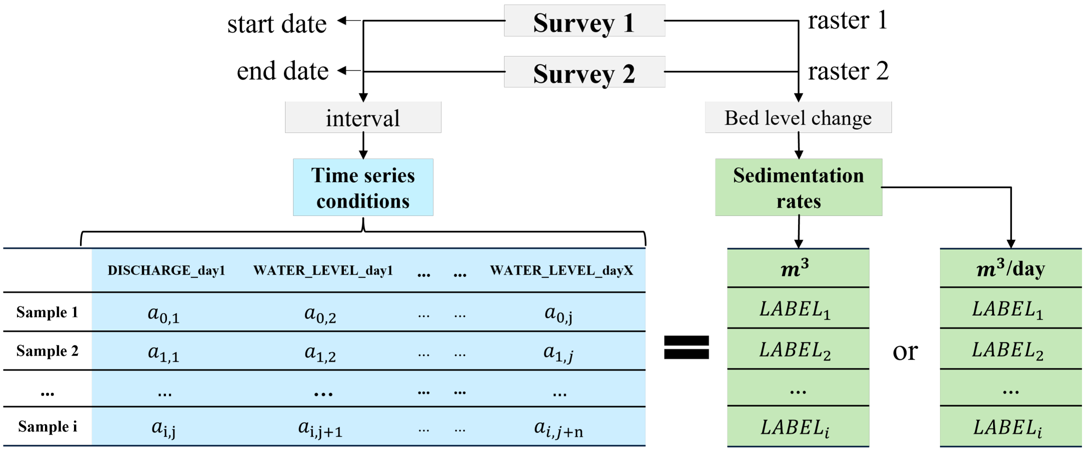
3.5.2. Sample creation

The literature review showed that SL ML models are trained with labelled data samples. These samples are not readily available prior to this research and must be constructed manually. The data available for the samples consists of the surveys, hydro-meteo variables, and logs on dredging operations discussed in this section. Of these three data types, surveying is the only one that grants insight into the morphological variation below the water surface. Therefore, the surveys will form the basis for a sample as they determine which period of conditions are taken as input and what the value of the dependent variable is. The period is to be determined by the previously mentioned interval between sequential surveys and, ideally, is constant. In this research, two sequential surveys at identical locations are called pairs. The sedimentation of a pair can be derived by rasterising the surveys, interpolating over the grid, and then taking the difference in bed level. The detailed reasoning behind this approach is discussed in the previous section.

To illustrate this approach, Botlek Centrale Geul is surveyed on January 1, 2024, and February 1, 2024, resulting in a start and end date. The conditions at the survey location between these dates, including the dredging volumes, are extracted from a dataset and ascribed to that pair. The rasterisation step provides the net sedimentation in m^3 in this period, resulting in a labelled sample. SR in m^3/day is determined by dividing the difference between the net sedimentation and the dredging volumes by the time between the survey pairs. In this case, that time is 31 days. This choice is explained in the next subsection. The resulting equation is:

$$SR = \frac{\Delta Sedimentation - \Delta Dredging}{Time\ interval} \tag{3.1}$$

These steps are taken for all pairs until a complete set is created, as shown in Figure 3.8. Since most hydro-meteo variables are measured at multiple locations, the variables of the measurement location closest to the dredging area are selected. Table B.2 shows which measurement locations are ascribed to which dredging area.



	DISCHARGE_day1	WATER_LEVEL_day1	...	...	WATER_LEVEL_dayX
Sample 1	$a_{0,1}$	$a_{0,2}$	...	...	$a_{0,j}$
Sample 2	$a_{1,1}$	$a_{1,2}$	...	...	$a_{1,j}$
...	...	...	...	...	...
Sample i	$a_{i,j}$	$a_{i,j+1}$	...	...	$a_{i,j+n}$

m^3
$LABEL_1$
$LABEL_2$
...
$LABEL_i$

m^3/day
$LABEL_1$
$LABEL_2$
...
$LABEL_i$

Figure 3.8: Overview of the method used to create samples.

Section 2.2 showed that RFR models require the samples to be shaped as a single row in a Pandas DataFrame when time series are considered. For SVR, the input must be shaped as an array. Therefore, the time series between two sequential surveys must be transposed into lagged variables to allow the algorithms to capture temporal dependencies. This is shown by including the day number in the column names in Figure 3.8. The resulting table is similar to the one Mussumeci and Codeço Coelho (2020) implemented in Figure 2.4.

Interval selection

It has already been mentioned that the interval between surveys influences the number of features per sample. For every extra day in between surveys, an extra feature per hydro-meteo variable is added. This is visible in Figure 3.8 where the last entry in the time series is denoted with *'_dayX'*. The time interval between sequential surveys is different for every pair but SVR and RFR require samples of constant length. That is why a suitable interval value must be determined.

Figure 3.9 shows the interval distribution between sequential surveys. The median and peak at 32.5 days show that the surveys (Table 3.1) taken at the high-priority areas from Section 3.1.1 are the dominant factor in determining this interval. The problem opposed by selecting a constant interval is represented by the second peak in Figure 3.9. This peak represents the 111 surveys with an average spacing of 58.7 days. Therefore, selecting the median interval will remove approximately 25 days of conditions for a substantial portion of the survey pairs. The interval length is seen as an influential variable, and setting the fitting value is part of the model development in Section 3.6.

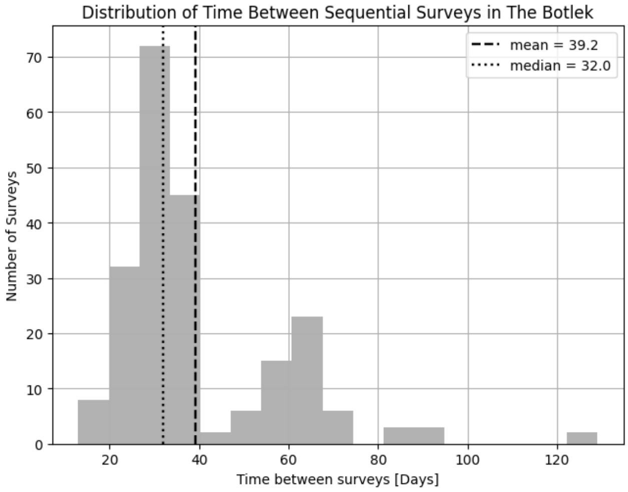
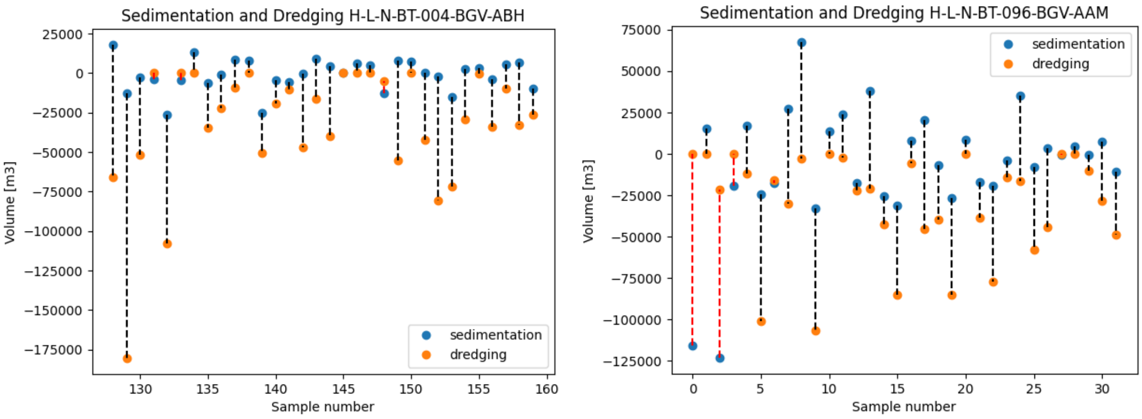


Figure 3.9: The distribution of the survey intervals of all dredging areas in the Botlek.

The first set of samples is created by assuming a mean survey interval of 31 days. This interval is close to the average of the three main surveying areas from Table 3.1 and aligns with the peak in Figure 3.9. This approach results in a total number of 205 samples. It is evident that 205 samples is lower than the 290 surveys available. The 1e Werkhaven, 2e Werkhaven, and Welplaathaven are left out of the first sample set because the average survey intervals for these areas are 50, 64, and 213 days, respectively. Including these surveys in the sample set will skew the assumed interval of 30 days for Botlek Centrale Geul, Botlek Mond, Botlek Vak 3, and 3e Petroleumhaven. Botlek Vak 3 also has an average survey interval of 64 days but is considered more relevant in maintenance dredging, which is why it is included.

3.5.3. Sample and dredging data cleaning

The samples contain both the dredging and sedimentation volumes of the period of the sample. Returning to the example from the previous section, this means that all dredging operations in January 2024 at Botlek Centrale Geul are accumulated and added to the sample row. The dredging volumes are added as a negative value, given that sediment leaves the system. The dredging data is manually registered in the maintenance records. However, some dredging operations are not registered as these are not considered maintenance but infrastructural operations, meaning that, for example, the NGD is permanently changed or a channel is widened. This results in a misalignment between what the survey and the dredging records show. Figure 3.10 shows the sedimentation and dredging volume per sample, both in m^3. The orange dots denote the dredging volumes, and the blue dots denote the net sedimentation acquired from subtracting two surveys.



(a) Botlek Mond 1

(b) 3e Petroleumhaven North

Figure 3.10: Net sedimentation and dredging volumes from samples linked to the dredging areas Botlek Mond 1 (3.10a) and 3e Petroleumhaven North (3.10b). The x-axis provides the sample number. The sedimentation values are the net bed level changes between the sequential surveys in a sample pair in m^3. The dredged volume is the amount of sediment dredged in between the surveys in m^3. A red line indicates if the sedimentation is lower than the dredged volume.

The dredging data has not been thoroughly analysed before the sample creation because the sedimentation acquired from the surveys was needed to provide context. According to the Asset Management department of the POR, sedimentation cannot be significantly lower than dredging volume as this implies unrealistic erosion. In these cases, the dashed lines are red to indicate a possible inaccuracy. Figure 3.10a shows the samples at the Botlek Mond 1 dredging area. Almost all samples have higher sedimentation than dredging, which is good. However, the second sample (sample number 129) contains a high dredging volume with only a moderate negative value for sedimentation. This can be ascribed to the deepening operations being finished around that time. The polygon from Botlek Mond 1 cuts off before het Scheur, thus not showing the significant change in depth due to dredging (see FigureB.21a). Figure 3.10b shows the results for 3e Petroleumhaven North. Here, the first and third samples show a different behaviour because the sedimentation is significantly lower than the dredged volume. In these cases, all sediment around a cable was removed (see Figure B.21b) but not registered as a maintenance operation, leading to a mismatch.

The same plots for all dredging areas are shown in Figure B.22 and B.23. The same outlier analysis as described above was possible for all plots given that the total number of samples is limited to 205. Only outliers that showed a significant difference between the SR and the dredged volume were removed. This resulted in a remaining total of 181 samples. Among these, there are still samples that imply 'erosion' with a negative SR.

3.5.4. Dredging areas in sample set

This section provides a concise overview of the characteristics of the dredging areas used in the sample set. The goal of this overview is to provide better context to the results from Chapter 4. From now on in this report, the dredging areas will be mentioned by their respective area codes. The general areas only have acronyms if they contain a measurement station (see Figure 3.4. Figure 3.11 visualises which code belongs to which area. ACM, ABF, and AFO make up Botlek Vak 3. Botlek Centrale Geul Oost (BOTCGO) is referred to as ABG. AAO and AAM are situated in 3e Petroleumhaven and ABH and ABK in BOTM.

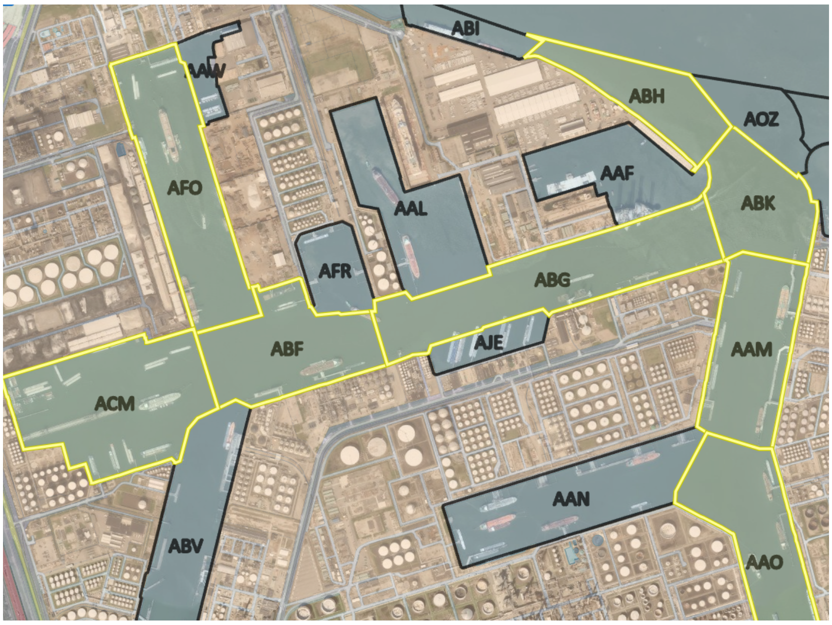


Figure 3.11: The yellow-bordered Dredging areas are included in the sample set.

Table 3.5 displays factors that contextualise the areas. The areas are ranked from highest mean SR ($\overline{SR}$) to lowest. The means are calculated with the SR values of the sample set. Generally, the areas closest to the Botlek entrance experience the highest $\overline{SR}$ with the exception of ABF in the central channel. ABF even outranks ABG while ABG is closer to the entrance. This inconsistency in SR for ABG is caused by the missing dredging data seen in Figure B.22d and explained in Section 5.2.

Table 3.5: Dredging areas ranked from highest to lowest '$\overline{SR}$'. '$\overline{SR}$' is the mean SR in m^3/day of the samples of that area. 'Mean dredged volume' is the mean volume dredged in the samples of the specific area in m^3. 'Distance to entrance' is the euclidean distance from the center of the dredging area to the center of the Nieuwe Maas channel in front of the Botlek entrance. The 'Surface area' column shows the area of the polygons from Figure B.20 in m^2 and 'Number of samples' the number of area specific samples in the total set of 181 samples.

Dredging area	$\overline{SR}$	Mean dredged Volume	Distance to Entrance	Surface area dredging polygon	Number of samples
AAM	1226.3	-33680.7	1180	160439	27
ABH	959.37	-26592.03	495	85279	30
ABF	772.4	-16313.9	2180	200583	13
ABG	673.61	-1501.3	1300	239829	27
ABK	590.69	-155531.2	577	125157	31
AAO	519.2	-11110.8	1730	87039	27
AFO	461.49	-6511.2	2440	217689	13
ACM	307.9	-3898.08	3010	124499	13

Conditional variation

Aside from the information in Table 3.5, it is challenging to contextualise the differences between the areas. Figure 3.4 and Table 3.2 already showed that the hydro-meteo variables are not area specific. Therefore, investigating the differences in slack water, stream direction, or variation in salinity across all areas in the Botlek is not possible, mainly because the 3e Petroleumhaven is not represented in the OSR or measurements. However, the measurement stations BOTCGW, 2WERKH, and BOTM do provide information on the variation in hydro-meteo conditions between some areas. For example, Figure 3.12 shows the fluctuation of salinity in the Botlek. 2WERKH has a slightly higher salinity than BOTCGW for the majority of the time. This is logical given its closer proximity to the entrance.

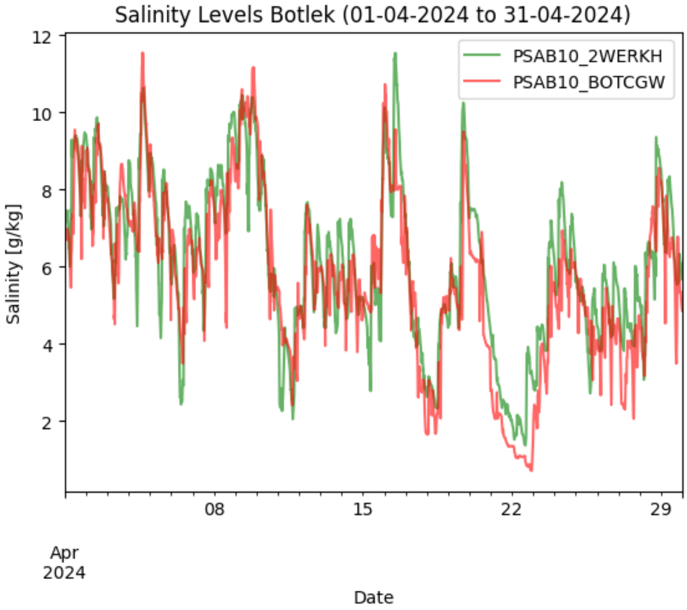
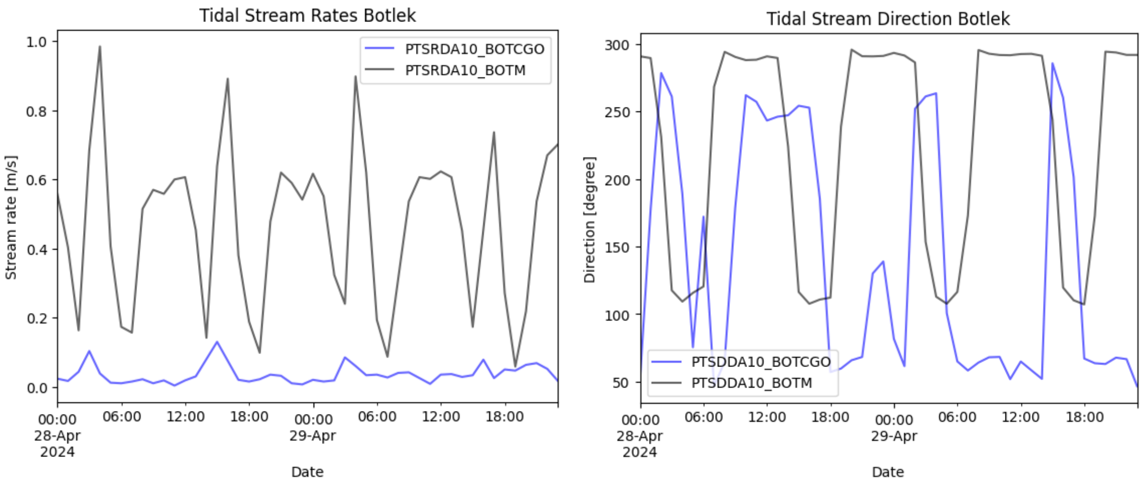


Figure 3.12: Salinity levels in April 2024 in the two areas of the Botlek. BOTCGW is located in ACM while 2WERKH is closer to the entrance in AAL (see Figure 3.11).

A more evident difference in conditions across the Botlek is noticeable in the tidal stream rate and direction in Figure 3.13. The stream rates at the entrance (BOTM) are significantly higher than the rates at the channel (BOTCGO), indicating a dissipation of tidal energy as the flow propagates into the basin. The cause of this dissipation likely is the bifurcation of the entrance flow into BOTCGO and the 3e Petroleumhaven. Moreover, there is a delay in flow reversal as BOTCGO has a lagged response compared to BOTM.



(a) Depth averaged stream rate [m/s]

(b) Depth averaged tidal stream direction [degree]

Figure 3.13: A visualisation of the fluctuation in tidal stream rates (3.13a) and direction (3.13b) in two areas of the Botlek. BOTCGO is located in ABG and BOTM in ABK.

3.6. Model engineering

The model engineering process aims to develop RFR and SVR models that can accurately predict SR. This process consists of feature engineering, hyperparameter tuning, and performance evaluation. Section 2.1 mentions that the common practice train-test split is 80% train and 20% test data. This split will be the standard during the entire modelling process.

3.6.1. Feature engineering

Input correlation and feature importance

Figure 3.8 shows the variables discharge and water level as examples. What subset of variables will result in the most accurate predictions is unknown and is a vital part of this research. Section 2.3 already provided tidal variation, salinity, and fluvial discharge as the dominant variables for sedimentation in the Botlek. The volume that is dredged in between surveys could also be influential, as removing sediment has a direct impact on the bed level. However, whether the developed models will perform optimally with this set of variables as input cannot be assumed and must be shown by performing model runs with different subsets. These subsets can be selected based on expected performance through literature, feature importance analysis provided by RFR, or a combination of both. The variable selection determines the length of a sample as every chosen parameter includes all lagged time series values of that parameter. A single entry in a sample row is called a feature, and reducing the number of features is generally favourable when training ML models (see Section 3.3). Therefore, analysing the importance of each parameter and its correlation to the dependent variable is an essential step in model development. It can show which features are redundant and assist in selecting smaller, more efficient subsets of parameters that could improve model performance. This process is mostly an iterative process during model development. However, the correlation matrix in Figure 3.14 already provides insight into which variables might be redundant due to a high mutual correlation.

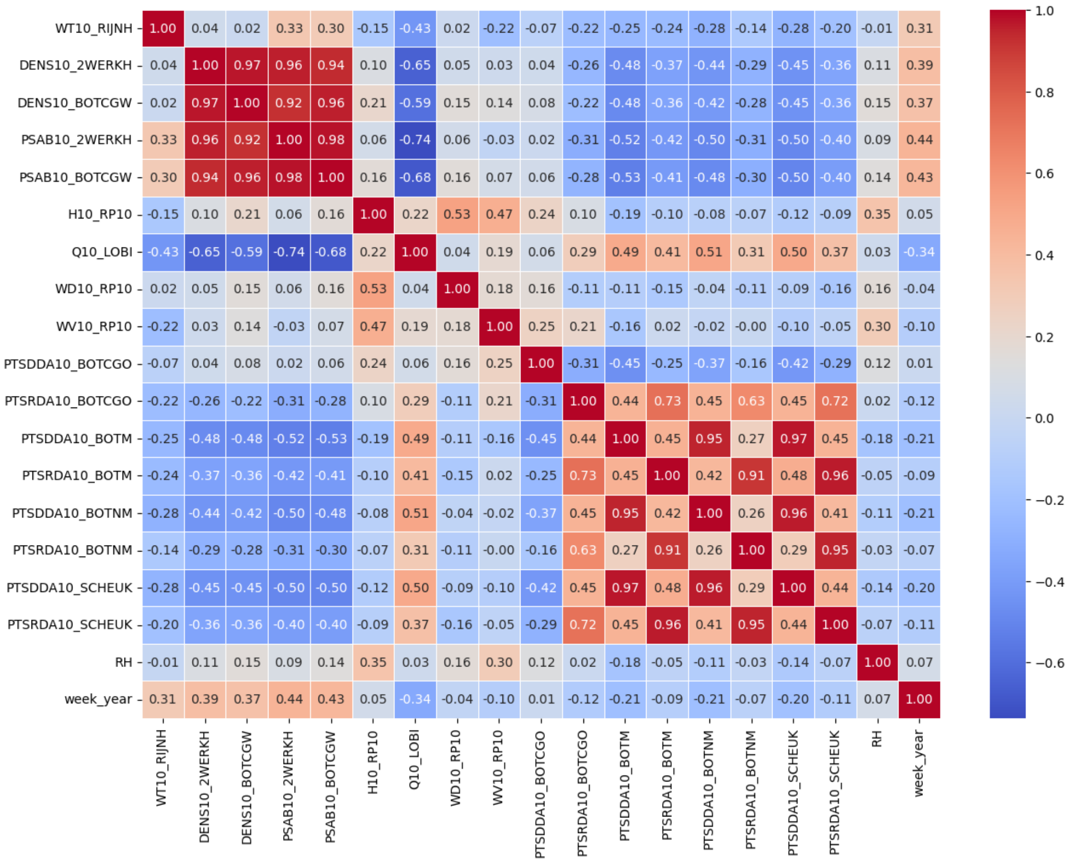


Figure 3.14: Pearson correlation matrix of all hydro-meteo variables

Depending on the measurement station, DENS10 and PSAB10 have a correlation ranging from 0.92 to 0.98. These high values are logical given that DENS10 was calculated with the GSW TEOS-10 package (TEOS-10 Developers, 2024) that converts water temperature (C°) and salinity (g/kg) into density (kg/m^3). Choosing DENS10 or PSAB10 helps reduce the number of features without losing significant information. Another noticeable aspect of the matrix is the strong negative correlation between Q10 and DENS10/PSAB10. The apparent explanation is that a higher freshwater discharge means less salt in the Botlek. The same can be said for the correlation between the PTSDDA10/PTSRDA10 and the DENS10/PSAB10 variables. The direction and stream velocity of the tide affect the salinity levels, bringing either fresh or saltwater inside the Botlek basin. The mutual high correlation between the different PTSRDA10 stations does not necessarily mean that only one station must be chosen as different stations are used for different samples (see Section 3.5.2).

Averaging

The smallest possible measurement frequency of all variables, except RH10, is 10 minutes. It is already discussed that this frequency results in too many features. A method to solve this is taking the average of that variable over a certain period. Whether that should be, for example, a daily or monthly average is unclear as the effect on the model performance must be determined. The downside of averaging is the significant loss of information. A daily average provides a reliable indication of the value of most variables like Q10, WT10, or WV10, but the longer the period becomes, the more detail is averaged out. Moreover, the monthly average of harmonic tidal movements results in a mean water level and does not show spring or neap tides. Reducing the number of features or preventing information loss is an important trade-off. The effect of this trade-off on the model performance needs to be analysed during development.

3.6.2. Hyperparameter tuning

The Scikit-learn (sklearn) library (Pedregosa et al., 2011) contains both RFR and SVR. Therefore, the model development will be done using sklearn. The hyperparameters in this report are denoted as defined by sklearn.

Parameter grids

Sklearn offers functions that make the overall tuning process less complex and time-consuming. Two of these functions are GridSearchCV and RandomizedSearchCV. The functions require a parameter grid containing the relevant hyperparameters of the chosen estimator, in this case, RFR and SVR. SearchCV then applies these parameters and finds the best combination possible by fitting on the training set and finding the combination that results in the highest cross-validation score. RandomSearchCV is usually less computationally expansive because it randomly samples from the grid. It is a reasonable starting point when there is no indication of the range of suitable hyperparameters. GridSearchCV tries all combinations, making it more suitable when the parameter space is smaller. It is often beneficial to start with the random approach to narrow the search space and then apply a grid search on a more detailed grid to investigate if that improves performance (Holihah, 2023).

Random states

The random state parameter controls the shuffling applied to a dataset before it is split. Therefore, choosing the same value will result in a reproducible output. This is relevant during development because it allows the application of different algorithms with different parameter grids on the same data split. Another angle is maintaining the same grid but changing the random state to show how a model performs on different subsets of the data.

Analysing the performance of different random states will be part of the model training. The random state should not be optimised. Therefore, it is not tuned like the hyperparameters. However, this approach does provide a comprehensive view of the average performance. The number of runs on different random states is limited by the length of the training time. For example, when the daily averages of all hydro-meteo variables are chosen as input, the runtime could be much longer than when the monthly average of a few variables is chosen.

RFR hyperparameters

The decision trees in an RFR repeatedly split the dataset into smaller datasets that reduce the variance in the dependent variable. The decisions to split the dataset are made in internal nodes, starting at the root node. When a split will not reduce the variance further, the internal node will not split and become a leaf node. The node is then called pure. The following hyperparameters influence this process:

- *N_estimators*: number of decision trees.
- *Max_depth*: the longest path between the root node and leaf node. When max_depth is not specified, the tree grows until all leaf nodes are pure, which can lead to overfitting.
- *Min_samples_split*: minimum number of samples required to split a node. When the value is too low, the tree splits the nodes until all nodes are pure.
- *Min_samples_leaf*: minimum number of samples required at the leaf node. By setting a higher value, only the final nodes with the minimum number of samples are considered leaf nodes.

- *Max_features*: number of features to consider when looking for the best split. Max_features can be defined by sqrt(n_features), log2(n_features), or None. In the case of None, max_features = n_features.

SVR hyperparameters

SVR try to find an optimal hyperplane that fits all training samples. It does so by finding a function as flat (simple) as possible while allowing a certain margin of error on the training data. The margin of error is called ϵ (epsilon). By ignoring values within that certain threshold, the algorithm is robust to outliers(MathWorks, 2024). SVR can handle nonlinear data by using different kernel functions. The following hyperparameters will be considered:

- *Kernel*: Specifies the kernel type for the algorithm. The default is a radial basis function (RBF). The other possibilities are sigmoid, precomputed, linear, or polynomial (poly).
- *Gamma*: Kernel coefficient for ’rbf’, ’poly’, and ’sigmoid’. It controls the influence of individual data points and determines how flexible the fit of the model is.
- *Epsilon*: Specifies the epsilon-tube. The tube is a margin of tolerance where no penalty is given for errors.
- *C*: Determines the penalty for errors larger than the epsilon margin. It is a regularisation parameter that controls the trade-off between minimising the prediction error and maximising the margin. A lower C means a wider margin that allows more errors, resulting in a simpler model, while a larger C leads to a complex model.

Below is the default RBF as defined by Anisa et al. (2024), with γ (*Gamma)* as kernel coefficient and $\|x_i - x_j\|^2$ the eucledian distance between two feature vectors:

$$K(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2) \tag{3.2}$$

Performance indicators

Hyperparameter tuning is not possible without an indication of the performance of a certain setting. Section 2.1 already mentioned common practice indicators. Their functions can be found there. For this study, the following indicators are chosen:

- R^2: the coefficient of determination. It measures the total variance in the dependent variable explained by the independent variables. It indicates how well the model captures the trends within the data.
- $RMSE$: the square root of the average squared error. $RMSE$ is more interpretable than MSE because it has the same unit as the dependent variable.
- MSE: the average squared error. Because the residuals are squared, MSE emphasises larger errors.

The evaluation will primarily be done on the R^2 and $RMSE$ scores. The MSE is added to provide extra context on the outlier errors.

3.6.3. Modelling phases

This section describes the phases undertaken to develop accurate LR, SVR and RFR prediction models. The model runs will be performed on the sample set with erosion samples and without, to investigate the impact of removing the erosion samples. Moreover, the models are trained using two dependent variables: SR in m^3 or m^3/day. The first unit is more practical for the POR. The second unit normalises the dependent variable which might result in better predictions. The phases build upon each other to refine the performance.

Phase 1.0: model training on dredging data only

In the first phase, ’Phase 1.0’, The LR, RFR, and SVR will be trained on only ’hoeveelheid_total’ and ’location_number’ first to illustrate whether adding hydro-meteo variables results in a better performance. Hoeveelheid_total is the name code for the total dredged volume in a sample. The ’location_number’ is a number ascribed to a certain dredging area. The LR and SVR will be run on three random states. The RFR will only run the best performing configuration due to runtime. These runs are performed for both dependent variables with all samples included and without the erosion samples. This phase already gives a clear indication on which dependent variable results in the best performance.

Phase 1.1 to 1.3: model training with hydro-meteo variables

After Phase 1.0, the hydro-meteo variables are added. All three algorithms will be trained on all features but with different time scales. All features means all lagged values of all hydro-meteo variables from Table 3.2, ’hoeveelheid_total’, and ’location_number’. By selecting daily, weekly, and monthly averages as features, the sample length for the three scenarios will differ by 302, 52, and 12 features, respectively. It is complex to estimate whether decreasing the number of features but losing information by averaging has a positive or negative impact on the performance. For this reason, all three sample types are used.

’Phase 1.1’ refers to the runs where the dependent variable is the total aggregated volume in m^3, as this is the most usable unit for the POR. In ’Phase 1.2’, SR is in m^3/day because this normalises the dependent variable. The runs will be performed with random_state set to 0, 20, and 42 to provide a general performance indication. This approach results in 54 runs, 18 for LR, RFR, and SVR each. The hyperparameter grids do not change during the different runs, and the grid search is done using RandomSearchCV. The grids for both SVR and RFR can be found in Table C.1 and C.2.

Phase 1.1 and 1.2 are run with all samples, including the ones containing erosion (see Figure 3.10), to investigate the impact of using these samples. After establishing the performance on all samples, ’Phase 1.3’ will rerun the SVR without the erosion samples (142) with either the SR in m^3 or m^3/day, depending on which unit provided the best performances. The runs in Phase 1.3 are first done with SVR and LR because preliminary runs showed that the runtimes were a maximum of a few seconds, whereas the RFR could take 30-40 minutes. RFR will also be rerun if the models from Phase 1.3 significantly outperforms 1.1 and 1.2.

Every RFR run provides new feature importance scores because of the difference in random state shuffling and length of the samples. All scores will be gathered and transformed into easy-to-analyse plots and overviews. These plots will help substantiate the choices in cutting certain features from the feature space in the following phases. The end of Phase 1 will be an analysis of the performance indicators, the hyperparameters chosen by RandomSearchCV, the feature importance scores, and the actual predicted SR values from the best and worst-performing models.

3.6.4. Conclusion data processing and model engineering

Sections 3.5 and 3.6 describe how the data and LR, RFR, and SVR models must be engineered to predict SR. Therefore, the third sub-question is answered:

How can the selected ML algorithms and features be configured to predict SR?

The development of the ML models starts with creating samples that align the dredging logs, survey data, and hydro-meteo variables. Each sample consists of the lagged time series of the hydro-meteo conditions between two sequential surveys at the same dredging area. The samples need to have a constant length. Therefore, a length of 31 days of conditions is chosen. This coincides with the surveying interval of the high-priority dredging areas. The dredged volume between these surveys shows how much sediment leaves the system. The label is the net bed level change in m^3 over two surveys. There are three sample types: daily, weekly, and monthly means of the hydro-meteo conditions. The number of features decreases each time scale from 302 to 52 to 12, respectively. Additionally, two dependent variables are considered during modelling. The first is SR in m^3 as this is the most practical unit for the POR. The second is SR in m^3/day because normalising the dependent variable is expected to improve model performance. Finally, the decision is made to train the ML models on sample sets with and without erosion. The many possible configurations are structured into phases:

1. **Phase 1.0**: The models are trained on dredging data only
2. **Phase 1.1 - 1.3**: The hydro-meteo variables are introduced while the different units of SR are tested.

The phases are summarised in Table 3.6. The performance of the configurations is tested over multiple random states to provide a more conclusive indication of the ML model performance.

Table 3.6: Summary of the ML training phases. The columns indicate what is included in each specific phase. ’Dependent variable’ in Phase 1.0 contains both units, while Phase 1.3 selects the best-performing unit of all preceding phases. The (RFR) and (+8) in the Phase 1.3 row are in brackets because the RFR is only included if the configurations from Phase 1.3 result in a better performance than the other phases.

	Algorithms	Hydro-meteo variables	Dependent variable	Random states	Number of Samples	Sample type(s)	Total runs
Phase 1.0	LR, RFR, SVR	hoeveelheid_total + location_number	m^3 , m^3/day	0, 20, 42	181 (with erosion), 142 (without erosion)	daily, weekly, monthly	27
Phase 1.1	LR, RFR, SVR	All	m^3	0, 20, 42	181 (with erosion)	daily, weekly, monthly	27
Phase 1.2	LR, RFR, SVR	All	m^3/day	0, 20, 42	181 (with eosion)	daily, weekly, monthly	27
Phase 1.3	LR, SVR, (RFR)	All	Best performing	0, 20, 42, 60	142 (without erosion)	daily, weekly, monthly	24 (+8)

4 Results

Chapter 4 covers the results provided by following the steps in the methodology. By doing so, it answers the following sub-question:

How do the selected ML algorithms perform across different configurations?

The configurations refer to the iterative modelling phases described in Section 3.6.3. This section explains that Phase 1.1 refers to the runs with m^3 as the dependent variable and Phase 1.2 to m^3/day. The mean performance over three random states and two dependent variables will indicate what feature and hyperparameter spaces can improve performance. Table 4.1 and 4.2 show the difference between the datasets for both independent variables.

Table 4.1: Dataset statistics of the test set where the dependent variable has unit m^3. The '$\tilde{SR}$' represents the median of the test set.

random state	$\tilde{SR}$	SR_{max}	SR_{min}
0	-2149.13	67481.52	-33025.22
20	-1657.42	18356.05	-35247.27
42	2380.04	46426.87	-35247.27

Table 4.2: Dataset statistics of the test sets where the dependent variable has unit m^3/day.

random state	$\tilde{SR}$	SR_{max}	SR_{min}
0	342.13	2612.67	-851.355
20	412.78	1880.91	-1174.91
42	331.57	1965.06	-1174.91

The large SR_{min} values in Table 4.1 indicate that samples still have a negative SR, meaning erosion. Phase 1.3 will rerun the best-performing dependent variable without these erosion samples. Phase 1.0 runs LR and SVR models on the same configurations as the phases described above but without the hydro-meteo variables and only on the dredging volumes. The results from this phase provide excellent context in determining if adding the hydro-meteo variables improves the performance.

The leading performance indicator is R^2, given that $RMSE$ and MSE depend on the different random states and dependent variables because these result in data splits with different statistics. The results for the RFR and SVR will contain plots with the predicted values, as this helps contextualise the impact of the hyperparameters. LR models do not have tuning, so only the LR results that have a similar or better performance than the other RFR or SVR runs will be shown in this chapter.

4.1. Phase 1.0: only dredging data in feature space

This section covers the runs of the LR, RFR, and SVR that are only trained on the dredging volumes (hoeveelheid_total) and a location number (location_number). Table 4.3 displays the mean performances of the runs where the dependent variable was SR in m^3. The individual runs can be found in Appendix C.1.1. None of the runs scores high on any of the performance indicators which means that the models are not usable in a practical application. A useful takeaway is that removing the erosion samples does improve the performance.

Table 4.3: Mean performance of LR and SVR models from Phase 1.0 with SR in m^3. This table summarises the tables in Appendix C.1.1.

Model	**Erosion Samples**	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
LR	Yes	0.05±0.10	$1.49*10^4 \pm 1.83*10^3$	$2.25*10^8 \pm 5.24*10^7$
SVR	Yes	0.07±0.08	$1.48*10^4 \pm 2.05*10^3$	$2.23*10^8 \pm 5.77*10^7$
LR	No	0.14±0.13	$1.22*10^4 \pm 1.17*10^3$	$1.49*10^8 \pm 2.57*10^7$
SVR	No	0.17±0.12	$1.19*10^4 \pm 8.92*10^2$	$1.42*10^8 \pm 2.53*10^7$

Table 4.3 displays the mean performances of the runs where the dependent variable was SR in m^3/day. The individual runs can be found in Appendix C.1.2. The performance improves significantly when the dependent variable is normalised. Especially the LR and SVR runs without the erosion samples are noticeable.

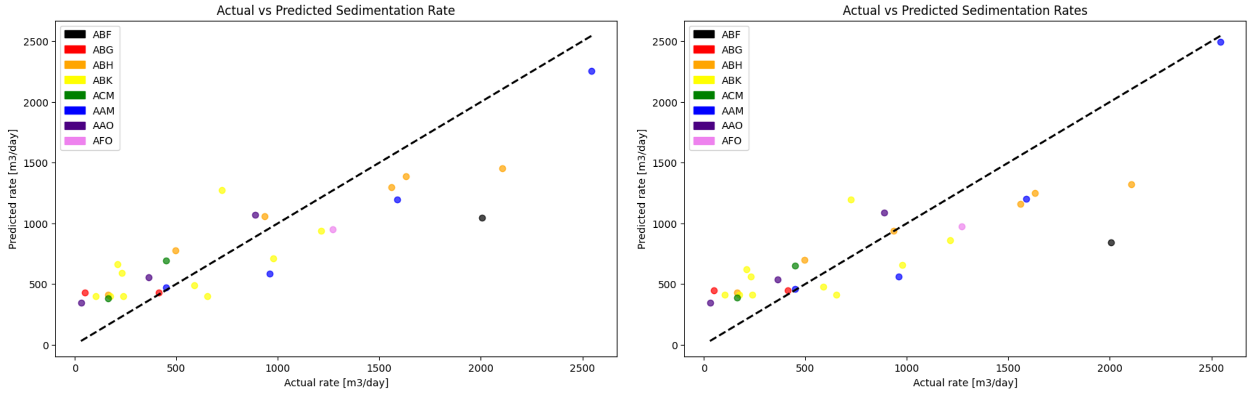
Table 4.4: Mean performance of LR, RFR, and SVR models from Phase 1.0 with SR in m^3/day. This table summarises the tables in Appendix C.1.2.

Model	Erosion Samples	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
LR	Yes	0.47±0.14	496.61±55.46	$2.17{*}10^5 \pm 5.31{*}10^4$
SVR	Yes	0.47±0.16	496.00±62.13	$2.17{*}10^5 \pm 6.23{*}10^4$
LR	No	0.65±0.06	405.76±42.27	$1.65{*}10^5 \pm 3.10{*}10^4$
SVR	No	0.65±0.02	407.95±29.27	$1.66{*}10^5 \pm 2.38{*}10^4$
RFR	No	0.55±0.06	460.08±19.00	$2.12{*}10^5 \pm 1.68{*}10^4$

Table 4.5 displays the best individual runs per algorithm from Phase 1.0. Given the variety in results for different random states, the mean performances are a better indication. Nonetheless, the LR and SVR show that the best performing models are already able to produce usable indications on SR.

Table 4.5: Best LR, SVR, and RFR models from Phase 1.0.

Algorithm	R^2	$RMSE$	MSE
LR	0.73	350.90	$1.23{*}10^5$
RFR	0.59	467.31	$2.18{*}10^5$
SVR	0.68	379.81	$1.44{*}10^5$



(a) Best performance LR ($R^2 = 0.73$)

(b) Best performance SVR($R^2 = 0.68$)

Figure 4.1: Scatter plot of the predictions of the best performing LR (4.1a) and SVR (4.1b) models from Table 4.5 plotted against the actual SR from the test set. Better model performance means that the points are clustered near the Identity Line (IL). The colour codes indicate the dredging area for which the model predicts the sedimentation and refer to the last three letters from the area codes from Figure B.20.

4.2. Phase 1.1: dependent variable in m^3

4.2.1. Results RFR

Table 4.6 displays the mean performances of the RFR models for all three sample types. The individual performances are in Appendix C.2. Sample type refers to the averaging interval and determines the number of features. This is explained in more detail in Section 3.6.3. The $\overline{R^2}$ scores are underwhelming for all sample types and show a relatively large standard deviation, indicating an inconsistent performance. The weekly runs outperform daily and monthly on all three indicators but show a more significant standard deviation on $\overline{RMSE}$ and $\overline{MSE}$.

Table 4.6: Summary of the performance of RFR models from Phase 1.1. This table summarises the RFR tables in Appendix C.2. 'Sample type' refers to the averaging interval and determines the number of features as explained in Section 3.6.3. The overbar indicates that these are the mean scores per sample type. Each entry represents the mean of three separate runs. The standard deviation is included as well.

Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Daily	0.13 ± 0.2	$1.40*10^4 \pm 7.55*10^2$	$1.97*10^8 \pm 2.12*10^8$
Weekly	0.18 ± 0.16	$1.37*10^4 \pm 1.18*10^3$	$1.89*10^8 \pm 2.92*10^7$
Monthly	0.09 ± 0.31	$1.43*10^4 \pm 8.79*10^2$	$2.03*10^8 \pm 2.75*10^7$

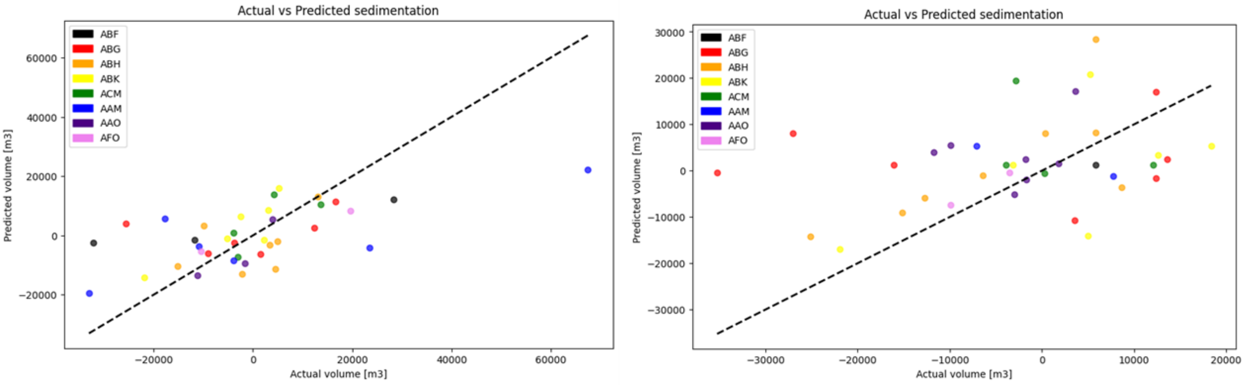
Another observation from Table 4.6 is that, in this case, the mean performance does not become better when fewer features are involved. The monthly runs have the worst mean performance indicators out of all. This can be attributed to the fact that the highest and lowest scores came from the monthly runs. These are shown in Table 4.7. The difference between the two runs explains the high standard deviation and results in the lower mean.

The best RFR of this section has a max_depth of 50, allowing it to capture more complex relations, while the min_leaf of 6 helps avoid overfitting by preventing the trees from growing too deep. The worst run even ends with a negative R^2, meaning that a linear fitted line would predict sediment volumes better than the RFR. The min_leaf of 1 indicates that the model overfitted by allowing small tree splits.

Table 4.7: Best and worst performing RFR models from Phase 1.1.

Performance	R^2	$RMSE$	MSE	**n_est**	**min_split**	**min_leaf**	**max_depth**	**random_state**
Best	0.38	$1.40*10^4$	$1.97*10^8$	300	2	6	50	0
Worst	-0.23	$1.35*10^4$	$1.82*10^8$	450	2	1	10	20

Figure 4.2 shows the predictions from the best and worst models. The predictions from Figure 4.2b do not remotely follow the IL and are scattered around the graph. Especially the outliers from ABG significantly deviate. In Figure 4.2a, the predictions follow a pattern closer to the IL except for the AAM outlier and some negative samples. These outliers have a more substantial effect on the $RMSE$, especially when there is such a significant difference between the outliers and the mean sedimentation volumes.



(a) Best performance ($R^2 = 0.38$)

(b) Worst performance ($R^2 = -0.23$)

Figure 4.2: Scatter plot of the predictions of the best (4.2a) and worst (4.2b) performing RFR models from Table 4.7 plotted against the actual volume from the test set. Better model performance means that the points are clustered near the IL. The colour codes indicate the dredging area for which the model predicts the sedimentation and refer to the last three letters from the area codes from Figure B.20.

4.2.2. Results SVR

Table 4.8 displays the mean performances of the SVR models for all three sample types. The individual performances are in Appendix C.2. Again, the scores are underwhelming for all sample types. However, in Phase 1.1, the SVR models outperform the RFR and show improved consistency for $\overline{R^2}$. There is less difference in the performance of the three sample types, but the weekly runs still outperform daily and monthly, similar to the RFR.

Table 4.8: Summary of the performance of SVR models in Phase 1.1. This table summarises the SVR tables in Appendix C.2. The overbar indicates the mean scores per sample type. Each entry represents the mean of three separate runs. The standard deviation is included as well.

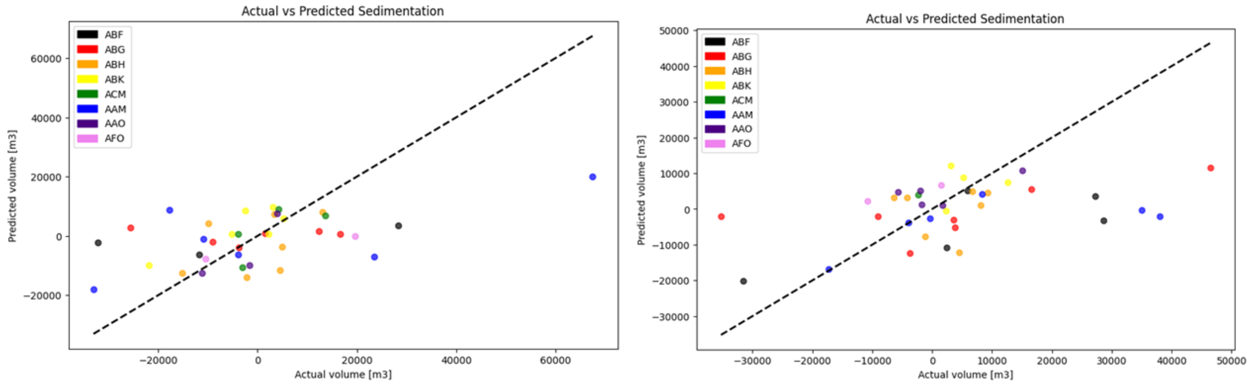
Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Daily	0.17±0.04	$1.40{*}10^4 \pm 2.22{*}10^3$	$1.99{*}10^8 \pm 5.34{*}10^7$
Weekly	0.22±0.05	$1.35{*}10^4 \pm 2.14{*}10^3$	$1.92{*}10^8 \pm 5.94{*}10^7$
Monthly	0.18±0.07	$1.39{*}10^4 \pm 2.53{*}10^3$	$1.94{*}10^8 \pm 5.88{*}10^7$

Table 4.8 shows a slight improvement with fewer features and this time, monthly is not ranked lowest. Still, weekly has more features and outperforms monthly. The best-performing model from Table 4.9 resulted from a weekly run, while the worst, again, resulted from a monthly. The difference in R^2 scores between RFR and SVR in Phase 1.1 is noticeable. RFR scores 0.10 higher on the best performance, while the gap between the worst SVR and RFR is 0.35 in favour of SVR.

Table 4.9: Best and worst performing SVR in Phase 1.1

Performance	R^2	$RMSE$	MSE	**C**	**epsilon**	**gamma**	**random_state**
Best	0.27	$1.52{*}10^4$	$2.32{*}10^8$	5	0.3	0.001	0
Worst	0.12	$1.52{*}10^4$	$2.23{*}10^8$	5	0.3	0.01	42

C and epsilon are equal for the best and worst SVR. Therefore, the gamma, random state, and sample type made a difference. The higher gamma for worst performance means that the SVR was less flexible in fitting and likely overfitted on the training data. The difference in predicted values between both SVR models in Figure 4.3 is not as straightforward as for the RFR models in Figure 4.2. The values around zero are clustered around the IL, but most values, especially the outliers, are still off. The best performances in Phase 1.1 occurred with the random state equal to 0. Comparing Figure 4.2a and 4.3a shows that SVR and RFR predicted the outliers from this split approximately the same.



(a) Best performance ($R^2 = 0.27$)

(b) Worst performance ($R^2 = 0.12$)

Figure 4.3: Scatter plot of the predictions of the best (4.3a) and worst (4.3b) performing SVR models from Table 4.9, plotted against the actual volume from the test set. Better model performance means that the points are clustered near the IL.

4.2.3. Results LR

The individual Phase 1.1 LR runs from Appendix C.2 are summarised in Table 4.10. LR was not able to capture any relations as the three performance indicators show even worse results that the SVR and RFR from this phase. For this reason, the LR results are not further elaborated in this section.

Table 4.10: Summary of the performance of LR models in Phase 1.1. The overbar indicates the mean scores per sample type. This table summarises the LR tables in Appendix C.2. Each entry represents the mean of three separate runs. The standard deviation is included as well.

Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Daily	-1.03±0.54	$2.14{*}10^4 \pm 9.79{*}10^2$	$4.57{*}10^8 \pm 3.68{*}10^7$
Weekly	-0.34±0.081	$1.72{*}10^4 \pm 2.08{*}10^3$	$2.98{*}10^8 \pm 6.05{*}10^7$
Monthly	0.063±0.24	$1.45{*}10^4 \pm 1.50{*}10^3$	$2.14{*}10^8 \pm 5.02{*}10^7$

4.3. Phase 1.2: dependent variable in m^3/day

4.3.1. Results RFR

Table 4.11 summarises the RFR performances for Phase 1.2. The individual performances are in Appendix C.2.1. Scaling the dependent variable from volume to SR improved the $\overline{R^2}$ compared to the RFR and SVR from Phase 1.1. The consistency did not improve much as the standard deviation is comparable. For $\overline{RMSE}$ and $\overline{MSE}$, the consistency of the Phase 1.2 RFR models is worse as the standard deviation range is relatively more extensive than those of Phase 1.1. The monthly runs resulted in the best mean performance. The errors are slightly larger than in the daily and weekly runs, but the standard deviation is much lower. Even though monthly performed best, the effect of decreasing the number of features on the RFR models is, again, not remarkable.

Table 4.11: Summary of the performance of RFR models in Phase 1.2. This table summarises the RFR tables in Appendix C.2.1. The overbar indicates the mean scores per sample type. Each entry represents the mean of three separate runs. The standard deviation is included as well.

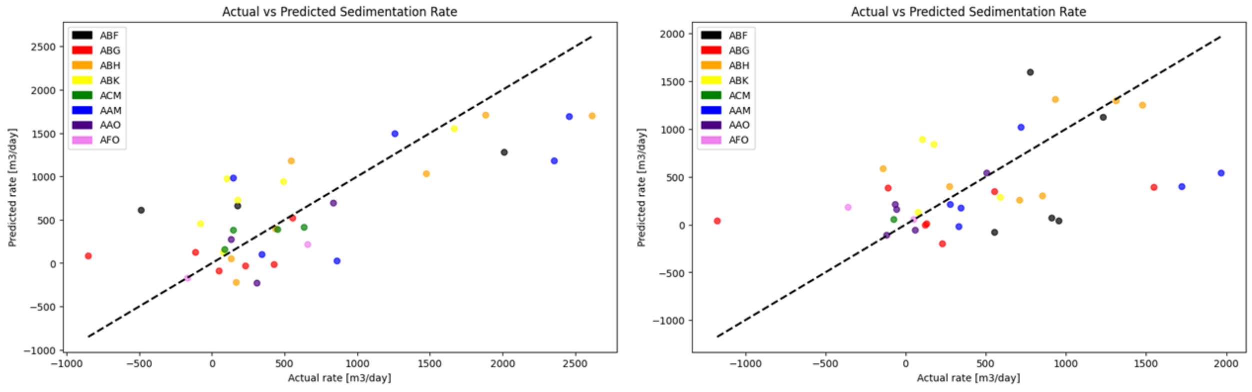
Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Daily	0.39±0.21	531.64±67.00	$2.85*10^5 \pm 5.68*10^4$
Weekly	0.39±0.21	514.94±72.82	$2.69*10^5 \pm 7.34*10^4$
Monthly	0.43±0.17	519.09±38.89	$2.71*10^5 \pm 3.90*10^4$

The performance increase for both Phase 1.2 RFR models from Table 4.12 compared to Table 4.7 is significant. A weekly run produced the best-performing model, slightly beating a monthly run on $RMSE$ and MSE (see Table C.26). The worst performance came from a daily run.

Table 4.12: Best and worst performing RFR models from Phase 1.2

Performance	R^2	$RMSE$	MSE	**n_est**	**min_split**	**min_leaf**	**max_depth**	**random_state**
Best	0.57	532.27	$2.83*10^5$	400	8	2	40	0
Worst	0.13	587.74	$3.45*10^5$	100	8	4	10	42

The higher number of decision trees allowed the best-performing model to generalise better as the output is averaged over more trees. The lack of depth in the worst-performing model likely prevented the RFR from capturing the more complex relations. The prediction difference is visible in Figure 4.4. The outliers in Figure 4.4a follow the IL much better than in Phase 1.1 and have similar residuals to the samples with a smaller SR. The more even spread shows that the model has a better generalisation. Both models struggle with predicting negative SR, as most predictions are positive. This is an essential observation as it indicates that the models cannot discover clear patterns in samples that contain erosion as the dependent variable, proving that these samples are unreliable.



(a) Best performance ($R^2 = 0.57$)

(b) Worst performance ($R^2 = 0.13$)

Figure 4.4: Scatter plot of the predictions of the best (4.4a) and worst (4.4b) performing RFR models from Table 4.12, plotted against the actual SR from the test set. Better model performance means that the points are clustered near the IL.

4.3.2. Results SVR

Table 4.13 summarises the SVR performances for Phase 1.2. All model runs can be found in Appendix C.2.1. The Phase 1.2 SVR models outperformed all others from Phase 1.1 and 1.2 so far. Phase 1.2 SVR did become less consistent than

Phase 1.1 regarding the $\overline{R^2}$. In return, the consistency in the errors became relatively better than those of the Phase 1.1 SVR (Table 4.8). The monthly runs show the most promising results, outperforming daily and weekly on all performance indicators.

Table 4.13: Summary of the performance of SVR models in Phase 1.2. This table summarises the SVR tables in Appendix C.2.1. The overbar indicates the mean scores per sample type. Each entry represents the mean of three separate runs. The standard deviation is included as well.

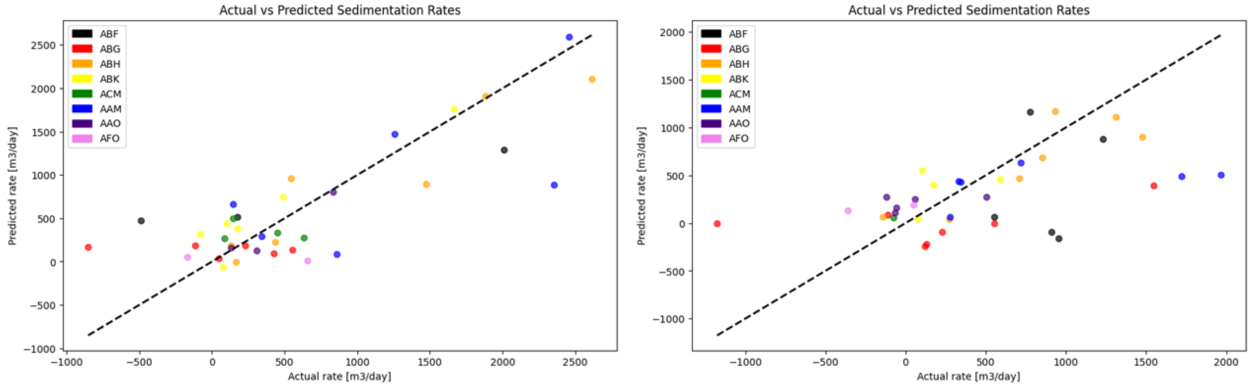
Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Daily	0.43±0.16	511.30±15.92	$2.62{*}10^5{\pm}1.36{*}10^4$
Weekly	0.48±0.16	490.98±38.99	$2.42{*}10^5{\pm}3.05{*}10^4$
Monthly	0.51±0.22	474.84±63.20	$2.29{*}10^5{\pm}7.34{*}10^4$

The best and worst performing models in Table 4.14 are unique in Phase 1 because both resulted from monthly runs, and the hyperparameters are the same. However, the runs were performed on different random states: 0 and 42. Looking at Tables 4.7, 4.9, and 4.12, it can be concluded that 0 always resulted in the best performance, both for Phase 1.1 and 1.2. The bias towards a specific random state should be investigated by taking a larger variety of random states. This provides a more generalised performance indication over the entire dataset.

Table 4.14: Best and worst performing SVR models from Phase 1.2

Performance	R^2	$RMSE$	MSE	C	epsilon	gamma	random_state
Best	0.67	468.21	$2.19{*}10^5$	50	0.001	0.001	0
Worst	0.23	553.95	$3.07{*}10^5$	50	0.001	0.001	42

Figure 4.5a shows the improved predictive accuracy. The points are clustered closer to the IL than all models before. The SVR can predict some outliers nearly perfectly compared to Figure 4.4a but still struggles with generalising. The grouping of the points around 0 to 500 m^3/day is also much closer to the IL. Again, the negative SR values are misinterpreted and predicted as positives or near zero.



(a) Best performance ($R^2 = 0.67$)

(b) Worst performance ($R^2 = 0.23$)

Figure 4.5: Scatter plot of the predictions of the best (4.5a) and worst (4.5b) performing SVR models from Table 4.12 plotted against the actual SR from the test set. Better model performance means that the points are clustered near the IL.

4.3.3. Results LR

Table 4.15 shows the same performance improvement as for the RFR and SVR in this phase. The individual runs can be found in Appendix C.2.1. The daily samples still prove to be difficult, even when the SR is normalised. The monthly runs are only just outperformed by the SVR from Table 4.13.

Table 4.15: Summary of the performance of LR models in Phase 1.2. This table summarises the LR tables in Appendix C.2.1. The overbar indicates the mean scores per sample type. Each entry represents the mean of three separate runs. The standard deviation is included as well.

Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Daily	-0.069±0.17	711.72±27.43	$5.08{*}10^5{\pm}4.72{*}10^4$
Weekly	0.34±0.25	551.21±63.38	$3.06{*}10^5{\pm}5.75{*}10^4$
Monthly	0.48±0.22	484.09±48.48	$2.38{*}10^5{\pm}4.97{*}10^4$

4.4. Phase 1.3: excluding erosion samples and SR in m^3/day

It is clear from the summarising tables in Section 4.2 and 4.3 that the performances from Phase 1.2 are better than Phase 1.1. Therefore, the runs in Phase 1.3 are performed with SR in unit m^3/day. An extra random state is added to provide a better general performance as Section 4.3 showed that a bias towards a specific random state can occur. Additionally, the parameter grid is slightly extended (Table C.30) because the short runtime of SVR allows running more iterations. The statistics of the random states used in this phase are shown in Table 4.16.

Table 4.16: Dataset statistics of test sets without the erosion samples.

random state	$\tilde{SR}$	SR_{max}	SR_{min}
0	452.28	2612.66	59.05
20	587.50	2544.97	30.366
42	570.15	2456.23	44.076
60	553.01	2544.97	8.31

4.4.1. Results SVR

Table 4.17 summarises the SVR performances for Phase 1.3. All model runs can be found in Appendix C.3. The results have improved significantly compared to Phase 1.2. Again, the monthly runs perform best on all three performance indicators. The standard deviation in $\overline{R^2}$ across the different sample types is relatively small compared to Phase 1.2, whereas the error deviation is comparable. Moreover, the results are better than the best baseline performances from LR and SVR in Phase 1.0.

Table 4.17: Summary of the performance of SVR models trained in Phase 1.3. This table summarises the SVR tables in Appendix C.3. The overbar indicates the mean scores per sample type. Each entry represents the mean of four separate runs. The standard deviation is included as well.

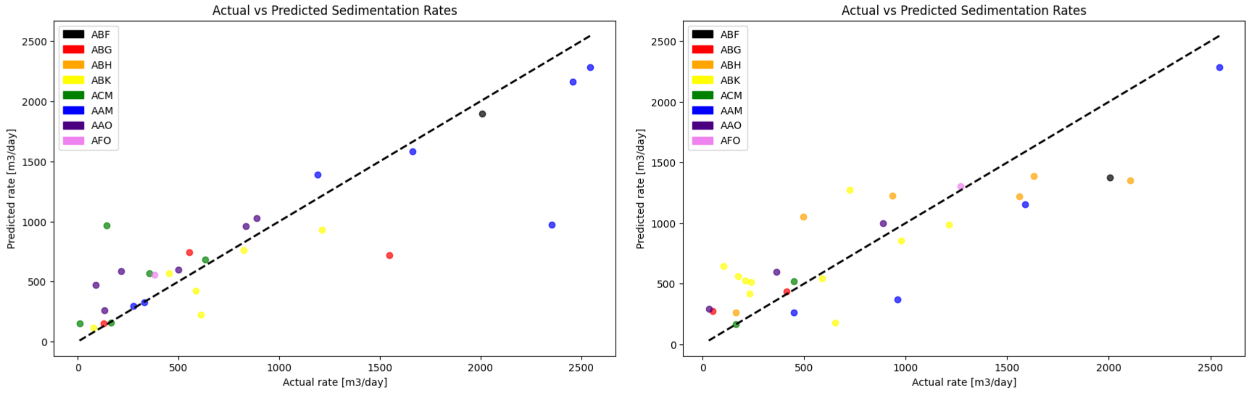
Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Daily	0.58±0.059	455.98±9.81	$2.08{*}10^5 \pm 8.38{*}10^3$
Weekly	0.63±0.10	426.93±52.47	$1.84{*}10^5 \pm 4.66{*}10^4$
Monthly	0.69±0.053	388.78±27.59	$1.52{*}10^5 \pm 2.35{*}10^4$

The two best-performing models are highlighted in Table 4.18 instead of focusing on the best and worst-performing models like the previous sections. Both models are produced by a monthly run. The reason for the focus on the best performing is the similarity in the R^2 scores but the difference in errors. The best-performing model stems from a random state equal to 60 and has a $RMSE$ 33.8 higher than the second-best model from a random state equal to 20.

Table 4.18: Best performing SVR models from Phase 1.3.

Performance	R^2	$RMSE$	MSE	**C**	**epsilon**	**gamma**	**random_state**
Best	0.74	382.91	$1.47{*}10^5$	1000	0.1	0.001	60
2nd Best	0.73	349.11	$1.22{*}10^5$	50	0.001	0.001	20

The prediction plots in Figure 4.6 visualise the difference in errors. Both SVR models show that the predictions are relatively close to the IL compared to the previous phases. The difference in $RMSE$ could be ascribed to the three severely mispredicted outliers in Figure 4.6a. Aside from these outliers, the models show to have generalised better than the previous phases.



(a) Best performance ($R^2 = 0.74$)

(b) Second best performance ($R^2 = 0.73$)

Figure 4.6: Scatter plot of the predictions of the best (4.5a) and second best (4.5b) performing SVR models from Table 4.18, plotted against the actual SR from the test sets. Better model performance means that the points are clustered near the IL.

4.4.2. Result RFR

Table 4.19 summarises the RFR performances for Phase 1.3. All model runs can be found in Appendix C.3. Daily runs are not performed here as the monthly and weekly performances already showed that the SVR from Phase 1.3 significantly outperforms the RFR and the daily RFR runs take approximately 30 minutes. The overall scores did improve compared to the RFR and SVR from Phase 1.2. The monthly runs showed the best mean performance.

Table 4.19: Summary of the performance of RFR models in Phase 1.3. This table summarises the RFR tables in Appendix C.3. The overbar indicates the mean scores per sample type. Each entry represents the mean of four separate runs. The standard deviation is included as well.

Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Weekly	0.51±0.07	489.36±18.08	$2.40{*}10^5 \pm 1.47{*}10^4$
Monthly	0.55±0.07	454.15±23.33	$2.07{*}10^5 \pm 2.43{*}10^4$

The best performing RFR in Table 4.20 outperforms its predecessors in Phase 1.1 and 1.2 but not the SVR from this and the previous phase. Figure 4.7 shows that the RFR still struggles with the outliers.

Table 4.20: Best and worst performing RFR models from Phase 1.3

Performance	R^2	$RMSE$	MSE	**n_est**	**min_split**	**min_leaf**	**max_depth**	**random_state**
Best	0.62	449.92	$2.02{*}10^5$	300	10	4	50	0

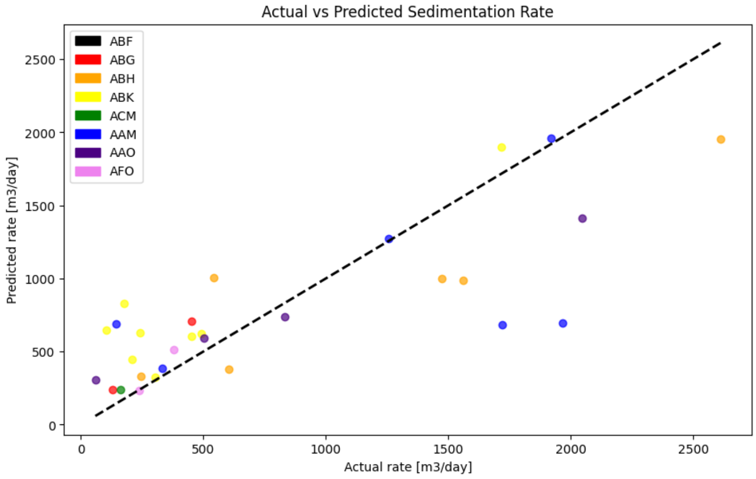


Figure 4.7: Scatter plot of the predictions of the best performing RFR model ($R^2 = 0.62$) from Table 4.20, plotted against the actual SR from the test sets. Better model performance means that the points are clustered near the IL.

4.4.3. Results LR

Table 4.21 summarises the LR performances of Phase 1.3. The individual model runs can be found in Appendix C.3. LR still struggles with larger number of features. The monthly runs are closer to the results from Phase 1.0 and the SVR from this phase.

Table 4.21: Summary of the performance of LR models in Phase 1.3. This table summarises the LR tables in Appendix C.3. The overbar indicates the mean scores per sample type. Each entry represents the mean of four separate runs, and the standard deviation is included as well.

Sample type	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
Daily	0.28±0.23	588.38±64.45	$3.49*10^5 \pm 7.80*10^4$
Weekly	0.44±0.096	542.68±41.30	$2.96*10^5 \pm 4.35*10^4$
Monthly	0.60±0.071	441.58±15.01	$1.95*10^5 \pm 1.28*10^4$

Figure 4.8 visualises the predictions of the best performing LR from Phase 1.3. The SVR from Figure 4.6a was trained on the same random state and performed better overal. The LR is able to predict the higher SR values closer to the IL while the SVR does better on the lower values.

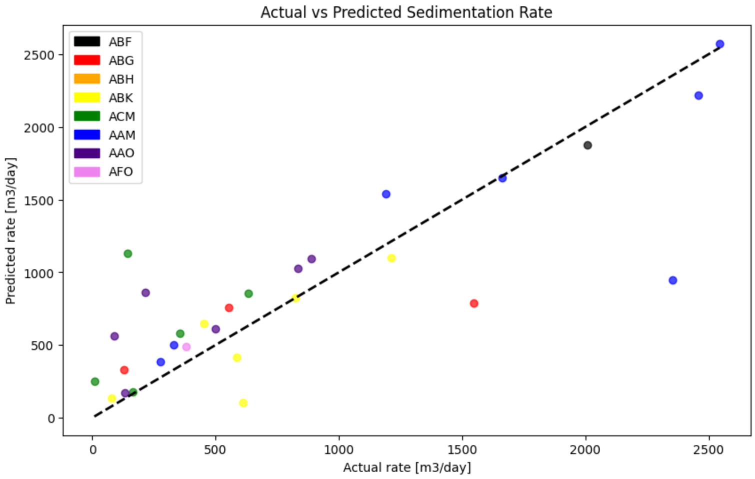


Figure 4.8: Scatter plot of the predictions of the best performing LR model ($R^2 = 0.69$) from Phase 1.3, plotted against the actual SR in m^3/day from the test sets. Better model performance means that the points are clustered near the IL.

4.4.4. Predictive accuracy per dredging area

The prediction versus actual SR plots show a colour code for each specific area but are challenging to analyse due to the many points. That is why the mean performance per area is calculated to indicate which areas are predicted the best. The runs were performed with SVR, monthly samples, and the Phase 1.3 configurations, as this combination has resulted in the best performance so far. Table 5.1 in the next chapter shows the results.

These scores do not agree with the mean score of the monthly runs from Table 4.17. There are two reasons for this:

- Sklearn can only calculate the R^2 for more than two values; otherwise, it returns a Nan-value. Some random states only added one or two samples of a particular area to the test set, resulting in the loss of test samples.
- Some areas can have significant negative R^2 scores similar to the LR in Phase 1.1. All negative R^2 scores were set to zero to prevent these areas from skewing the mean.

For these reasons, the scores from Table 5.1 provide an indication but not a definitive answer to the best predicted areas.

4.5. Feature importance scores

The mean feature importance scores resulting from all three phases are plotted in Figure 4.9. Each bar represents the mean of the nine runs from its respective phase except for the Phase 1.3 bars, as Phase 1.3 only trained RFR on weekly and monthly samples. An essential note is that Sklearn indicates that the predictive performance of the RFR should be high enough to get reliable feature importance scores (Sklearn, 2024a). Phase 1.3 produced the most accurate models and is, therefore, leading.Phase 1.3 produced the most accurate models. Therefore, the hydro-meteo variables are ranked from high to low based on the Phase 1.3 scores. In this study, the score is an aggregated score, meaning that for the daily and weekly runs, the scores of all lagged features of a specific hydro-meteo variable are aggregated into one value. The scores per phase can be found in Appendix C.3.1 and Figure 5.2.

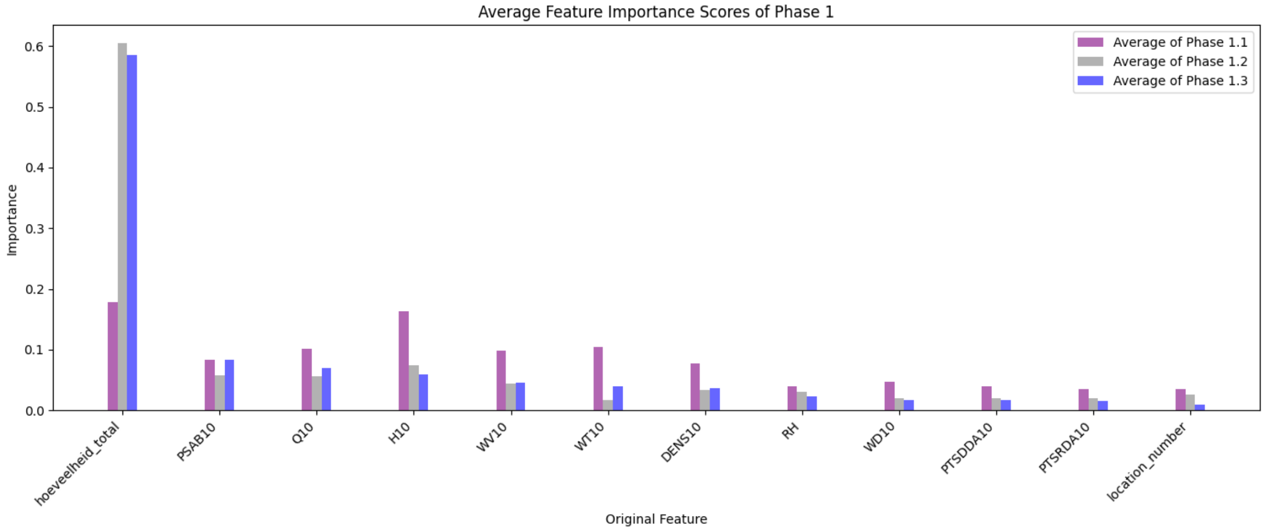


Figure 4.9: Mean feature importance scores of Phase 1. Each bar represent the mean of all 9 runs in a phase. 'Hoeveelheid_total' represents the total volume that is dredged in between two sequential surveys. The codes represent the hydro-meteo variables from 3.2.

The analysis on the feature importance scores acquired from the RFR models will be done in Section 5.1.4 in the next chapter.

4.5.1. Reduced feature set runs

The effect of removing hydro-meteo variables will be covered by Section 5.3. Table 4.22 shows the result of the first iteration where only PSAB10, H10, Q10, WV10, hoeveelheid_total, and location_number were used as features. The runs were performed on monthly samples as these consistently produce the highest performance. The tables summarises the individual runs from Appendix C.3.2.

Table 4.22: Summary of the performance of runs with less features and monthly samples

Algorithm	$\overline{R^2}$	$\overline{RMSE}$	$\overline{MSE}$
RFR	0.64±0.01	415.08±18.26	$1.73{*}10^5{\pm}1.66{*}10^4$
SVR	0.68±0.02	387.09±33.18	$1.51{*}10^5{\pm}2.52{*}10^4$
LR	0.60±0.03	437.25±14.73	$1.91{*}10^5{\pm}1.38{*}10^4$

4.6. Conclusion results

The results demonstrated the performance of LR, SVR, and RFR for predicting SR in the Botlek while using hydro-meteo variables and dredging volumes as input. If properly trained, the models should be able to predict how much sediment accumulated after a month of conditions and dredging. The models were evaluated across several phases that had changing configurations. Table 3.6 summarises the configurations of these phases.

Phase 1.0 provided a benchmark performance by training the models on dredging volumes only but with the configurations of the phases that follow. The best performance was produced by LR and SVR models that were trained to predict SR in m^3/day and without the samples with a negative SR (erosion). Both models reached a mean R^2 score of 0.65 across three random states.

In Phase 1.1, the hydro-meteo variables were added as input. None of the algorithms reached the performance of Phase 1.0. The SR in $\overline{m^3}$ proved to be a challenge, especially for the LR. The SVR showed the best mean performance. However, the maximum $\overline{R^2}$ of 0.22 showed that the results are not usable. There was no clear trend in reducing the number of features as the daily, weekly, and monthly samples types all showed varying performances.

Normalising the SR to m^3/day significantly improved the performance in Phase 1.2. Again, the SVR was able to produce the best performance. The best R^2 score reached in this phase was 0.67 and produced by a monthly run. The reduction in features resulted in an evident performance improvements. The less complex setup of Phase 1.0 still outperformed Phase 1.2. The prediction scatter plots showed that the models started to generalise better with each iterative modelling step, except for the erosion samples in the test set. The best working models structurally predicted negative SR values as positive, thus showing that these samples do not contribute to a better performance.

All runs in Phase 1.3 were performed with SR in m^3/day as Phase 1.2 proved that this resulted in a higher accuracy. Removing the erosion samples improved the performance of all three algorithms compared to the other phases. The SVR models trained on monthly samples from Phase 1.3 are able to beat Phase 1.0 on mean performance while weekly comes close. Moreover, LR and RFR, similar to the preceding phases, do not reach the same performance level as SVR or Phase 1.0.

The feature importance scores and location specific performance scores provide a basis for feature elimination and further model tuning. The implications of these scores and feature removal are discussed in the next chapter.

Table 4.23 summarises the findings of this chapter and shows the best performing configurations so far.

Table 4.23: Summary of best performing configurations. The 'Sample Type' rows of the two runner ups contain a '-' because no hydro-meteo variables were included as features. Hoeveelheid_total and location_number are uniform accros all sample types.

	Algorithm	**Sample Type**	**Unit SR**	**Input**	$\overline{R^2}$	$\overline{RMSE}$
Best performing	SVR	monthly	m^3/day	All variables + 'hoeveelheid_total'	0.69±0.053	388.78±27.59
Runner up	SVR	-	m^3/day	'hoeveelheid_total' + 'location_number'	0.65±0.02	407.95±29.27
Runner up	LR	-	m^3/day	'hoeveelheid_total' + 'location_number'	0.65±0.06	405.76±42.27

5 Discussion

The results from Chapter 4 are contextualised into technical and practical insights. In Section 5.1, the performance accuracy, feature importance scores, reducing the feature space, and hyperparameter configurations are further analysed. Next, the limitations of the modelling approach are outlined. By linking the limitations and the overall performance, suggestions on what can improve the accuracy of the hydro-meteo prediction models can be formed in Section 5.3. This answers the fifth sub-question:

What factors could enhance the performance of ML methods in predicting SR?

The value of this study for the POR is that any developed models have a practical implementation. The findings of this research process can assist in improving the efficiency of the maintenance operations. Therefore, the final sub-question is answered in Section 5.4:

What practical insights and implementations can be gained from the model results regarding sediment behaviour and maintenance efficiency in an estuarine harbour?

5.1. Result analysis

The conclusion on the results from Section 4.6 is observational. The underlying meaning and practical implication of these results are analysed in this section.

5.1.1. Accuracy

The R^2 was the leading performance indicator until now. The $RMSE$ is more challenging to compare when the errors are calculated on different random states. Nonetheless, the $RMSE$ can provide a better context to the actual predictions than the R^2 score. Take the residuals in Figure 5.1 of the best-performing SVR, for example.

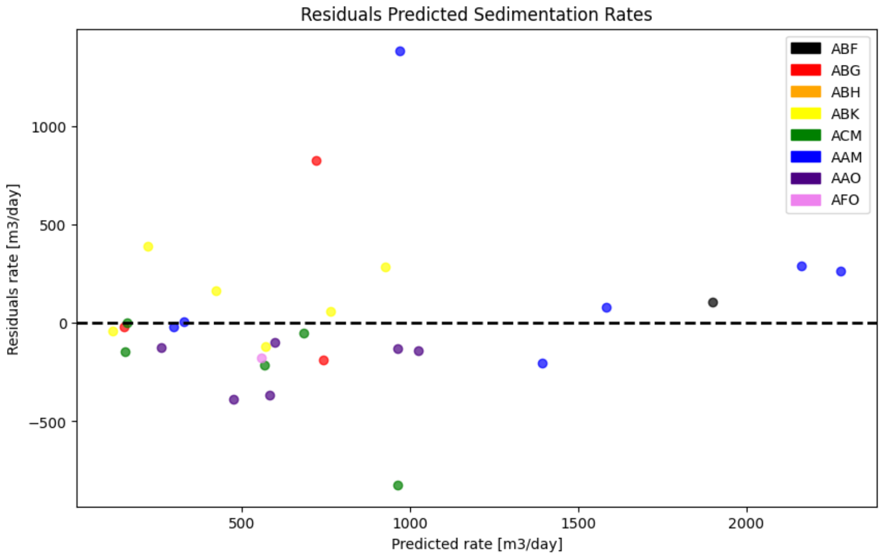


Figure 5.1: The residuals of the predictions of the best performing SVR model from Figure 4.6a. The dashed line represents an error of zero. The colour codes again refer to the dredging areas.

The R^2 score of this model is 0.74, which indicates a reasonably strong fit. This is evident from the many points scattered around the zero line. Most residuals do not exceed 300 to 350, except for some significant outliers. Although the R^2 suggests a reasonable fit on a daily scale, on a monthly scale, these errors can accumulate into 9,000 to 10,500 m^3. In context, the mean hoeveelheid_total per dredging area in the sample set ranges from 4,728 to 35,727 m^3.

An important takeaway is the generalisation of the model. Besides the extremes, the high SR values show the same error margin as the smaller values. The few outliers result in the $RMSE$ of 382.91 $m3/day$ while the MAE is 245.03. Percentage-wise, the relative monthly error will be lower for high SR. Despite this generalisation, the model consistently underestimates the high SR values because of the positive residuals. The majority of the lower SR have negative residuals, indicating overestimation. This phenomenon occurs in most prediction scatter plots in Chapter 4. When an SVR model is implemented, these findings can be integrated into the predictions by adjusting the low and high SR accordingly.

Accuracy per area

Section 4.4.4 mentioned that the mean accuracy per area was calculated with the best-performing SVR configurations and explained the reason why the $\overline{R^2}$ scores do not coincide with the ones from Table 4.23. Table 5.1 shows that the two best-predicted areas are ABH and ABK, both situated at the entrance of the Botlek. Overall, it seems that smaller dredging areas are predicted more accurately. ABG and ABF, both areas in the central channel, showed the worst overall performance by far.

Table 5.1: The mean performance per dredging area of 10 SVR runs with different random states. 'Dredging area' refers to the last three letters of the dredging codes, as seen in Figure B.20. '$\overline{SR}$' is the mean SR in m^3/day for samples corresponding to the dredging area and its standard deviation. 'Surface area dredging polygon' is the area in m^2 of the dredging area polygons from Figure B.20. The areas are ranked from best to worst.

Dredging area	$\overline{R^2}$	$\overline{RMSE}$	$\overline{SR}$	Surface area dredging polygon
ABH	0.73	337.57	957.37±757.46	85279
ABK	0.55	344.83	590.69±483.41	125157
AAO	0.51	240.28	519.2±459.52	87039
ACM	0.44	84.85	307.9±319.24	124499
AAM	0.41	240.69	1226.3±774.79	160439
AFO	0.40	135.87	461.49±381.02	217689
ABF	0.00	528.70	772.40±603.78	200583
ABG	0.00	943.40	673.61±776.02	239829

The first argument for the difference in performance is that the data quality varies over the areas. Figure B.22d visualises the poor match between the dredging volumes and the sedimentation for ABG. The dredging volumes are often zero, and the sedimentation dots show inconsistent behaviour. The reason for this is a practical oversight. The sediment trap in the central channel is dredged separately from the channel. This dredging area is called ABJ and is not shown on PortMaps which is why it was not noticed until the final phase of the research. The ABJ volumes are much higher than the ABG volumes and match the sedimentation trends in the ABG samples. On the contrary, Figure B.22e shows a clean data pattern for ABH.

Areas ABH to AFO all show a relatively reliable data pattern. Another reason for the performance differences could be the selected sample length of 31 days. ABH, ABK, and AAO are surveyed every 31 days, while the interval for ACM, AAM and AFO (Botlek Vak 3) is closer to 60 days. The samples fit the interval of the three best-performing areas, thereby better capturing the conditions leading up to the SR within the samples. Another consequence of the shorter interval is more surveys and, therefore, more samples representing these areas.

5.1.2. Complexity and performance

Phase 1.0 shows that simpler models do not necessarily perform significantly worse in predicting SR. The mean performances of the LR and SVR trained on dredging data were not that inferior to the SVR models from Phase 1.3. Simpler models are more suitable for direct implementation into the daily operations of port authorities from a practical point of view. For instance, processing the dredging data into input features instead of all hydro-meteo variables would be less time-consuming. Additionally, fewer features and simpler configurations require less computational resources. Section 3.6.3 already explained that taking the monthly means results in fewer features but a significant loss of information. This loss does not seem to impact the performance of the models. Aside from Phase 1.1, the monthly sample types consistently outperformed daily and weekly. However, considering the 'Large p, Small n' issue of Section 3.3.1, the daily and weekly SVR models from Phase 1.3 do not perform incredibly poorly.

With or without hydro-meteo variables

The performance similarities between the simple and complex models raise the question of whether including the hydro-meteo variables makes sense. It certainly does when considering the noticeable decrease in errors in Table 4.23. Here, the complex SVR model with all features has a lower $RMSE$ than the simple model. The previous section highlighted that minor improvements in errors still significantly impact the total predicted accumulation when scaled back to monthly rates. Another reason for including the hydro-meteo variables follows from this argument: only limited tuning has been done in this study due to time restrictions. The simple LR from Phase 1.0 are basically at maximum performance and already outperformed by the Phase 1.3 SVR models, while these could still be improved by feature selection and hyperparameter tuning. This process could make the difference in errors even more significant and the results more usable.

An additional argument is that there is room for improvement in the quality and quantity of the hydro-meteo data. This will be discussed in Sections 5.1.4 and 5.2. Lastly, not including the hydro-meteo variables removes a significant part of the practical implementations of the SR forecast model. The practical implementations are covered later in this chapter.

5.1.3. Optimal hyperparameters

The parameter grids used for SVR and RFR were constant throughout Phases 1.1 and 1.2. In Phase 1.3, additional parameters were added to the SVR grid. The selected hyperparameters can provide context to how the algorithms learned from the data. The varying effect across the random states suggested that the data splits impact the performance. Still, insights can be gathered by considering the general results from the RandomSearchCV for RFR and GridSearchCV for SVR in Appendix C.1.

RFR

The first two noticeable hyperparameters are max_depth and n_estimators. Usually, a larger number of trees reduces the variance as the prediction is averaged over more estimators (Hub, 2023). The results agree with this, as the better-performing models with a lower score variance generally had more trees, ranging from 300 to 500. The max_depth was also higher for these models, indicating deeper trees that can better capture complex relations. Deeper trees can lead to overfitting, which can be prevented by setting the min_samples_split parameter high enough. The better-performing models tend to have a value between 4 to 10. The min_samples_leaf parameter varies over all results. In Phase 1.3, where the RFR scored best, the hyperparameter varies from 1 to 6.

SVR

The regularisation parameter C for the best-performing models in Phase 1.0 is 500 to 1000. When the hydro-meteo variables are added, RandomSearchCV ends up with 10, 50 or 100. A higher C means outliers are more heavily penalised, which could lead to overfitting. However, gamma is consistently 0.001 or 0.01, making the SVR less flexible and preventing overfitting. The epsilon parameter is often smaller as well. This is visible in the scatter plot of Figure 4.6 where the SVR captured the overall trend to a certain extend. The most accurate models with the lowest $RMSE$ usually have 0.001 or 0.1 as epsilon. This means that even smaller errors are not ignored by the SVR models, which is beneficial for increasing the practical use of the model.

5.1.4. Feature importance score analysis

The results show that the models can partially learn relations and patterns when all hydro-meteo variables are included. The best-performing models were all trained on monthly sample types, indicating that fewer features resulted in better performance. The feature importance scores from the RFR runs in Phase 1 might offer support in decreasing the number of features even more to boost performance. The importance score indicates how much a feature decreases the impurity of a node in a decision tree. The scores from Phase 1.3 are shown in the figure below, as these scores are the most reliable. The four highest scores will be analysed based on the literature from Section 2.3.

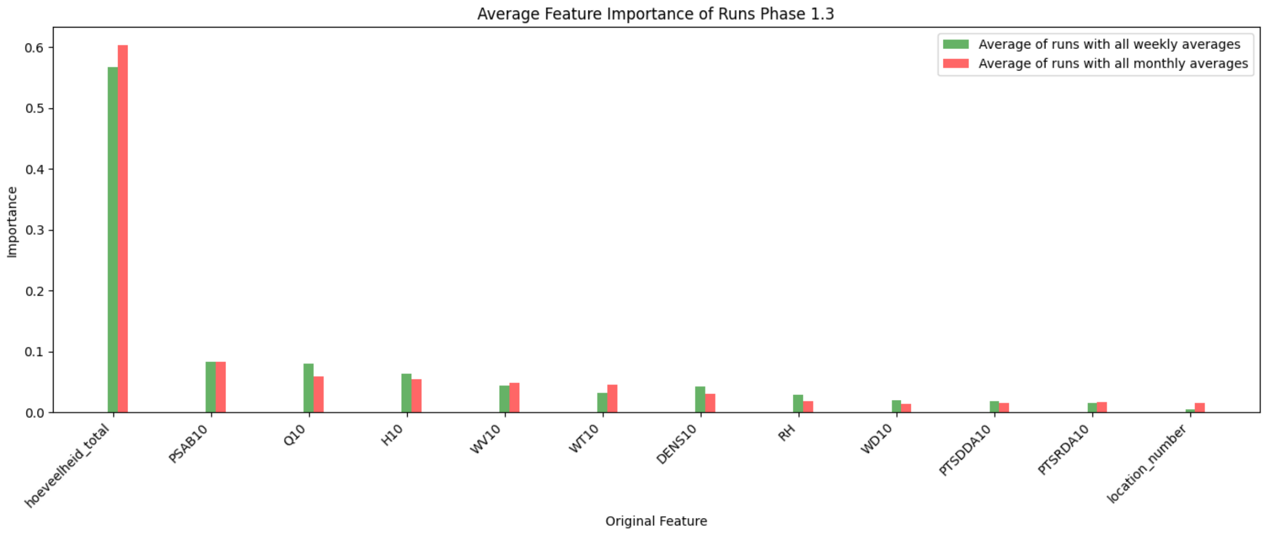


Figure 5.2: Mean feature importance scores of the RFR models from Phase 1.3. The hydro-meteo variables are ranked from highest to lowest score.

1. Hoeveelheid_total

The RFR models ascribe the highest importance scores to 'Hoeveelheid_total' in m^3. Section 4.1 already showed the importance of the dredging data as Phase 1.0 resulted in reasonably performing models that the Phase 1.3 SVR model outperformed. The high importance score was expected as dredging removes large amounts of sediment from the system, thus significantly influencing the bathymetry. In Figure 4.9, it is visible that Phase 1.1 did ascribe a much lower score compared to Phase 1.2 and 1.3. A probable reason for this is that Phase 1.1 had a very low accuracy compared to the other phases, thereby misinterpreting its relevance. Another explanation is that SR is normalised in Phases 1.2 and 1.3, resulting in a large difference between SR and hoeveelheid_total. The RFR models could ascribe a higher score due to this difference. Separate runs where hoeveelheid_total was normalised into a daily dredging rate (m^3/day) show that this is not the case. These runs resulted in approximately identical feature importance scores. Figure D.2 shows two of these runs.

2. Predicted Salinity at Bottom (PSAB10)

Section 2.3.1 stated that, according to the dissertation of De Nijs (de Nijs, 2012), the near-bed density currents are the dominant factor in the transport of SPM into the Botlek and that the stratified water column allows the SPM to settle. The position of the salt wedge is, therefore, an essential factor for sedimentation. The RFR models seem to be able to find this relation within the data and ascribed the second highest importance score to PSAB10. Unfortunately, there are only two locations where PSAB10 is predicted inside the Botlek. Table B.2 shows which stations were ascribed to which dredging area. There is a noticeable difference between the two locations as the mean PSAB10 at 2WERKH is 10.29 g/kg and 9.66 g/kg at BOTCGW. These salinity levels had to be generalised over a larger area, meaning that specific information on the location of the salt wedge inside the Botlek was not present. While the RFR still acknowledges the importance of this variable, dredging area-specific PSAB10 could show the variation of salinity across the Botlek and might result in a better performance per area.

3. Discharge (Q10)

Q10 has the third highest importance score. Q10 is not measured in or near the Botlek. Instead, Q10 is the discharge at Lobith delayed by four days (see section 3.3.1). Q10 is different from PSAB10 and other variables because Q10 refers to the discharge in the Nieuwe Maas and is not a variable inside the Botlek. Section 2.3.2 states that the accumulated sediment in the Botlek is primarily of fluvial origin. This does not necessarily imply that a high discharge results in more SPM and, therefore, more sedimentation in the Botlek. De Nijs (de Nijs, 2012) highlights the importance of the salt wedge location and the ETM on SR. A high river discharge mitigates the salt intrusion process (Huismans et al., 2024), thus changing the position of the salt wedge. The RFR recognises the interplay between these factors through the high importance scores for Q10 and PSAB10.

4. Tidal variation (H10)

H10 is given the fourth highest importance score. The coupling of PSAB10, Q10, and H10 by the RFR models is logical, as the mentioned interplay of the salt wedge and the river results from the tidal and riverine forcing. Section 2.3.1 stated that tidal advection controls the phase at which turbid and saline water reach the Botlek. However, the monthly mean of H10 represents the mean water level. With the averaging, the short-term effects of the tidal excursions or wind setup are smoothed out. RFR models notice that the water level variation affects SR as the importance scores for the weekly runs are higher for both H10 and Q10. The weekly mean shows the variation better than the monthly.

H10 also has the same shortcoming as PSAB10 because H10 is measured only at RP10, a radar post near the entrance of the Botlek. This means that variation across the Botlek is not considered with this hydro-meteo variable. Nonetheless, the minor fluctuations in the monthly means are enough to learn the effect that H10 has on SR partially.

5.1.5. Impact of reduced feature set

The general conclusion of the results is that reducing the number of features by taking monthly means improves the performance. This implies that the reduction in complexity outweighs the loss of information on hydro-meteo fluctuations. Following this assumption, further reducing the number of features by removing insignificant hydro-meteo variables could boost performance even more. Section 5.1.5 displayed the results that test this hypothesis. Table 4.22 from that section shows the mean performances of all algorithms on monthly sample sets that only had PSAB10, H10, Q10, WV10, hoeveelheid_total, and 'location_number' as input instead of all variables. These variables were selected because they were ranked highest on their feature importance score.

Starting with the significant improvement of the RFR models, the results indicate that removing variables can boost performance. The $\overline{R^2}$ goes up almost 0.1 and the errors decrease compared to the Phase 1.3 RFR. There is almost no notable difference in the absolute performance scores compared to Phase 1.3 for the LR and SVR. The effect on the RFR could be larger because RFR is an ensemble technique. The prediction is the average of many decision trees. This is elaborated in Section 2.1. Removing the chance that irrelevant features are selected for leaf nodes makes each tree more reliable before the average prediction is constructed. This not only decreases variance in performance but also boosts accuracy. The performance variance is the second noticeable aspect of Table 4.22. The standard deviation of the R^2 across all algorithms is lower, meaning that the different random states produce a more consistent accuracy. One explanation could be that removing the variables with a lower importance score reduced the noise more than it removed information.

The change to the original feature set is only the first iterative modelling step inspired by the feature importance scores. The many other configurations of including or excluding certain features could lead to an even more consistent and accurate performance of one or all algorithms.

5.2. Model limitations

The model development methodology is based on a few assumptions to overcome practical or theoretical limitations. These limitations stem from data availability, model structure, and practical challenges caused by the complexity of maintaining the infrastructure of a large port.

5.2.1. Data availability and quality

The available data enabled the development of a model that can be applied to the majority of the Botlek. Section 5.4 will contextualise that the accuracy is within an acceptable range, but the better the performance, the higher the practical use. The best-performing sample set had a total of 142 samples. A suggestion to boost the performance would be to add more data to the training set, as this might help to prevent overfitting. The reason for this argument is that small training sets are one of the main reasons for overfitting (Ying, 2020). However, Figure 5.3 indicates that adding more data might not be the solution for the Botlek prediction model.

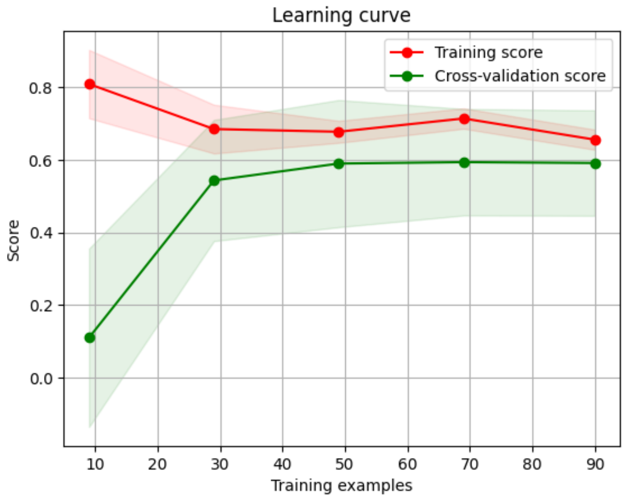


Figure 5.3: Learning curve of the best performing SVR (R^2=0.73). The curve is plotted using the standard learning_curve() function of Sklearn. The training score indicates how well the model fits the training data. The CV-score shows if the model can generalise on unseen validation sets within the training data.

The CV-score stabilises at 0.6 when half of the training examples are used, indicating that adding more training samples does not necessarily improve the accuracy (Sklearn, 2024b). The learning curve would be an argument to shift the focus to other model improvement factors like feature engineering. Despite this conclusion, Table 5.1 showed that this accuracy varies significantly over the dredging areas. In this area, the model can be refined by adding more data. The practical use of the model would be much higher if the model were able to predict SR with the same confidence interval for every area. Improvement of area-specific accuracy is partially reliant on the factors below.

Surveys

The level to which individual areas can be considered depends on how many samples can be labelled. The surveys are the only data source that provide the actual SR. This limits the possibility of training the ML models for individual areas. If it were not for the deepening in 2018 and 2019, the number of surveys that represent the bathymetry in the present conditions would be large enough to train the models on seperate or similar areas.

Hydro-meteo stations

The number of hydro-meteo stations across the Botlek is limited. This influences the quality of hydro-meteo data specific to a dredging area. Section 5.1.4 already mentioned that the data loses its temporal nature when averaged over more extended periods. Still, the models can learn from these weekly and monthly values. The variation is even further reduced by generalising the limited stations over the entire Botlek, precisely when minor differences between areas could improve the ability to learn site-specific relations.

Missing variables

Another limitation within the data influences the overall performance: the absence of turbidity data. De Nijs (de Nijs, 2012) stated the significant role of the ETM in transporting sediment into the Botlek. Turbidity could provide an essential context of the availability of SPM in the time frame leading up to a particular SR.

A potentially important variable that is available but not used is vessel traffic. Frequent traffic could stir up sediment, which can then be transported elsewhere. The vessel traffic could be incorporated by including AIS data or operation logs of the terminals in the Botlek.

5.2.2. Dredging area grouping

The decision to train the models on multiple areas was imposed by the need for more available survey data. It resulted in the performance differences in Table 5.1. Aside from area-specific performance, it also affects general performance. Even though the locations were distinguished by their respective numbers, the ability to learn patterns specific to individual areas was likely reduced. There could be a few explanations for this.

Sample interval

A uniform surveying interval of 31 days was assumed. This coincides with the survey interval of the areas expected to have a higher SR (see Table 3.1). The interval is essential in the sample creation as it determines the time frame of conditions that are ascribed to a certain SR. The dredging areas in Botlek Vak 3 (60 day interval) were added to the sample set under the assumption that there would otherwise not be enough data to learn from. Of the 142 samples in the most reliable set, 27 have an actual surveying interval of 60 days. The SR values of these samples are not adjusted by dividing them into two months, and the dredging data is also collected over two months. Only the hydro-meteo conditions span the chosen 30 days. Dividing the SR by two to scale the sedimentation could have provided a more realistic value for the models to train on.

Removing Vak 3 from the current training set could have been another development step. However, with the already low quantity of samples, this is not efficient. Instead, the choice of excluding the remaining dredging areas with a 60-day interval could have been changed by including them and then splitting all areas into two groups based on the intervals. The expectation was that the two sets would be too small and need more training data, which is why this approach was not selected in the first place. Figure I am running a few minutes late; my previous meeting is running over.5.3 shows that this assumption was unjust, as the performance of the SVR stagnated before all training samples were used.

Location within the Botlek

The positioning of each dredging area within the Botlek significantly influences sedimentation dynamics. Table 3.5 in Section 3.5.4 ranked the dredging areas with the highest $\overline{SR}$. AAM and ABH are both close to the entrance and experience the highest $\overline{SR}$ while AFO and ACM have the lowest $\overline{SR}$ and are positioned furthest away. Note that the $\overline{SR}$ of ACM, ABF, and AFO should be lower due to the interval inconsistency mentioned in the previous paragraph.

The table is meant to illustrate the impact of the positioning of the dredging areas on SR. In the current setup, merging these areas does not benefit the performance. The lack of area-specific hydro-meteo data and the missing dredging volumes of some areas might worsen the impact of merging even more. If all areas are represented evenly and in larger quantities in the training set, the ML models could perhaps capture separate patterns. Otherwise, grouping dredging areas based on their respective locations and characteristics could help the models to generalise the patterns within the data.

5.3. Conclusion model improvement factors

Sections 5.1 and 5.2 offer a great insight into why the models perform as they do. It is expected that the accuracy and consistency of the models can be improved by adjusting the modelling approach and through further tuning. The time limit of this study restricts further model development for now. The conclusion to the fifth sub-question can provide the guidelines for future development efforts. The question that is answered is:

What factors could enhance the performance of ML models in predicting SR?

Feature engineering and data quality

The initial model runs showed that incorporating the complete set of hydro-meteo variables as input improved performance compared to only training on dredging data. This aligns with the expectation that the ML models can recognise the impact of the variables on SR. Next, Section 5.1.4 proved that reducing the feature space by selecting the most essential variables contributes to a more consistent performance across the dataset. While this is promising, many combinations of variables, including those not currently available, like turbidity, can still be tested. Of all three algorithms, RFR benefited most from this feature space engineering.

Increasing the coverage of the hydro-meteo stations builds upon adding potentially meaningful variables. The current setup means that some dredging areas are not represented in the OSR or in measurements, which forces a generalisation of the data across the Botlek. This diminishes the variation between areas, preventing the models from capturing the subtle differences between the areas. This variation can be introduced by increasing the number of measurement stations.

The oversight of dredging area ABJ (see section 5.1.1) resulted in unreliable samples for ABG, a critical area in the Botlek. Adding the dredging volumes to better match the SR in the samples makes the dataset more reliable and increase the area specific accuracy.

Model adjustments

The current models are based on a sample set that includes dredging areas with varying surveying intervals (30 or 60 days), while the chosen time frame for the samples is set to 30 days. This introduces inconsistencies as the SR of more extended periods are treated the same as the 30-day SR values. Moreover, the areas are not filtered based on their position within the Botlek. It was shown that this results in a significant difference in accuracy per dredging area.

In addition to the area-specific hydro-meteo data, more uniform sample sets can be constructed if a clear division between areas with the same characteristics can be made. This would allow training separate models on these sub-groups to capture the location-coupled patterns better.

Hyperparameter tuning and ML selection

The hyperparameter grids used during training were coarse and not specified on commonly reoccurring values. The RFR grid only allowed the Default max_features and the kernel function of the SVR was consistently set as a radial basis function. The computational time of the SVR was only a few seconds. The RFR has shown to less flexible in increasing the grid size. The monthly runs did significantly reduce the runtime from almost 40 minutes to a few. These run times allow for a more detailed and focused GridSearchCV operation, especially with the Databricks resources of the POR.

LR, RFR, and SVR are the only algorithms applied in this study. In Section 3.4, XGB was discarded as the tuning process is regarded more complex. This section also highlighted the high accuracy provided by XGB. Adding this algorithm to the development process might result in a model that can beat the best performing SVR.

5.4. Practical implementation

The results have been shared with the POR AM department and the two thesis supervisors. A. van Hassent spoke on behalf of the dredging desk at AM, while E.B.J. Hupkes and J.E. Vettorato represent PRISMA and the POR Data Science department, respectively. By discussing the method behind the model development and contextualising the results, conclusions can be made regarding the practical use of an SR prediction model. This entire section serves as the conclusion to the last sub-question:

What practical insights and implementations can be gained from the model results regarding sediment behaviour and maintenance efficiency in an estuarine harbour?

The introduction explained that the port authority does not have real-time information on bathymetry or expected high SR. The bathymetry surveys are the only indication they have of the state of the waterbed, and the frequency of these surveys is monthly or bimonthly. Nonetheless, AM states that surveys are indispensable as they provide the most detailed image for the PortMaps used by the customers, thereby improving navigational safety. Moreover, the dredging vessels need the surveys to remove sediment efficiently. While a model will not replace dredging, it can still significantly contribute to proactive dredging, an aspect of maintenance that AM labels essential.

5.4.1. Proactive dredging

Proactive dredging means the authorities can anticipate where and when high SR levels are likely to occur. This is something that a predictive model could enable. The OSR already forecasts many hydro-meteo variables, including the critical PSAB10. The tidal variation H10 is periodical, so the tidal cycle can be predicted to a certain extent. Q10 is another essential variable that can be anticipated at least four days ahead (see Section 3.3.1) or even further based on seasonality.

A trend-based SR forecast system can be developed by integrating the predictive models with the information from the OSR. Rather than focusing on the absolute sediment volumes, the goal would be to identify trends in accumulation in the near future. To contextualise, AM considers a deviation of ±30% in predicted volume still operationally useful. This is enough to identify potential navigational dangers or obstructions of operations. Figure 5.4 shows that a significant share of the predictions from the best performing SVR fall well within this range.

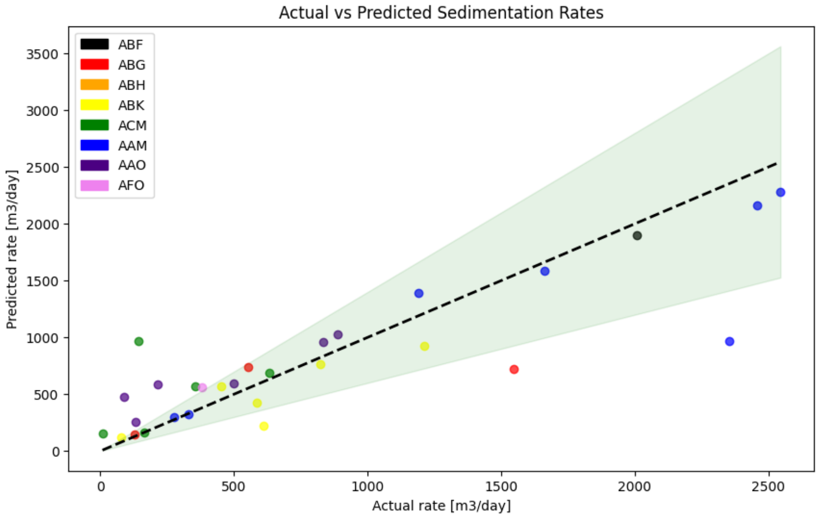


Figure 5.4: The scatter plot of the best performing SVR from Figure 4.6a. The green envelop is added to show the desired margin of error of ±30%.

From all predicted values, only six points can be considered to be far out of the range. This number could be reduced further through model tuning and when the consistent underestimation of high SR and overestimation of low SR (Section 5.1.1) are adjusted before the results are implemented. When the outliers are smoothed out, the developed models could already be integrated into the operational day-to-day.

An early-warning system in the form of the SR prediction model increases the ability to be proactive in channel maintenance. The dredging activities can be anticipated before the images of the surveys come through, making the operations more efficient and preventing the workload from piling up. Improving the operational efficiency by using trend-based predictions is practical in at least two other areas: navigational safety and finances.

Navigational safety

The navigational safety in any port depends, among other things, on a reliable NGD. Vessels should be able to enter the channels without worrying about grounding. Currently, the images provided by the surveys visualise the bathymetry on a (bi)monthly interval. This gives a periodic but not real-time understanding of the sediment accumulation. This limitation can create challenges if an unexpected period of extreme SR occurs. The predictive model, especially when integrated with the hydro-meteo forecasts, would allow the POR to anticipate these events.

With real-time information, the port authorities can mitigate the risk of encountering unknown shoals or shallow areas. AM indicated that the surveys are practical for the dredging vessels but not necessary. If the model can provide a sediment accumulation within the acceptable margin of error, the dredging vessels can remove the expected bumps without surveying first. This directly improves safety because the grounding risk has decreased. Moreover, the operational flow of the POR remains smooth because the risk was eliminated before the surveys were conducted. A smooth operational flow without bottlenecks in the channels also contributes to safety.

Financial benefits

Currently, the dredging budget is evenly divided over the year, with some flexibility for the months where more dredging activities are expected. The actual costs are adjusted monthly to anticipate any budgetary shortcomings. By integrating a trend-based prediction model, the POR can anticipate peak dredging periods more accurately. This would allow AM to distribute its resources efficiently. Budget adjustments can be made when dredging is needed during unexpected critical periods. Moreover, the ability to predict SR can also improve the dredging efficiency in the shorter term. AM can plan the dredging activities more effectively when high SR is anticipated. This could prevent redundant trips that cost resources and obstruct the operation flow.

Another benefit highlighted by the POR is the potential reduction of insurance premiums. A properly working SR prediction model increases navigational safety by reducing the risk of vessel grounding or vessel delays. This can help negotiate lower rates by demonstrating improved risk management. Aside from insurance, a reliable port could attract traffic and more clients to invest in long-term contracts.

5.4.2. Maintenance recommendations

This research resulted in the predictive models. The process behind the development provided some valuable insights regarding the maintenance operations and the possible integration of the predictive models. The integration recommendations do not cover the steps to improve the accuracy and consistency of the models as these are mentioned in Section 5.3 and Chapter 7.

Monitoring improvements

The measurement stations in the POR currently cover the general area, but the station density could be increased. If an SR prediction model were to be integrated, its performance would benefit from dredging-area-specific hydro-meteo conditions. The differences between the bordering dredging areas could be insignificant, but they show the variation across a basin. Also, the visiting vessels navigate the port with the assistance of the OSR forecasts. Increasing the level of detail in the OSR would contribute to the already mentioned navigational safety.

The data analysis showed that dredging data, particularly around the deepening operations of the Botlek, can be inconsistent or incomplete. The figures in Section B.3 show that the sedimentation acquired from the surveys does not always match the pattern of the dredging volumes. Of course, the sedimentation volumes are calculated based on the polygon areas and not over the entire dredging areas, so a margin of error cannot be prevented. Despite this error margin, the difference between the dredged volumes and the sedimentation was, in many cases, still significant and, therefore, unusable for the SR prediction model.

Model integration

This study focused on developing a model for the dredging areas in the Botlek. Currently, the code that trained the LR, RFR, and SVR algorithms does not allow a direct application of the best-performing model. The main reason is that no data pipeline feeds up-to-date information to predict SR. For this to happen, the following components need to be developed:

- **Continuous OSR information:** The data had to be manually exported from the Hystorical Data Store HydroMeteo, the data lake, and the dredging database to create the training samples. An automated data pipeline should be developed to ensure that the SR model can make real-time or frequent predictions.

- **Automated data processing:** A system must be able to preprocess the data from the automated pipeline automatically. This includes normalising the input and cleaning missing values. Moreover, hydro-meteo variables must be processed into daily, weekly, or monthly means, depending on which performs best after further model development.
- **Forecast interface:** A user-friendly interface must be developed to make the models accessible for AM. The interface could display the SR trends along with the confidence interval. The results can assist them in proactive maintenance.
- **Historical data integration:** The SR prediction model can improve if retrained with updated historical data. The model performance can be refined by integrating newly gathered hydro-meteo, dredging, and sedimentation data.

By implementing these components, the predictive models could operate in real time. The models can then be used to support decision-making around dredging operations and navigational safety and improve the overall maintenance strategy of the POR.

6 Conclusion

This research aims to enhance the efficiency of the POR maintenance operations by developing ML models capable of predicting SR by learning patterns in hydro-meteo conditions and dredging data. Therefore, the main research question of this thesis was:

To what extent can ML methods be utilised in predicting SR in an estuarine harbour, considering the dynamic interplay of marine and riverine influences?

The estuarine harbour of interest is the Botlek because this harbour experiences high SR and is situated at the transition from fresh to saline water. The method and findings behind developing SR prediction models were captured into six sub-questions, starting with:

What data is available and relevant for predicting SR using ML in estuarine harbours?

The research identified three main data types that provide critical insights into sediment accumulation and the environmental conditions: Multibeam bathymetry surveys, hydro-meteo variables, and dredging data. The surveys are essential in determining the sediment build-up over time and are dredging area-specific. The areas in the Botlek are surveyed every 30 or 60 days, depending on if an area is considered high priority due to high SR. Hydro-meteo variables such as salinity, river discharge, and tidal variation partially explain the conditions contributing to sedimentation. The variables are measured or predicted at multiple locations throughout the Botlek. Lastly, the dredging data provides historical records of dredging operations between surveys. The dredging volumes can, therefore, explain inconsistent sedimentation patterns in the surveys. Integrating these data types is the key to developing ML models that can accurately predict SR.

Which ML algorithms are most suitable for predicting SR?

The limited data availability and the goal of developing a practical SR prediction model make it that the ML algorithms must be able to handle small datasets, have an interpretable process and result, and have shown high predictive accuracy in similar applications. The algorithms that are viable candidates based on the literature in Section 3.4 are ANN, RFR, SVR, and XGB. By reviewing the criteria, RFR and SVR are considered the most suitable algorithms for predicting SR within the scope of this thesis. Additionally, LR can provide a baseline performance.

The choice for RFR was based on its ability to produce feature importance scores. The scores can contribute to interpretability and assist in determining which hydro-meteo variables have an essential influence on SR. Moreover, the ensemble nature of RFR reduces the risk of overfitting, a common challenge when working with small datasets. SVR is a more complex algorithm, but Section 3.4 highlights its effectiveness on small datasets and non-linear relations. Combining RFR and SVR ensures interpretability and a robust approach to predicting SR in the Botlek.

How can the selected ML algorithms and features be configured to predict SR?

Configuring the ML algorithms and data to predict SR starts with constructing training samples that align the survey data, hydro-meteo variables, and the dredging logs. Two sequential surveys at the same dredging area can provide the net bed level change, which can be converted into an SR value. This SR is the label of a sample.

The interval between the sequential surveys provides the time frame over which the hydro-meteo time series can be extracted. This is based on the assumption that the conditions in this time frame correlate with the observed bed level change. RFR and SVR require the training data to be shaped as a one-dimensional array. Therefore, the time series were transposed into lagged features to maintain the temporal dependencies of the data. Finally, the total volume dredged between the surveys was added to the sample row. The hydro-meteo variables were aggregated into daily, weekly and monthly mean values to create three sample types and reduce the sample length. This is because decreasing the number of features reduces model complexity and might boost the predictive accuracy. The trade-off is that averaging results in a loss of periodic variation in the hydro-meteo variables.

The varying survey interval limits the proposed sample creation method. Since RFR and SVR require uniform sample lengths, a constant survey interval must be selected. The downside to a constant value is that including too many areas with varying intervals would introduce inconsistencies. This study set the interval to 30 days to comply with the high-priority areas. The 60-day areas, except for three areas in Botlek Vak 3, were excluded from the sample set to prevent skewing the model. The three areas were added to compensate for the loss in data quantity.

The ML model development was split into phases to investigate the impact of different data configurations. In every phase, all three sample types were tested:

- **Phase 1.0:** designed to establish a baseline by training models on dredging data without hydro-meteo variables. This approach assesses how much of SR variability can be explained by dredging alone.
- **Phase 1.1:** introduced hydro-meteo variables to test if this would enhance the predictive accuracy. The dependent variable in this phase was SR in m^3 as this is the most practical unit for the port authorities.
- **Phase 1.2:** used the same input but normalised the dependent variable into SR in m^3/day to explore whether normalisation would improve the performance.
- **Phase 1.3:** builds upon the preceding phases by selecting the best-performing configurations and applying them to the most reliable sample set. In Phases 1.1 and 1.2, the dataset contained samples with a negative SR (erosion).

The systematic approach ensured that each aspect of the data and model configurations can be tested. This increases the understanding of the factors influencing model performance and the sedimentation process in the Botlek.

How do the selected ML algorithms perform across different configurations?

The performance of the ML algorithms varied significantly across the different phases. Overall, the best performance was achieved in Phase 1.3. The combination of SVR with SR in m^3/day, the monthly sample types, hydro-meteo variables, the dredging data, and excluding negative SR values results in the best performance with an $\overline{R^2}$ of 0.69 and a $\overline{RMSE}$ of 388.78 over four different dataset splits. SVR outperforms LR and RFR not only in Phase 1.3 but in every phase.

Removing the erosion samples and normalising SR had a crucial role in improving the accuracy and consistency of the models. Phase 1.0 already provided a performance close to the one achieved in Phase 1.3. Adding the hydro-meteo variables did not improve the performance as much as the normalisation. However, it showed that it contributes to higher predictive accuracy and consistency, especially considering that the models from Phase 1.3 can still be tuned.

The RFR provided feature importance scores that form a basis for further feature reduction. The Phase 1.3 RFR models have the highest performance, which is why these scores are the most reliable for eliminating redundant hydro-meteo variables. The input variables with the highest importance scores are dredging volumes, salinity, discharge in the Nieuwe Maas, and tidal variation. This ranking is in agreement with the findings from the literature in Section 2.3. A first iterative model run was performed with only the highest-ranked variables. The RFR significantly benefits from this initial step, while the absolute mean performance scores of LR and SVR remain constant. However, it can be concluded that the continued decrease in features improves the consistency of all algorithms across the different dataset splits.

What factors could enhance the performance of ML models in predicting SR?

The results show a significant variance in performance between individual dredging areas. By analysing the area-specific accuracy, it is evident that this relies on the selected interval length and the position of the dredging area in the Botlek.

Dredging areas with a 30-day survey interval performed significantly better, except for the areas in the central channel. This underperformance is likely caused by the poor dredging data quality for this region and perhaps the sediment trap in the channel. While the ML models can partially capture the patterns and correlations for the 60-day areas in Vak 3, their performance is not as accurate as for the 30-day areas.

A logical step forward would be to group the dredging areas with the same survey interval into separate training sets. This can prevent the models from unnecessarily generalising area-specific patterns over the Botlek. Furthermore, this improvement can be strengthened by providing hydro-meteo data specific to each dredging area. Currently, the limited number of measurement stations does not allow specific conditions. Instead, the information has to be generalised over the Botlek, which diminishes the local variations between dredging areas. Incorporating these suggestions might decrease the difference in performance between areas.

What practical insights and implementations can be gained from the model results regarding sediment behaviour and maintenance efficiency in an estuarine harbour?

The SR prediction models enable proactive dredging and improve navigational safety. The POR indicates that trend-based predictions with a confidence interval of ±30% are enough to be of practical use. Most predictions in this study, with the exception of outliers, fall within this range.

Currently, the dredging operations rely on the images provided by the (bi)monthly surveys. Forecasting SR will allow the authorities to anticipate when and where high SR is likely to occur in real time. This prevents unexpected peaks in workload and allows for more efficient operations. This proactive method assists in effectively distributing the resources, thereby avoiding operational shortcomings during seasons with high SR or at the end of the year. Moreover, with enhanced system knowledge and the ability to forecast SR, the risk of bottlenecks and vessel grounding is significantly reduced, ensuring smoother and safer port operations.

With all sub-questions covered, an answer to the main research question can be given:

To what extent can ML methods be utilised in predicting SR in an estuarine harbour, considering the dynamic interplay of marine and riverine influences?

The research demonstrates that the developed ML models can be utilised to predict SR in estuarine harbours to a certain extent. Among the tested algorithms, SVR emerged as the most reliable algorithm, consistently outperforming RFR and LR in almost every research phase. This success is likely due to its ability to handle small datasets and high-dimensional relations.

The most straightforward model configuration, which focused on dredging data only, already produced somewhat reasonable predictions. However, incorporating the estuary dynamics by including the hydro-meteo variables enhanced model performance. The larger feature space allowed the models to better capture the conditions influencing the sedimentation process. Feature importance scores provided by the RFR ranked salinity, discharge, and tidal variation as essential factors in the sedimentation process, aligning with the findings of preceding research on the dynamics in the Botlek. Furthermore, the models showed that using monthly means of the hydro-meteo conditions currently results in the most accurate predictions. Excluding non-essential variables resulted in less variance across the different dataset splits. This indicates that a reduced, more refined selection of hydro-meteo features benefits the performance. A variable that is currently unavailable but potentially influential in the feature set is turbidity.

The models achieved promising accuracy with a significant share of the predictions already within or near the ±30% error margin mentioned by the POR. However, additional work is required to integrate them into the maintenance operations. Challenges such as refining dredging area-specific models and expanding the coverage of the hydro-meteo stations need to be addressed. Moreover, developing a real-time data pipeline is crucial if the models are to be added to the decision-making process of the asset management department. The research shows that, with further development and research, ML can deliver high-potential results that can be used as a practical SR prediction tool for port maintenance.

7 Recommendations

This chapter outlines several recommendations based on the findings of this research. These recommendations aim to improve the predictive accuracy of SR prediction models, assist in integrating the models into maintenance operations, and expand the application to different regions within the POR. Extensive coverage of the practical recommendations and model limitations can be found in Sections 5.2 and 5.3.

Retraining on different sub groups

The research shows significant differences in model performance based on dredging areas, particularly between regions with 30-day and 60-day survey intervals. Splitting the training data based on these intervals is recommended. Grouping dredging areas with similar intervals could reduce the variability in performance and allow the models to better capture patterns specific to that group or area. Location-based grouping can also be effective, considering the effect the positioning within the Botlek has on sedimentation.

Additionally, the assumption that splitting data into smaller groups would result in training sets that were too small was not necessarily valid. The results showed that the accuracy becomes constant halfway through the training process of the best-performing models. The remaining training data was, therefore, redundant. The suggestion is to use all available data by including the 2e Werkhaven and 1e Werkhaven. These areas should not be incorporated into the existing sample set. Instead, add these to the 60-day interval sets and retrain the models on the two separate, more reliable sample sets to potentially achieve better model performance.

Model accuracy and additional algorithms

The current models were developed with limited hyperparameter tuning. The focus was primarily on a coarse grid search rather than a detailed exploration of the hyperparameter space. Given the available computing resources of the POR, more extensive and detailed hyperparameter tuning for RFR and SVR is recommended.

Moreover, feature engineering could still result in improvements. The feature importance scores from the RFR currently resulted in only one iterative run with a reduced feature set. Therefore, it is recommended to test the many available feature combinations. An additional engineering step could be to explore the possibility of creating a sample set using biweekly means. This would reduce the period over which the variables are averaged while still ensuring a minimum amount of features.

Finally, incorporating new algorithms like XGB should be considered. The literature review has shown that XGB has already been applied for sedimentation prediction. Moreover, it indicates its high accuracy. XGB may outperform SVR and RFR when applied to the context of this study.

Model integration

Integrating the developed models into the maintenance operations will require several development steps. First, establishing a real-time data pipeline is crucial before continuous monitoring to predict SR is possible. The pipeline should automate the collection of the bathymetry surveys, hydro-meteo data, and dredging logs. Additionally, automated data preprocessing tools need to be incorporated into the pipeline. These tools must handle data cleaning, normalisation (for the SVR), and the transformation of the hydro-meteo variables into the daily, weekly, or monthly averages.

To make the models accessible, a user-friendly interface must be created so that the SR predictions are readily available for AM. The interface could display the predicted SR trends, the confidence intervals of the predictions, and locations at risk of high accumulations.

Finally, it is recommended to implement continuous learning to ensure long-term and constantly improving accuracy of the models. As more dredging, hydro-meteo, and survey data become available, a feedback loop would allow the model to evolve and refine its predictions. Any significant changes in infrastructure or bathymetry might limit the continuous learning loop.

Include turbidity and time of year

Two features that could add essential information are turbidity and a number that refers to the time of year. Including a feature that represents the month in which the surveys have been conducted might allow the ML models to capture the seasonal variation of SR. The availability of SPM is crucial for sedimentation in the Botlek. The current hydro-meteo data does not include a factor that represents this. Measuring turbidity levels can provide an indirect measure of the SPM and add crucial information on SPM availability for the ML models.

Project predictions over bathymetry

The predicted SR values are specific to a dredging area but do not specify where the sedimentation will likely occur. Additionally, translating an SR value into actual bathymetry changes is challenging without a visual aid. It would be valuable to map the expected SR over the bathymetry while considering area-specific patterns to make the model output more insightful. Instead of projecting a uniform increase in bathymetry, the expected accumulation can be ascribed to areas prone to sedimentation. The development of this approach would take up a significant amount of time as the sedimentation patterns of every dredging area must be analysed and configured into maps with sedimentation probabilities.

Dredging data Botlek Centrale Geul

When discussing the results with AM, it was discovered that the ABJ dredging data for the central channel was overseen. Most sediment in the central channel is removed from the sediment trap called ABJ. However, the sediment trap is not visualised in PortMaps as a dredging area, so it was not noticed. The majority of the dredging operations are done in the trap and not in the ABG area. This resulted in a mismatch between the sediment accumulation required from the surveys and the dredging volumes. The dredging volumes were often zero.

The area-specific accuracy analysis identified ABG as underperforming. The oversight of ABJ is the reason for this. The ABJ data should be added to the samples to create a more accurate dataset. Incorporating the dredging volumes into the samples with the developed sample pipeline is straightforward. This is expected to boost the performance, especially when the revised ABG samples are included in a 30-day-only sample set.

References

Allen, G.P., J.C. Salomon, P. Bassoulet, Y. Du Penhoat, and C. de Grandpré (1980). “Effects of tides on mixing and suspended sediment transport in macrotidal estuaries”. In: *Sedimentary Geology* 26.1-3, pp. 69–90. DOI: 10.1016/0037-0738(80)90006-8.

Anisa, Yuan, Winda Erika, and Fadhillah Azmi (2024). “Enhancing Student Performance Prediction Using a Combined SVM-Radial Basis Function Approach”. In: *International Journal of Innovative Research in Computer Science and Technology (IJIRCST)* 12.3. Article ID IRP1448, pp. 1–5. ISSN: 2347-5552. DOI: 10.55524/ijircst.2024.12.3.1. URL: https://www.ijircst.org.

Awad, Mariette and Rahul Khanna (2015). “Support Vector Regression”. In: *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA: Apress, pp. 67–80. ISBN: 978-1-4302-5990-9. DOI: 10.1007/978-1-4302-5990-9_4. URL: https://doi.org/10.1007/978-1-4302-5990-9_4.

Basak, Debashis, Srimanta Pal, and Dipak Chandra Patranabis (2007). “Support Vector Regression”. In: *Neural Information Processing - Letters and Reviews* 11.10.

Belyadi, Hoss and Alireza Haghighat (2021). “Chapter 5 - Supervised learning”. In: *Machine Learning Guide for Oil and Gas Using Python*. Ed. by Hoss Belyadi and Alireza Haghighat. Gulf Professional Publishing, pp. 169–295. ISBN: 978-0-12-821929-4. DOI: https://doi.org/10.1016/B978-0-12-821929-4.00004-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128219294000044.

Bosboom, J. and M. J. F. Stive (2023). *Coastal Dynamics*. Delft University of Technology, Delft, The Netherlands.

Britannica, Encyclopædia (2024). *Pycnocline*. https://www.britannica.com/science/pycnocline.

Brownlee, J. (2020a). *Understand Your Problem and Get Better Results Using Exploratory Data Analysis*.

— (Aug. 2020b). *What is the Difference Between Test and Validation Datasets?* https://machinelearningmastery.com/difference-test-validation-datasets/.

— (Oct. 2021). *How to Choose an Optimization Algorithm*. https://machinelearningmastery.com/tour-of-optimization-algorithms/.

Bruijn, L.H. de (2018). “Maintenance dredging in the Port of Rotterdam”. MA thesis. Delft University of Technology. URL: https://repository.tudelft.nl/islandora/object/uuid%3Aef075688-d855-4f99-9de4-c562e4211335.

Chang, Audrey Chingzu and Jyh-Ping Hsu (2006). “A polynomial regression model for the response of various accelerating techniques on maize wine maturation”. In: *Food Chemistry* 94.4, pp. 603–607. ISSN: 0308-8146. DOI: https://doi.org/10.1016/j.foodchem.2004.11.048. URL: https://www.sciencedirect.com/science/article/pii/S030881460500018X.

de Nijs, MAJ (2012). “On sedimentation processes in a stratified estuarine system”. English. Dissertation (TU Delft). Delft University of Technology. ISBN: 978-94-6191-562-7.

Deltares (2024). *Sediment management for accessible ports and waterways*. Accessed: 2024-10-03. URL: https://www.deltares.nl/en/expertise/areas-of-expertise/future-proof-infrastructure/ports-and-waterways/sediment-management-for-accessible-ports-and-waterways.

Dhumne, S. (Apr. 2023). *Introduction to XGBoost Classifier*. https://medium.com/@shruti.dhumne/introduction-to-xgboost-classifier-ba9a2d6d5d5.

Dongare, A.D., R.R. Kharde, and Amit D. Kachare (2012). “Introduction to Artificial Neural Network”. In: *International Journal of Engineering and Innovative Technology (IJEIT)* 2.1. ISSN: 2277-3754.

Doroudi, S., A. Sharafati, and S. H. Mohajeri (2021). “Estimation of Daily Suspended Sediment Load Using a Novel Hybrid Support Vector Regression Model Incorporated with Observer-Teacher-Learner-Based Optimization Method”. In: *Complexity* 2021, pp. 1–13. DOI: 10.1155/2021/5540284.

Eisma, D., P. Bernard, G.C. Cadée, V. Ittekkot, J. Kalf, R. Laane, J.M. Martin, W.G. Mook, A. Van Put, and T. Schuhmacher (1991). “Suspended-matter particle size in some west-European estuaries; part I: Particle-size distribution”. In: *Netherlands Journal of Sea Research* 28.3, pp. 193–214.

El Hamdi, A. (2012). “Sedimentation in the Botlek Harbour - A research into driving water exchange mechanisms.” MA thesis. Delft University of Technology. URL: https://repository.tudelft.nl/islandora/object/uuid%3A55c2fb12-1d77-47af-ab04-5ffc3adb3abe.

Elnabwy, M. T., E. Elbeltagi, M. M. E. Banna, M. Y. Elsheikh, I. Maatouq, J. Hu, and M. R. Kaloop (2022). “Conceptual prediction of harbor sedimentation quantities using AI approaches to support integrated coastal structures management”. In: *Journal of Ocean Engineering And Science*. DOI: 10.1016/j.joes.2022.06.005.

EPSG (2024). https://epsg.io/transform#s$_srs = 4326t_srs = 28992x = NaNy = NaN$.

Feng, Xuezhi, Chaoqi Zhu, J. Paul Liu, and Yonggang Jia (2023). "Sediment Dynamics in Coastal and Marine Environments: Scientific Advances". In: *Water* 15.7. ISSN: 2073-4441. DOI: 10.3390/w15071404. URL: https://www.mdpi.com/2073-4441/15/7/1404.

França, Reinaldo Padilha, Ana Carolina Borges Monteiro, Rangel Arthur, and Yuzo Iano (2021). "Chapter 3 - An overview of deep learning in big data, image, and signal processing in the modern digital age". In: *Trends in Deep Learning Methodologies*. Ed. by Vincenzo Piuri, Sandeep Raj, Angelo Genovese, and Rajshree Srivastava. Hybrid Computational Intelligence for Pattern Analysis. Academic Press, pp. 63–87. ISBN: 978-0-12-822226-3. DOI: https://doi.org/10.1016/B978-0-12-822226-3.00003-9. URL: https://www.sciencedirect.com/science/article/pii/B9780128222263000039.

Fuladipanah, Mehdi, H. Md. Azamathulla, Ozgur Kisi, Mehdi Kouhdaragh, and Vishwandham Mandala (2024). "Quantitative forecasting of bed sediment load in river engineering: an investigation into machine learning methodologies for complex phenomena". In: *Water Supply* 24.2, pp. 585–600. DOI: 10.2166/ws.2024.017. URL: https://doi.org/10.2166/ws.2024.017.

Gaurav (Mar. 2022). *An Introduction to Gradient Boosting Decision Trees*. https://www.analyticsvidhya.com/blog/2022/03/an-introduction-to-gradient-boosting-decision-trees/.

GEOJSON (2024). https://geojson.io/#map=13.89/51.88645/4.28433.

Geraeds, M.E.G. (2020). "The hydrodynamics of an eco-innovative sediment reuse project in the Rotterdam Waterway: Gaining insight into the physics and the predictive capability of two operational hydrodynamic models". Master's thesis. MA thesis. Delft University of Technology. URL: http://resolver.tudelft.nl/uuid:687f4d25-2b04-4c26-b927-20af2d262ab0.

Geyer, W. Rockwell (1993). "The Importance of Suppression of Turbulence by Stratification on the Estuarine Turbidity Maximum". In: *Estuaries* 16.1, pp. 113–125.

Goldstein, Evan B., Giovanni Coco, and Nathaniel G. Plant (2019). "A review of machine learning applications to coastal sediment transport and morphodynamics". In: *Earth-Science Reviews* 194, pp. 97–108. ISSN: 0012-8252. DOI: https://doi.org/10.1016/j.earscirev.2019.04.022. URL: https://www.sciencedirect.com/science/article/pii/S001282521830391X.

Gonzalez Rodriguez, L., A. McCallum, D. Kent, et al. (2023). "A review of sedimentation rates in freshwater reservoirs: recent changes and causative factors". In: *Aquatic Sciences* 85, p. 60. DOI: 10.1007/s00027-023-00960-0. URL: https://doi.org/10.1007/s00027-023-00960-0.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press. Chap. 5, pp. 111–115.

Google Cloud (2024). *What is Supervised Learning?* https://cloud.google.com/learn/what-is-supervised-learning.

Guo, N., W. Gui, and W. et al. Chen (2020). "Using improved support vector regression to predict the transmitted energy consumption data by distributed wireless sensor network". In: *Journal of Wireless Communications and Networking* 2020, p. 120. DOI: 10.1186/s13638-020-01729-x.

Guo, Yakun (2022). "Hydrodynamics in Estuaries and Coast: Analysis and Modeling". In: *Water* 14.9. ISSN: 2073-4441. DOI: 10.3390/w14091478. URL: https://www.mdpi.com/2073-4441/14/9/1478.

H20 (2022). *Discharge Rhine approaches lowest ever recorded*. https://www.h2owaternetwerk.nl/h2o-actueel/waterafvoer-rijn-nadert-laagste-afvoer-ooit.

Hengl, Tomislav, Madlene Nussbaum, Marvin N. Wright, Gerard B.M. Heuvelink, and Benedikt Gräler (2018). "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables". In: *PeerJ* 6, e5518. DOI: 10.7717/peerj.5518. URL: https://peerj.com/articles/5518/.

Hinwood, J.B. and E.J. McLean (2018). "Tidal inlets and estuaries: Comparison of Bruun, Escoffier, O'Brien and attractors". In: *Coastal Engineering* 133, pp. 92–105. DOI: https://doi.org/10.1016/j.coastaleng.2017.12.008. URL: https://www.sciencedirect.com/science/article/pii/S0378383917301916.

Holihah, Hesti (2023). "Hyperparameter Tuning Showdown: Grid Search vs. Random Search - Which is the Ultimate Winner?" In: *Medium*. URL: https://medium.com/@hestisholihah01/hyperparameter-tuning-showdown-grid-search-vs-random-search-which-is-the-ultimate-winner-5927b322e54d.

Hub, Tilburg Science (2023). *Random Forest — Machine Learning: Supervised Learning*. URL: https://tilburgsciencehub.com/topics/analyze/machine-learning/supervised/random_forest/.

Huismans, Y., L. Leummens, S.M.T. Rodrigo, and S.C. Laan (2024). "The effectiveness of fresh-water pulses to mitigate salt intrusion into the Lek River". In: *Netherlands Centre for Coastal Research*. URL: https://www.nck-web.org/boa-2024/835-the-effectiveness-of-fresh-water-pulses-to-mitigate-salt-intrusion-into-the-lek-river.

Huynh, PH, VH Nguyen, and TN Do (2020). "Improvements in the Large p, Small n Classification Issue". In: *SN COMPUT. SCI.* 1, p. 207. DOI: 10.1007/s42979-020-00210-2.

IBM (2024a). *What is geospatial data?* https://www.ibm.com/topics/geospatial-data.

— (2024b). *What is machine learning?* https://www.ibm.com/topics/machine-learning.

Janssen, W. (2024). *Wind in The Netherlands*. https://www.weerplaza.nl/weerinhetnieuws/klimaat/wind-in-nederland/6820/.

Kadam, Vidya S., Shweta Kanhere, and Shrikant Mahindrakar (2020). “Regression Techniques in Machine Learning & Applications: A Review”. In: *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* 8.X, pp. 826–834. ISSN: 2321-9653. URL: `https://www.ijraset.com`.

Kim, N., D. H. Kim, and S. Park (2023). “Prediction of the Turbidity Distribution Characteristics in a Semi-Enclosed Estuary Based on the Machine Learning”. In: *Water* 16.1. DOI: `10.3390/w16010061`.

Kirichek, Alex, Ronald Rutgers, Marco Wensveen, and Andre van Hassent (2018). “Sediment Management in the Port of Rotterdam”. In: *Proceedings of the Department of Asset Management, Port of Rotterdam, the Netherlands*. Port of Rotterdam. Netherlands.

KNMI (2024). *Uurgegevens van het weer in Nederland*. https://www.knmi.nl/nederland-nu/klimatologie/uurgegevens.

Latif, S.D., K. L. Chong, A. N. Ahmed, Y. F. Huang, M. Sherif, and A. El-Shafie (2023). “Sediment load prediction in Johor river: deep learning versus machine learning models”. In: *Applied Water Science* 13.3. DOI: `10.1007/s13201-023-01874-w`.

Loyola-González, Octavio (2019). “Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View”. In: *IEEE Access* 7, pp. 154096–154113. DOI: `10.1109/ACCESS.2019.2949286`.

MathWorks (2024). *Understanding Support Vector Machine Regression*. Accessed: 2024-08-29. URL: `https://nl.mathworks.com/help/stats/understanding-support-vector-machine-regression.html`.

Matveichev, D. (Aug. 2023). *Binary Classification: Understanding Activation and Loss Functions with a PyTorch Example*. `https://hackernoon.com/binary-classification-understanding-activation-and-loss-functions-with-a-pytorch-example`.

Mhasshah, A., B. N. Bockelmann-Evans, and S. Pan (2017). “Effect of hydrodynamics factors on sediment flocculation processes in estuaries”. In: *Journal Of Soils And Sediments (Print)* 18.10, pp. 3094–3103. DOI: `10.1007/s11368-017-1837-7`.

Mitchell, P.J., M.A. Spence, J. Aldridge, A.T. Kotilainen, and M. Diesing (2021). “Sedimentation rates in the Baltic Sea: A machine learning approach”. In: *Continental Shelf Research* 214, p. 104325. ISSN: 0278-4343. DOI: `https://doi.org/10.1016/j.csr.2020.104325`. URL: `https://www.sciencedirect.com/science/article/pii/S0278434320302788`.

Mohammed, Ammar and Rania Kora (2023). “A comprehensive review on ensemble deep learning: Opportunities and challenges”. In: *Journal of King Saud University - Computer and Information Sciences* 35.2, pp. 757–774. ISSN: 1319-1578. DOI: `https://doi.org/10.1016/j.jksuci.2023.01.014`. URL: `https://www.sciencedirect.com/science/article/pii/S1319157823000228`.

Al-Mukhtar, M. (2019). “Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad”. In: *Environmental Monitoring And Assessment (Print)* 191.11. DOI: `10.1007/s10661-019-7821-5`.

Murphy, Kevin P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press. URL: `probml.ai`.

Musolf, A.M., E.R. Holzinger, and J.D. et al. Malley (2022). “What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics”. In: *Human Genetics* 141, pp. 1515–1528. DOI: `10.1007/s00439-021-02402-z`.

Mussumeci, Elisa and Flávio Codeço Coelho (2020). “Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression”. In: *Spatial and Spatio-temporal Epidemiology* 35, p. 100372. ISSN: 1877-5845. DOI: `https://doi.org/10.1016/j.sste.2020.100372`. URL: `https://www.sciencedirect.com/science/article/pii/S1877584520300502`.

Nadeem (Mar. 2022). *Cost Function & Loss Function*. `https://nadeem.medium.com/cost-function-loss-function-c3cab1ddff44`.

Nda, Muhammad, Mohd Shalahuddin Adnan, Mohd Azlan Bin Mohd Yusoff, and Ramatu Muhammad Nda (2023). “An Overview of Machine Learning Techniques for Sediment Prediction”. In: *Engineering Proceedings* 56.1. ISSN: 2673-4591. DOI: `10.3390/ASEC2023-16599`. URL: `https://www.mdpi.com/2673-4591/56/1/204`.

Neumann, S., A. Kirichek, and A. van Hassent (2024). “Agitation dredging of silt and fine sand with Water Injection Dredging, Tiamat and Underwater Plough: a case study in the Port of Rotterdam”. In: *Journal of Soils and Sediments*. https://doi.org/10.1007/s11368-024-03877-9. DOI: `10.1007/s11368-024-03877-9`.

Nijs, M. A. J. de, J. C. Winterwerp, and J. D. Pietrzak (2010). “The Effects of the Internal Flow Structure on SPM Entrapment in the Rotterdam Waterway”. In: *Journal of Physical Oceanography* 40.11, pp. 2357–2380. DOI: `10.1175/2010JPO4233.1`. URL: `https://doi.org/10.1175/2010JPO4233.1`.

Noori, R., B. Ghiasi, S. Salehi, M. E. Bidhendi, A. Raeisi, S. Partani, R. Meysami, M. H. Mahdian, M. Hosseinzadeh, and S. Abolfathi (2022). “An Efficient Data Driven-Based Model for Prediction of the Total Sediment Load in Rivers”. In: *Hydrology* 9.2. DOI: `10.3390/hydrology9020036`.

NSGI (2024). https://www.nsgi.nl/coordinatenstelsels-en-transformaties/overzicht-coordinatenstelsels.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pham, D. H. B., T. T. Hoang, Q. T. Bui, N. A. Tran, and T. G. Nguyen (2019). "Application of Machine Learning Methods for the Prediction of River Mouth Morphological Variation: A Comparative Analysis of the Da Dien Estuary, Vietnam". In: *Journal Of Coastal Research* 35.5, pp. 1024–1035. DOI: 10.2112/jcoastres-d-18-00109.1.

Phillips, H. (June 2023). *A Simple Introduction to Gradient Descent*. https://medium.com/@hunter-j-phillips/a-simple-introduction-to-gradient-descent-1c8a28b0deb4.

Piraei, R., S. H. Afzali, and M. Niazkar (2023). "Assessment of XGBoost to Estimate Total Sediment Loads in Rivers". In: *Water Resources* 37.13, pp. 5289–5306. DOI: 10.1007/s11269-023-03606-w.

Port of Rotterdam (2022). *Feiten en cijfers*. https://www.portofrotterdam.com/nl/nieuws-en-persberichten/havenbedrijf-rotterdam-test-baggeren-door-waterinjecties#:~:text=Het%20Havenbedrijf%20houdt%20de%20Rotterdamse%20v%20met%20e%20C3%A9%A9%20n%20meter%20bagger.

— (2023). *Eerste verkenning CO2*. Presentation only available within Port of Rotterdam.

— (2024a). https://www.portofrotterdam.com/en/up-to-date-information/weather-tides-and-water-depth.

— (2024b). *Operationeel Stromings Model (Operational Flow Model)*. Software only available within Port of Rotterdam.

— (2024c). *PortMaps*. Software only available within Port of Rotterdam.

Qi, Y. (2012). "Random forest for bioinformatics". In: *Ensemble Machine Learning*. Ed. by C. Zhang and Y. Ma. Berlin: Springer, pp. 307–323.

Rajaee, T. and H. Jafari (2020). "Two decades on the artificial intelligence models advancement for modelling river sediment concentration: State-of-the-art". In: *Journal Of Hydrology* 588, p. 125011.

Rajoub, Bashar (2020). "Chapter 3 - Supervised and unsupervised learning". In: *Biomedical Signal Processing and Artificial Intelligence in Healthcare*. Ed. by Walid Zgallai. Developments in Biomedical Engineering and Bioelectronics. Academic Press, pp. 51–89. DOI: https://doi.org/10.1016/B978-0-12-818946-7.00003-2. URL: https://www.sciencedirect.com/science/article/pii/B9780128189467000032.

Ren, Z., C. Liu, Y. Ou, P. Zhang, H. Fan, X. Zhao, H. Cheng, L. Teng, M. Tang, and F. Zhou (2023). "Deep Learning-Based Simulation of Surface Suspended Sediment Concentration in the Yangtze Estuary during Typhoon In-Fa". In: *Water* 16.1, p. 146. DOI: 10.3390/w16010146.

Restreppo, G.A., W.T. Wood, and B.J. Phrampus (2020). "Oceanic sediment accumulation rates predicted via machine learning algorithm: towards sediment characterization on a global scale". In: *Geo-Mar Lett* 40, pp. 755–763. DOI: 10.1007/s00367-020-00669-1.

Sepehri, Arash, Alex Kirichek, Marcel van den Heuvel, and Mark van Koningsveld (2024). "Smart, sustainable, and circular port maintenance: A comprehensive framework and multi-stakeholder approach". In: *Journal of Environmental Management* 370, p. 122625. ISSN: 0301-4797. DOI: https://doi.org/10.1016/j.jenvman.2024.122625. URL: https://www.sciencedirect.com/science/article/pii/S0301479724026112.

Sharafati, A., S. B. H. S. Asadollahi, D. Motta, and Z. M. Yaseen (2020). "Application of newly developed ensemble machine learning models for daily suspended sediment load prediction and related uncertainty analysis". In: *Hydrological Sciences Journal* 65.12, pp. 2022–2042. DOI: 10.1080/02626667.2020.1786571.

Sharma, Siddharth, Simone Sharma, and Anidhya Athaiya (2020). "Activation Functions in Neural Networks". In: *International Journal of Engineering Applied Sciences and Technology* 4.12. Published online April 2020, pp. 310–316. ISSN: 2455-2143. URL: http://www.ijeast.com.

Singh, Dalwinder and Birmohan Singh (2020). "Investigating the impact of data normalization on classification performance". In: *Applied Soft Computing* 97, p. 105524. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2019.105524. URL: https://www.sciencedirect.com/science/article/pii/S1568494619302947.

Sklearn (2024a). URL: https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-py.

— (2024b). *3.5. Validation curves: plotting scores to evaluate models*. URL: https://scikit-learn.org/stable/modules/learning_curve.html.

Svasek Hydraulics (2024). *Het Operationele Stromingsmodel Rotterdam (OSR)*. https://www.svasek.nl/project/het-operationele-stromingsmodel-rotterdam-osr/.

Tempel, A.H. (2019). "Sediment traps for reducing maintenance dredging costs in the port of Rotterdam". MA thesis. Delft University of Technology. URL: https://repository.tudelft.nl/record/uuid:546e1a84-71b0-4ffb-98fc-137477914bcf.

TEOS-10 Developers (2024). URL: https://teos-10.github.io/GSW-Python/.

TKI Deltatechnologie (2023). *PRISMA 3 – Programma Innovatief Sediment Management voor Havens*. https://tkideltatechnologie.innovemodule.nl/project/prisma-3-programma-innovatief-sediment-management-voor-havens/.

Üstün, B., W.J. Melssen, and L.M.C. Buydens (2007). "Visualisation and interpretation of Support Vector Regression models". In: *Analytica Chimica Acta* 595.1. Papers presented at the 10th International Conference on Chemometrics in Analytical Chemistry, pp. 299–309. ISSN: 0003-2670. DOI: https://doi.org/10.1016/j.aca.2007.03.023. URL: https://www.sciencedirect.com/science/article/pii/S0003267007004904.

Van den Berg, JH, JR Boersma, and A Van Gelder (2007). "Diagnostic sedimentary structures of the fluvial-tidal transition zone—evidence from deposits of the Rhine and Meuse". In: *Netherlands Journal of Geosciences—Geologie en Mijnbouw* 86.3, pp. 287–306.

Walsh, E. S., B. J. Kreakie, M. G. Cantwell, and D. Nacci (2017). "A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system". In: *PLOS ONE* 12.6. DOI: 10.1371/journal.pone.0179473.

Yan, Chunyan, Xiaoyan Shen, Feng Guo, et al. (2019). "A novel model modification method for support vector regression based on radial basis functions". In: *Structural and Multidisciplinary Optimization* 60, pp. 983–997. DOI: 10.1007/s00158-019-02251-5.

Ying, X. (2020). "An Overview of Overfitting and its Solutions". In: *Journal of Physics: Conference Series* 1168.2, p. 022022. DOI: 10.1088/1742-6596/1168/2/022022. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022.

Zhang, Danrong, Nimisha Roy, and J. David Frost (2024). "A data-driven approach to optimize the design configuration of multi-sleeve cone penetrometer probe attachments". In: *Computers and Geotechnics* 169, p. 106248. ISSN: 0266-352X. DOI: https://doi.org/10.1016/j.compgeo.2024.106248. URL: https://www.sciencedirect.com/science/article/pii/S0266352X24001848.

A
Appendix

A.1. Data description

Table A.1: Examples of surveys and order number ('Order'). The 'Function Place' is the code that specifies a surveying zone. 'Completion date' indicates the moment that the survey is conducted and processed.

Order	Ordertype	Start date	Description	Functieplaats	Completion date
70166190	ZM04	24-1-2018	Botlek Centrale Geul	H-L-N-BT-004-PLV-009	2-2-2018
70166666	ZM04	29-1-2018	Botlek Vak 3	H-L-N-BT-004-PLV-017	8-2-2018
70168704	ZM04	8-2-2018	Botlek mond + steinwegkade	H-L-N-BT-004-PLV-014	13-2-2018
70169004	ZM04	14-2-2018	1e Werkhaven	H-L-N-BT-167-PLV-001	20-2-2018
70169857	ZM04	21-2-2018	2e Werkhaven	H-L-N-BT-168-PLV-002	27-2-2018
70169858	ZM04	22-2-2018	3e Petroleumhaven	H-L-N-BT-096-PLV-003	22-2-2018
70170267	ZM04	2-3-2018	Botlek Centrale Geul	H-L-N-BT-004-PLV-009	20-3-2018
70170859	ZM04	13-3-2018	Botlek mond + steinwegkade	H-L-N-BT-004-PLV-014	21-3-2018
70171849	ZM04	5-4-2018	Botlek Vak 3	H-L-N-BT-004-PLV-017	16-4-2018
70172388	ZM04	4-4-2018	Welplaathaven	H-L-N-BT-145-PLV-027	19-4-2018
70172675	ZM04	10-4-2018	1e Werkhaven	H-L-N-BT-167-PLV-001	12-4-2018
70172676	ZM04	22-3-2018	3e Petroleumhaven	H-L-N-BT-096-PLV-003	26-3-2018
...	...	...	...	...	...

Table A.2: Names of measurements stations and their name codes.

Name location	Location code
Botlek Centrale Geul Oost	BOTCGO
Botlek Centrale Geul West	BOTCGW
Botlek mond + zwaaikom	BOTM
Botlek mond Nieuwe Maas	BOTNM
Hartelkering	HARK
Hoek van Holland	HOEK
Radarpost 10	RP10
Rijnhaven	RIJNH
Lekhaven	LEKH
Scheurkade	SCHEUK
2e Werkhaven	2WERKH

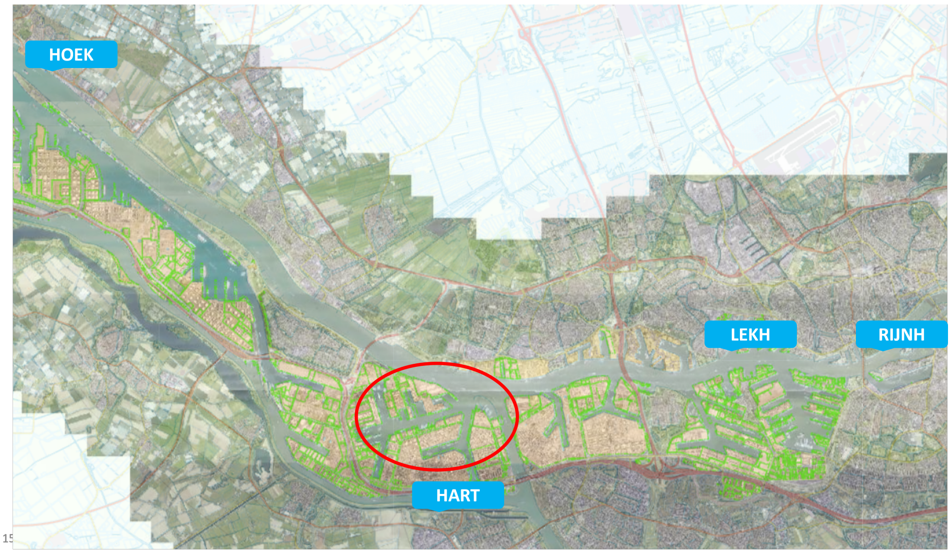


Figure A.2: Locations of temperature measurement stations. The red circle highlighst the Botlek harbour.

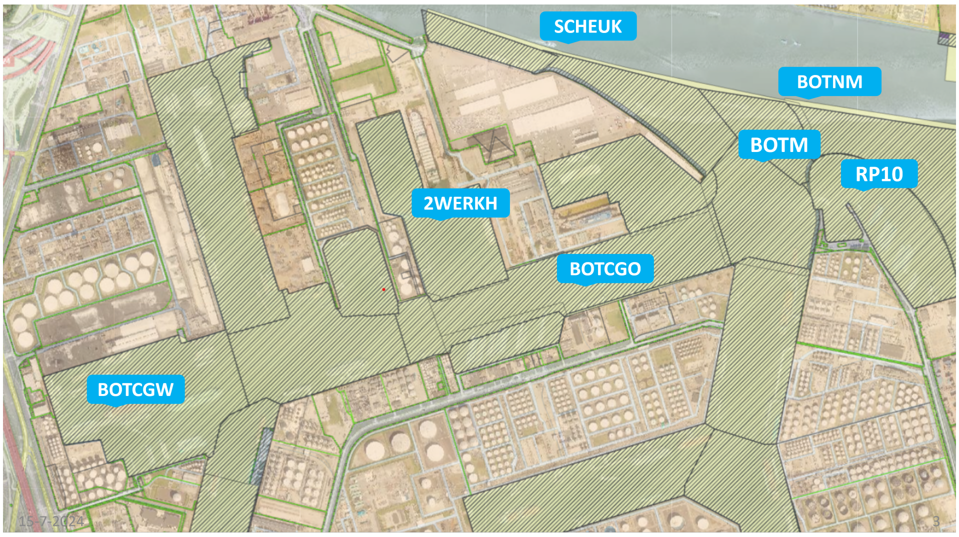
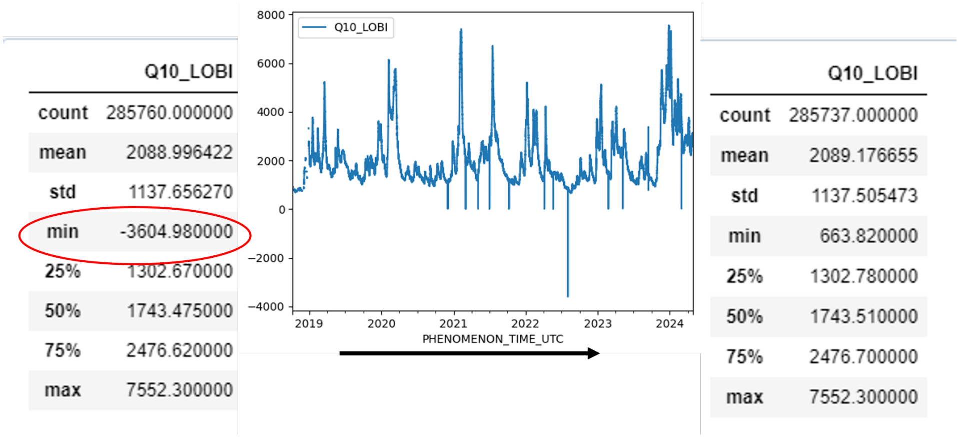


Figure A.1: Measurement and OSR prediction locations in Botlek harbour

B
Appendix

B.1. Parameter characteristics

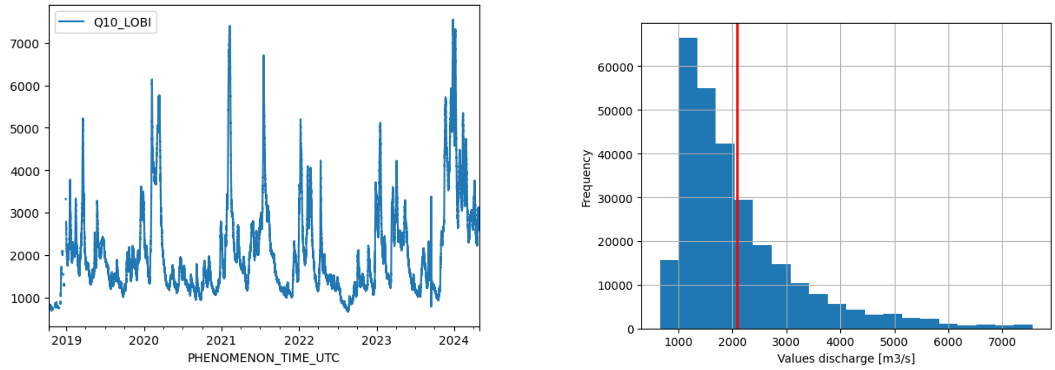
B.1.1. Discharge



	Q10_LOBI
count	285760.000000
mean	2088.996422
std	1137.656270
min	-3604.980000
25%	1302.670000
50%	1743.475000
75%	2476.620000
max	7552.300000

	Q10_LOBI
count	285737.000000
mean	2089.176655
std	1137.505473
min	663.820000
25%	1302.780000
50%	1743.510000
75%	2476.700000
max	7552.300000

Figure B.1: Interpolated all missing values and outliers below minimum discharge of 400 m3/s

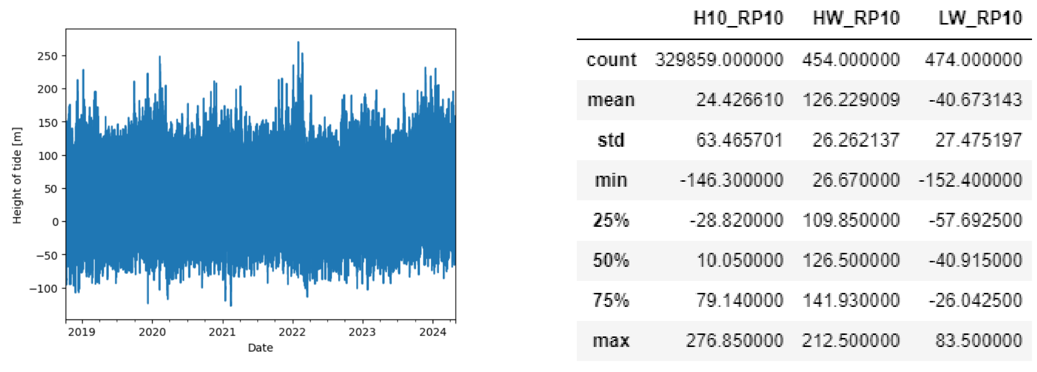


(a)

(b)

Figure B.2: Discharge time series (a) and distribution (b)

B.1.2. Height of tide



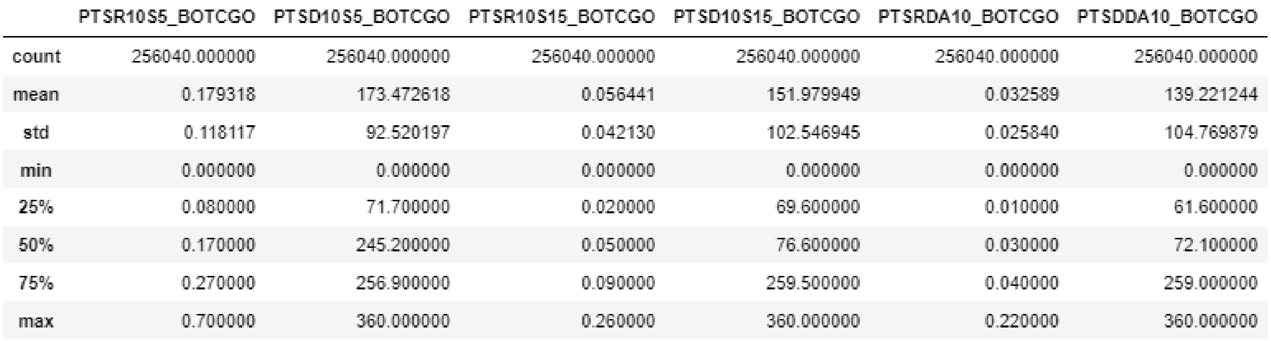
	H10_RP10	HW_RP10	LW_RP10
count	329859.000000	454.000000	474.000000
mean	24.426610	126.229009	-40.673143
std	63.465701	26.262137	27.475197
min	-146.300000	26.670000	-152.400000
25%	-28.820000	109.850000	-57.692500
50%	10.050000	126.500000	-40.915000
75%	79.140000	141.930000	-26.042500
max	276.850000	212.500000	83.500000

(a)

(b)

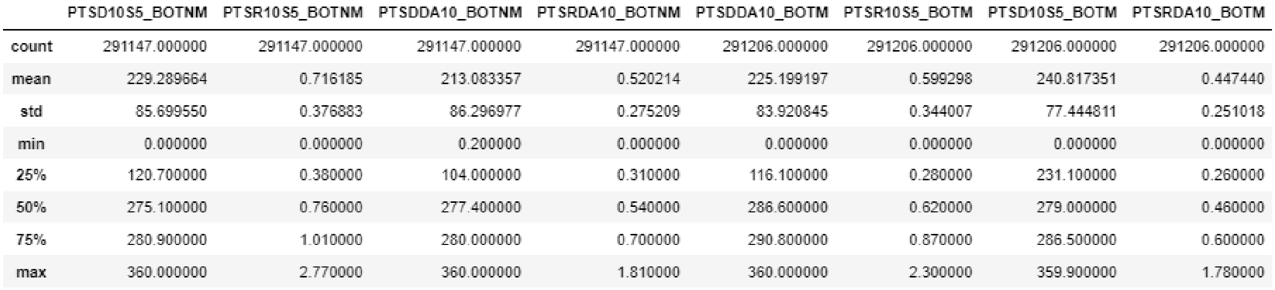
Figure B.3: Tidal variation (a) and characteristics (b)

B.1.3. Tidal stream and direction



	PTSR10S5_BOTCGO	PTSD10S5_BOTCGO	PTSR10S15_BOTCGO	PTSD10S15_BOTCGO	PTSRDA10_BOTCGO	PTSDDA10_BOTCGO
count	256040.000000	256040.000000	256040.000000	256040.000000	256040.000000	256040.000000
mean	0.179318	173.472618	0.056441	151.979949	0.032589	139.221244
std	0.118117	92.520197	0.042130	102.546945	0.025840	104.769879
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.080000	71.700000	0.020000	69.600000	0.010000	61.600000
50%	0.170000	245.200000	0.050000	76.600000	0.030000	72.100000
75%	0.270000	256.900000	0.090000	259.500000	0.040000	259.000000
max	0.700000	360.000000	0.260000	360.000000	0.220000	360.000000

Figure B.4: Characteristics PTSR10 and PTSD10 at BOTCGO.



	PTSD10S5_BOTNM	PTSR10S5_BOTNM	PTSDDA10_BOTNM	PTSRDA10_BOTNM	PTSDDA10_BOTM	PTSR10S5_BOTM	PTSD10S5_BOTM	PTSRDA10_BOTM
count	291147.000000	291147.000000	291147.000000	291147.000000	291206.000000	291206.000000	291206.000000	291206.000000
mean	229.289664	0.716185	213.083357	0.520214	225.199197	0.599298	240.817351	0.447440
std	85.699550	0.376883	86.296977	0.275209	83.920845	0.344007	77.444811	0.251018
min	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	120.700000	0.380000	104.000000	0.310000	116.100000	0.280000	231.100000	0.260000
50%	275.100000	0.760000	277.400000	0.540000	286.600000	0.620000	279.000000	0.460000
75%	280.900000	1.010000	280.000000	0.700000	290.800000	0.870000	286.500000	0.600000
max	360.000000	2.770000	360.000000	1.810000	360.000000	2.300000	359.900000	1.780000

Figure B.5: Characteristics PTSR10 and PTSD10 at BOTNM and BOTM.

BOTCGO

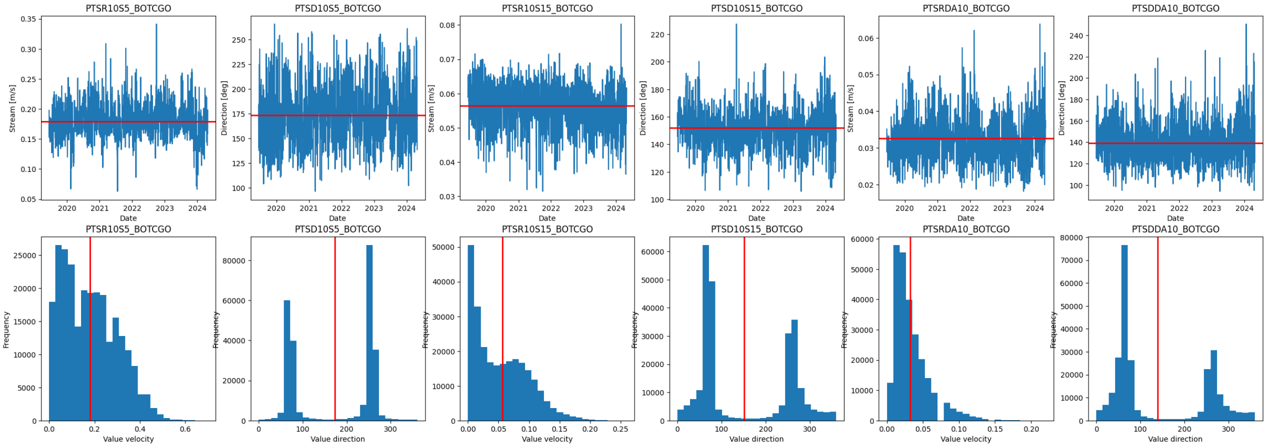


Figure B.6: Time series of PTSR10 and PTSD10 for BOTCGO and their accompanying distributions. The red line indicates the mean.

BOTM

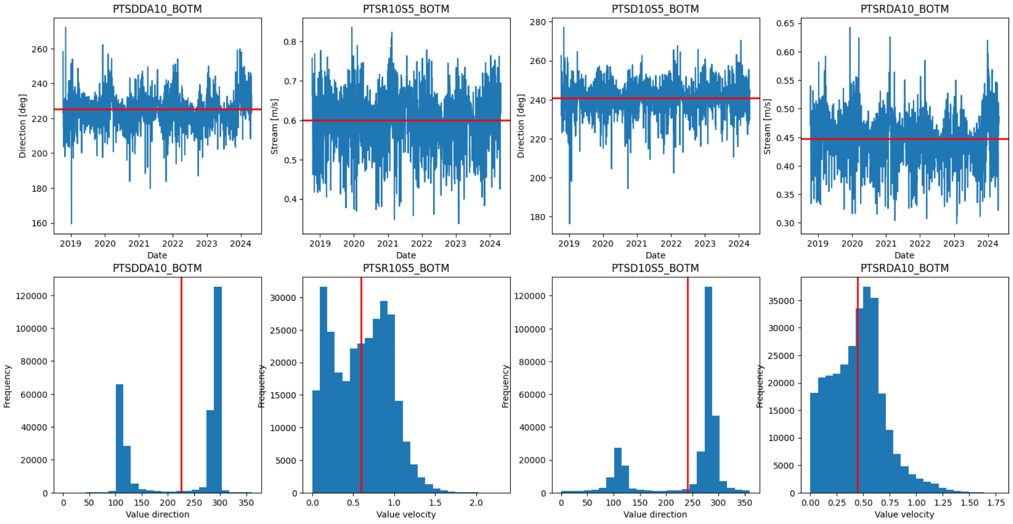


Figure B.7: Time series of PTSR10 and PTSD10 for BOTM and their accompanying distributions. The red line indicates the mean.

BOTNM

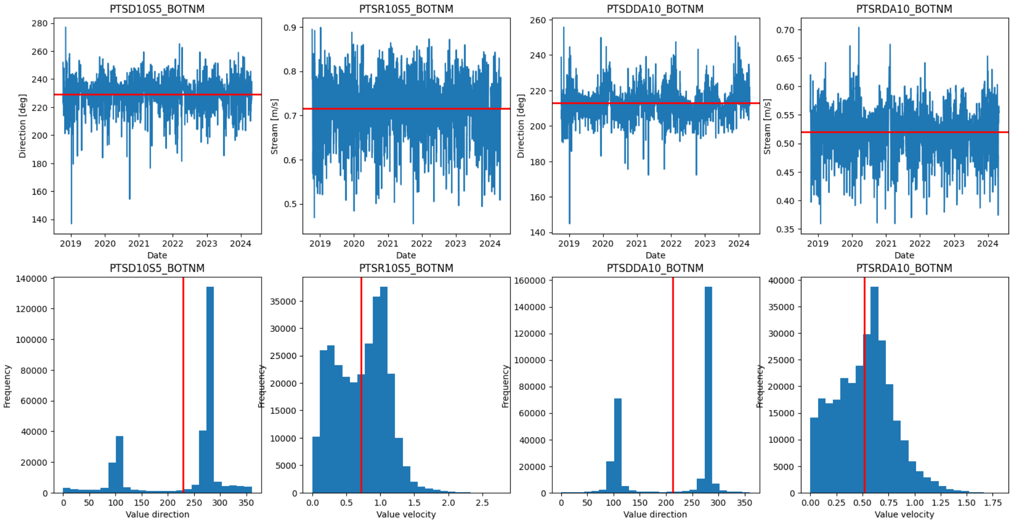
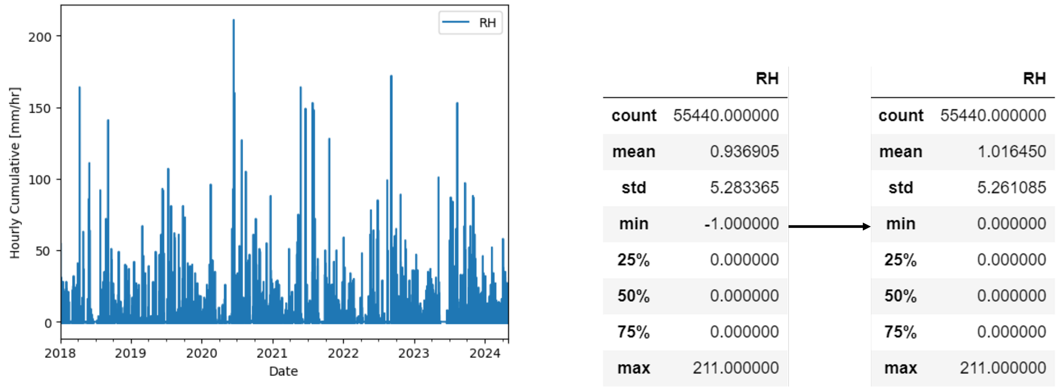


Figure B.8: Time series of PTSR10 and PTSD10 for BOTNM and their accompanying distributions. The red line indicates the mean.

B.1.4. Precipitation



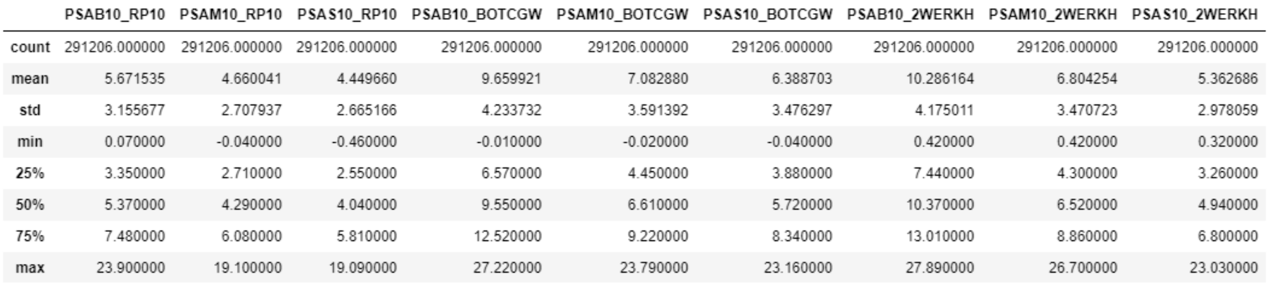
	RH
count	55440.000000
mean	0.936905
std	5.283365
min	-1.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	211.000000

	RH
count	55440.000000
mean	1.016450
std	5.261085
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	211.000000

(a) (b)

Figure B.9: Hourly cumulative rainfall at RP10 (a) and characteristics (b)

B.1.5. Salinity



	PSAB10_RP10	PSAM10_RP10	PSAS10_RP10	PSAB10_BOTCGW	PSAM10_BOTCGW	PSAS10_BOTCGW	PSAB10_2WERKH	PSAM10_2WERKH	PSAS10_2WERKH
count	291206.000000	291206.000000	291206.000000	291206.000000	291206.000000	291206.000000	291206.000000	291206.000000	291206.000000
mean	5.671535	4.660041	4.449660	9.659921	7.082880	6.388703	10.286164	6.804254	5.362686
std	3.155677	2.707937	2.665166	4.233732	3.591392	3.476297	4.175011	3.470723	2.978059
min	0.070000	-0.040000	-0.460000	-0.010000	-0.020000	-0.040000	0.420000	0.420000	0.320000
25%	3.350000	2.710000	2.550000	6.570000	4.450000	3.880000	7.440000	4.300000	3.260000
50%	5.370000	4.290000	4.040000	9.550000	6.610000	5.720000	10.370000	6.520000	4.940000
75%	7.480000	6.080000	5.810000	12.520000	9.220000	8.340000	13.010000	8.860000	6.800000
max	23.900000	19.100000	19.090000	27.220000	23.790000	23.160000	27.890000	26.700000	23.030000

Figure B.10: Characteristics salinty

BOTCGW

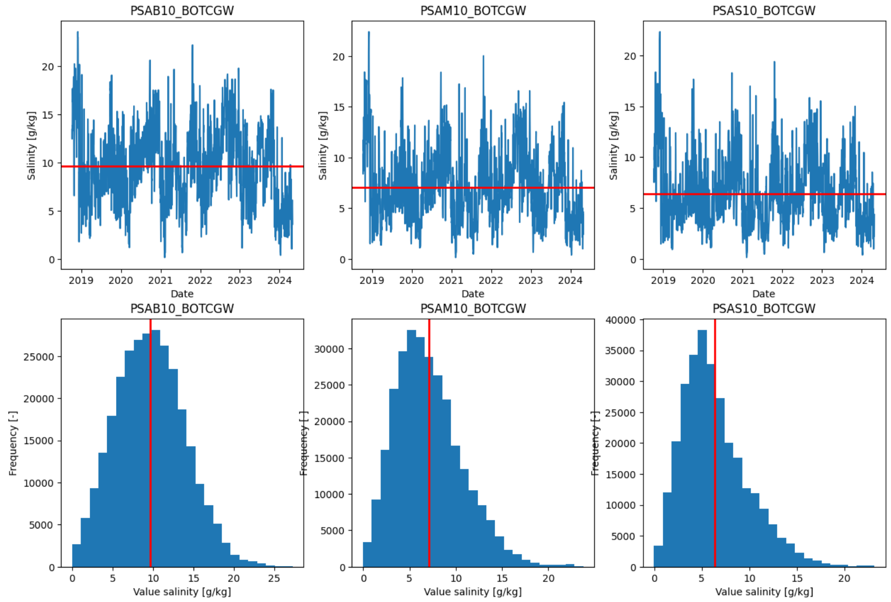


Figure B.11: Time series salinity at BOTCGW and its accompanying distributions. The red line indicates the mean.

RP10

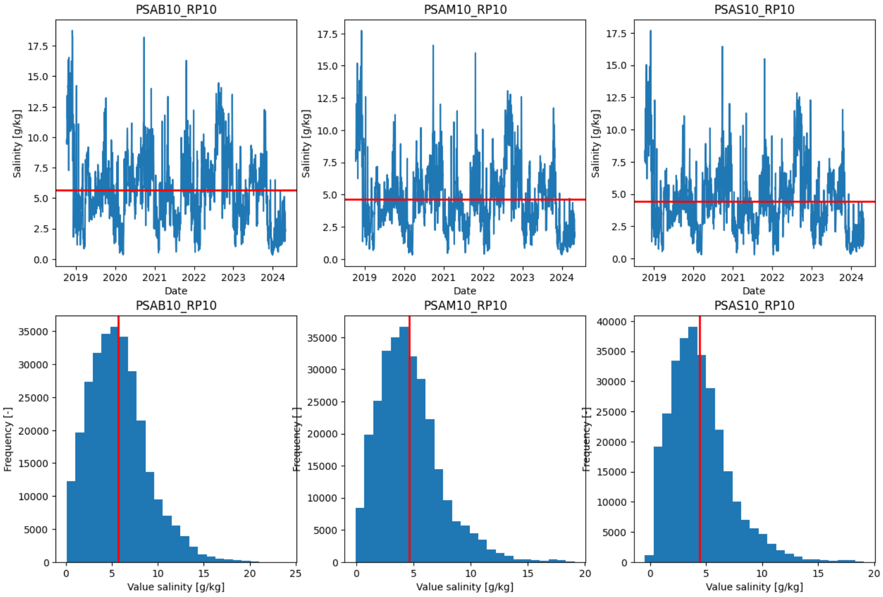


Figure B.12: Time series salinity at RP10 and its accompanying distributions. The red line indicates the mean.

2WERKH

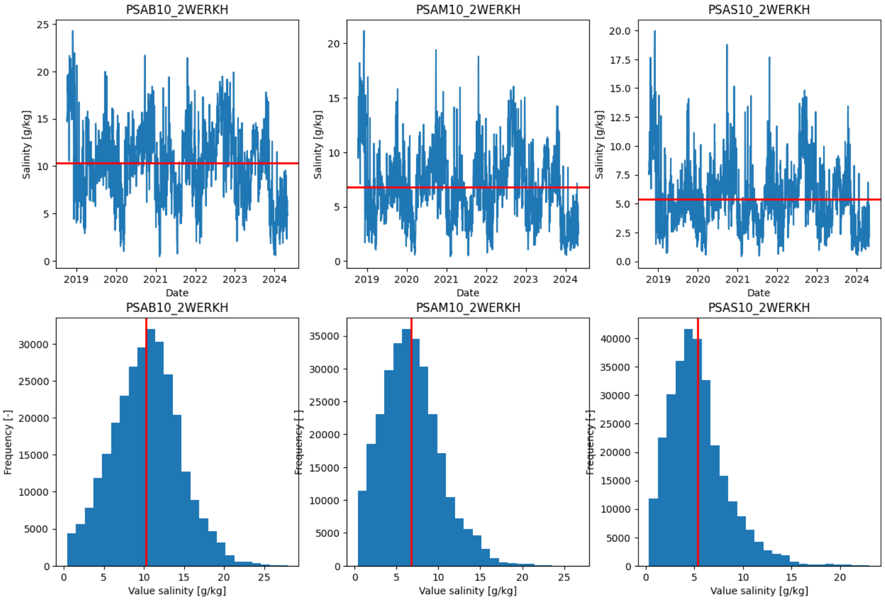
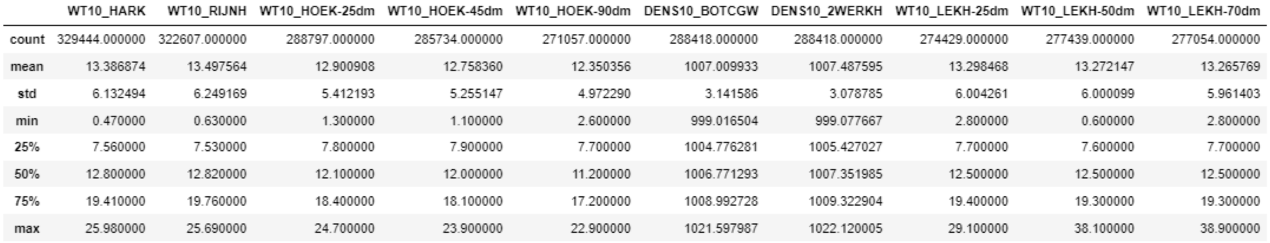


Figure B.13: Time series salinity at 2WERKH and its accompanying distributions. The red line indicates the mean.

B.1.6. Water temperature



	WT10_HARK	WT10_RIJNH	WT10_HOEK-25dm	WT10_HOEK-45dm	WT10_HOEK-90dm	DENS10_BOTCGW	DENS10_2WERKH	WT10_LEKH-25dm	WT10_LEKH-50dm	WT10_LEKH-70dm
count	329444.000000	322607.000000	288797.000000	285734.000000	271057.000000	288418.000000	288418.000000	274429.000000	277439.000000	277054.000000
mean	13.386874	13.497564	12.900908	12.758360	12.350356	1007.009933	1007.487595	13.298468	13.272147	13.265769
std	6.132494	6.249169	5.412193	5.255147	4.972290	3.141586	3.078785	6.004261	6.000099	5.961403
min	0.470000	0.630000	1.300000	1.100000	2.600000	999.016504	999.077667	2.800000	0.600000	2.800000
25%	7.560000	7.530000	7.800000	7.900000	7.700000	1004.776281	1005.427027	7.700000	7.600000	7.700000
50%	12.800000	12.820000	12.100000	12.000000	11.200000	1006.771293	1007.351985	12.500000	12.500000	12.500000
75%	19.410000	19.760000	18.400000	18.100000	17.200000	1008.992728	1009.322904	19.400000	19.300000	19.300000
max	25.980000	25.690000	24.700000	23.900000	22.900000	1021.597987	1022.120005	29.100000	38.100000	38.900000

Figure B.14: Characteristics water temperature at HARK, RIJNH, LEKH, and HOEK. The DENS10 columns are a construct of WT10 and PSAB10 (see Section 3.6).

HARK, RIJNH, AND HOEK

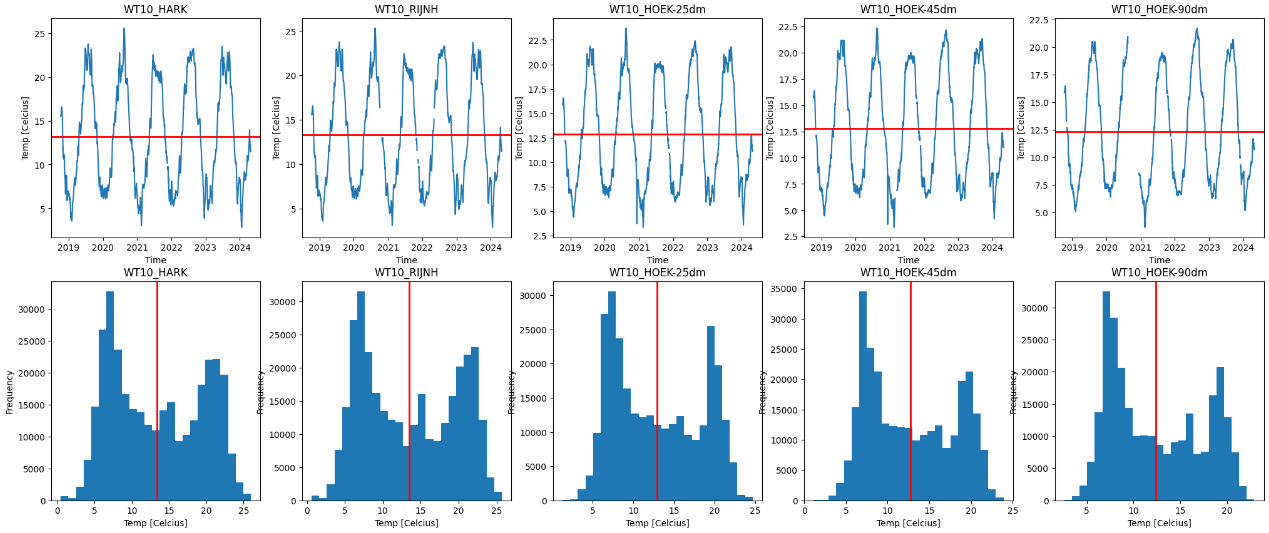


Figure B.15: Time series of water temperate at HARK,RIJNH, and HOEK and the accompanying distributions. The red line indicates the mean.

Lekhaven

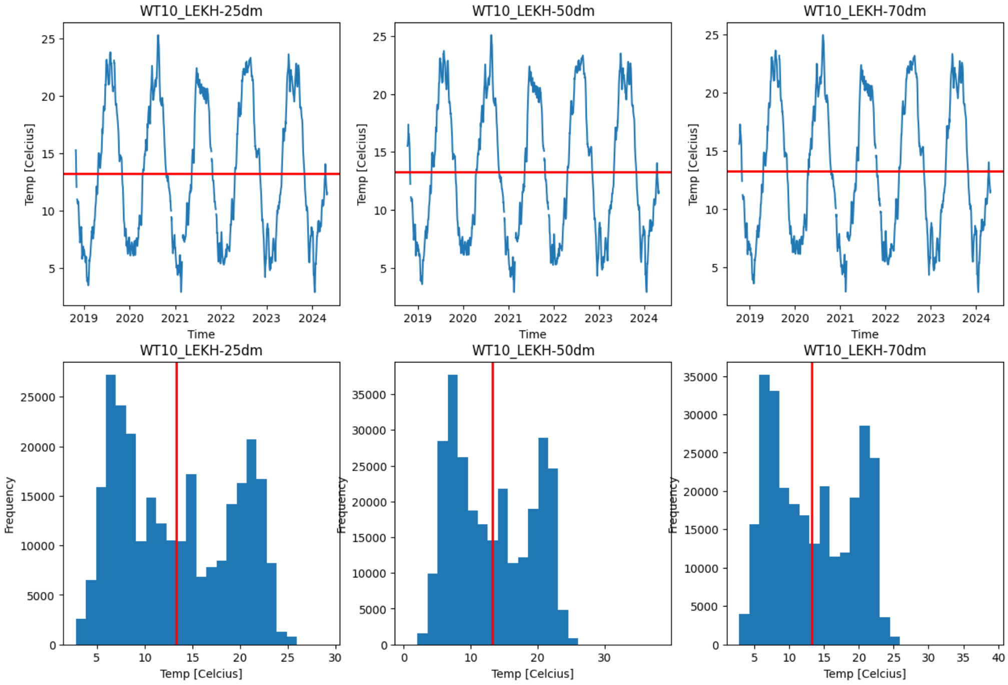
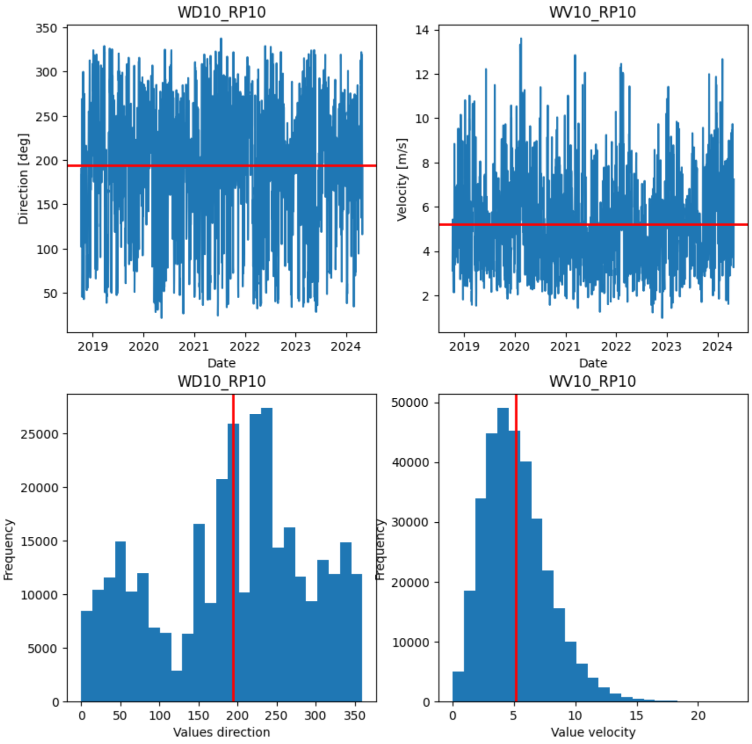
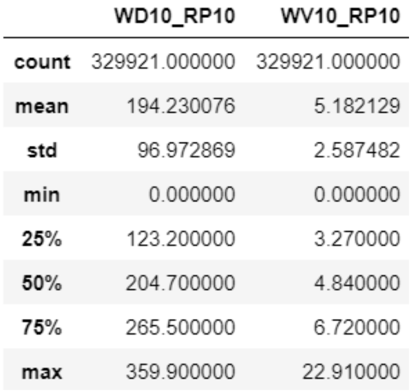


Figure B.16: Time series water temperature at LEKH and the accompanying distributions.

B.1.7. Wind direction and velocity



(a)

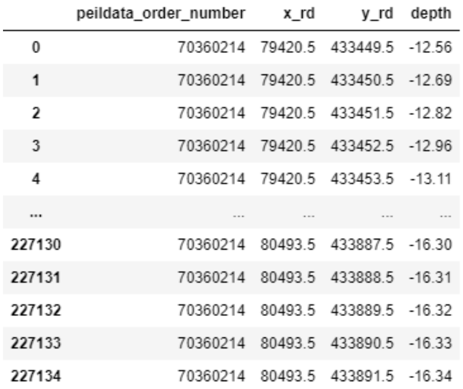


	WD10_RP10	WV10_RP10
count	329921.000000	329921.000000
mean	194.230076	5.182129
std	96.972869	2.587482
min	0.000000	0.000000
25%	123.200000	3.270000
50%	204.700000	4.840000
75%	265.500000	6.720000
max	359.900000	22.910000

(b)

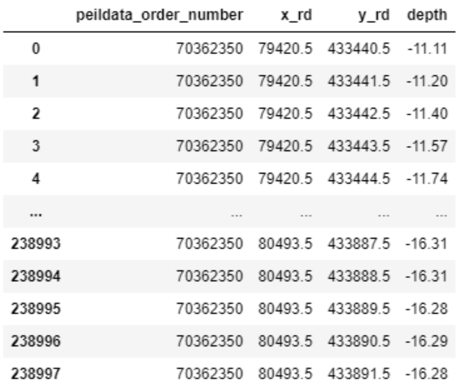
Figure B.17: Time series wind speed and direction at RP10 and the accompanying distributions (a) and the characteristics (b)

B.2. Survey formatting



	peildata_order_number	x_rd	y_rd	depth
0	70360214	79420.5	433449.5	-12.56
1	70360214	79420.5	433450.5	-12.69
2	70360214	79420.5	433451.5	-12.82
3	70360214	79420.5	433452.5	-12.96
4	70360214	79420.5	433453.5	-13.11
...	...	...	...	...
227130	70360214	80493.5	433887.5	-16.30
227131	70360214	80493.5	433888.5	-16.31
227132	70360214	80493.5	433889.5	-16.32
227133	70360214	80493.5	433890.5	-16.33
227134	70360214	80493.5	433891.5	-16.34

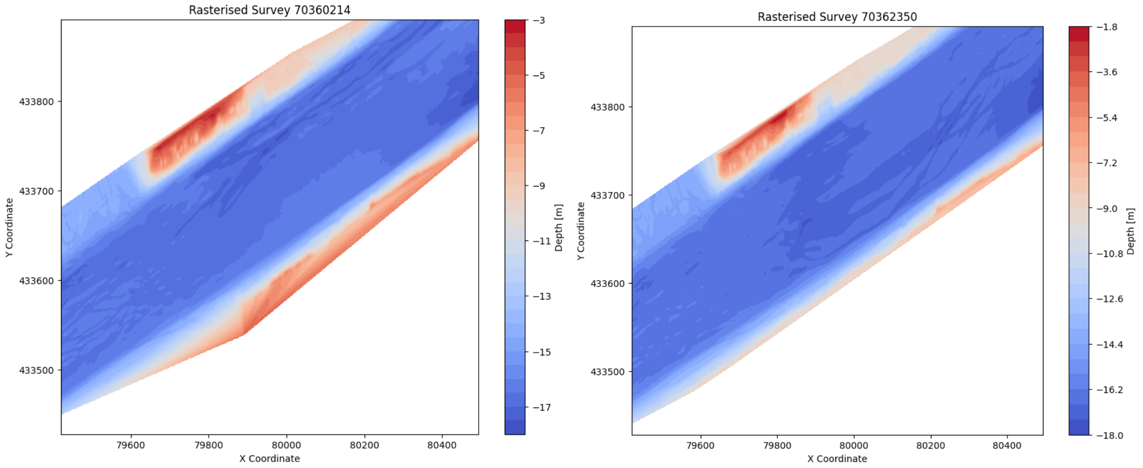
(a) Survey 70360214



	peildata_order_number	x_rd	y_rd	depth
0	70362350	79420.5	433440.5	-11.11
1	70362350	79420.5	433441.5	-11.20
2	70362350	79420.5	433442.5	-11.40
3	70362350	79420.5	433443.5	-11.57
4	70362350	79420.5	433444.5	-11.74
...	...	...	...	...
238993	70362350	80493.5	433887.5	-16.31
238994	70362350	80493.5	433888.5	-16.31
238995	70362350	80493.5	433889.5	-16.28
238996	70362350	80493.5	433890.5	-16.29
238997	70362350	80493.5	433891.5	-16.28

(b) Survey 70362350

Figure B.18: Surveys extracted over same polygon edges NOT WHITIN. The data points are sorted from the smallest x coordinate to the highest. Both tables show different coordinates and survey 70362350 (b) has significantly more data points.



(a) Survey 70360214

(b) Survey 70362350

Figure B.19: Surveys rasterised over same polygon.

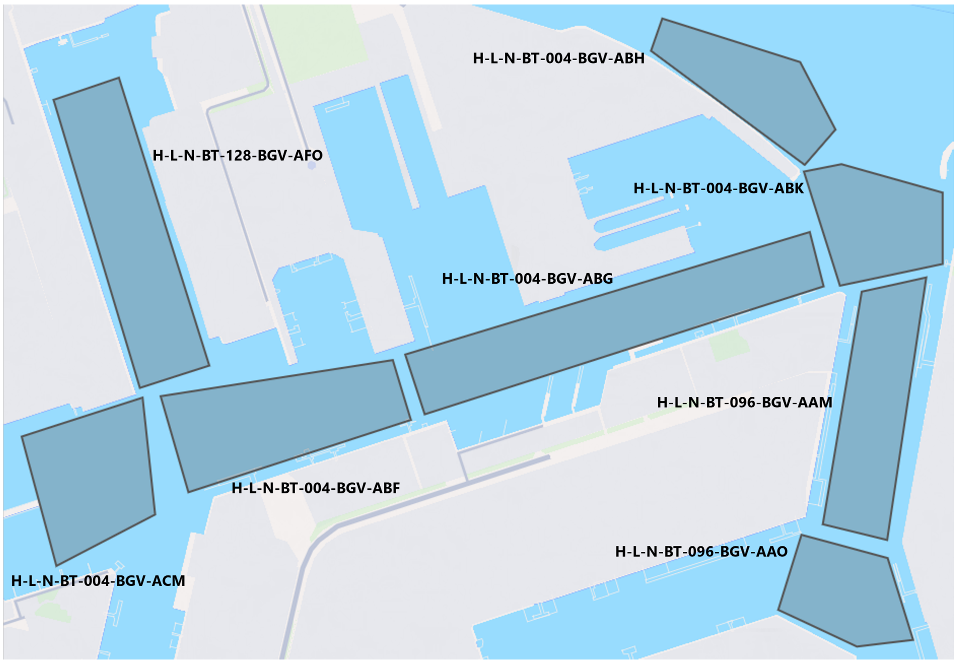


Figure B.20: Polygons defining each dredging area (GEOJSON, 2024)

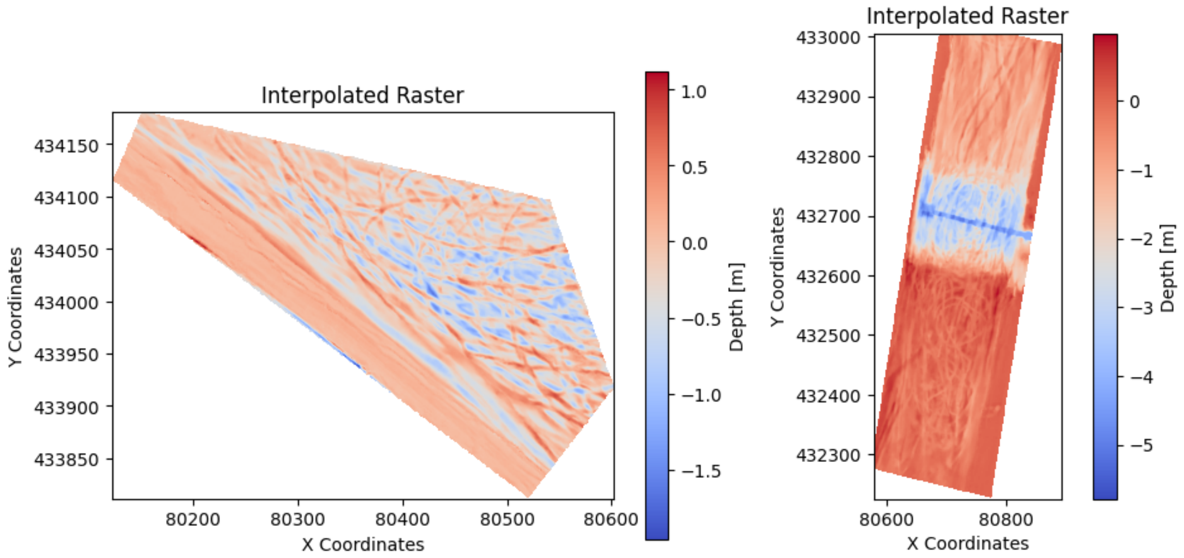
Table B.1: Polygons transformed from ESPG:4326 WGS 84 to EPSG:28992 Amersfoort/RD NEW with EPSG (EPSG, 2024). The polygons can be used in an SQL WITHIN query.

Dredging area	**Polygon**
H-L-N-BT-096-BGV-AAM	’POLYGON((80897 433784, 80685 433784.5, 80570 433054.5, 80775 433004, 80897 433784))’
H-L-N-BT-096-BGV-AAO	’POLYGON((80426 432778, 80659.5 432658, 80832 432703, 80768 432922.5, 80532 432994.,80426 432778))’
H-L-N-BT-004-BGV-ABG	’POLYGON((79264 433405.5, 80586 433794, 80542 433949.5, 79217 433572.5, 79264 433405.5))’
H-L-N-BT-004-BGV-ABK	’POLYGON((80524.5 434175.5, 80628 433791.5, 80951 433858.5, 80952.5 434088.5, 80630.5 434202.5, 80524.5 434175.5))’
H-L-N-BT-004-BGV-ABH	’POLYGON((80121.5 434483.5, 80520 434179.5, 80602.5 434286, 80541 434466, 80150 434550.5, 80121.5 434483.5))’
H-L-N-BT-004-BGV-ABF	’POLYGON((78428 433471.5, 78523.5 433153.5, 79256.5 433394.5, 79209 433574,78428 433471.5))’
H-L-N-BT-004-BGV-ACM	’POLYGON((78150.5 432974, 78427 433129.5, 78410.5 433465.5, 78041 433357.5,78150.5 432974))’
H-L-N-BT-128-BGV-AFO	’POLYGON((78148 434476, 78431 433503, 78644.5 433569.5, 78345.5 434536,78148 434476))’

Table B.2: Measurement stations ascribed to the dredging areas. The three letters refer to the last three letters of the full dredging area name code from Table B.1

Dredging area	**WT10**	**PSAB10**	**H10**	**Q10**	**WV10**	**WD10**	**PTSDDA10**	**PTSRDA10**	**RH**
AAM	RIJNH	2WERKH	RP10	LOBI	RP10	RP10	BOTCGO	BOTCGO	-
AAO	RIJNH	2WERKH	RP10	LOBI	RP10	RP10	BOTCGO	BOTCGO	-
ABG	RIJNH	2WERKH	RP10	LOBI	RP10	RP10	BOTCGO	BOTCGO	-
ABK	RIJNH	2WERKH	RP10	LOBI	RP10	RP10	BOTM	BOTM	-
ABH	RIJNH	2WERKH	RP10	LOBI	RP10	RP10	BOTM	BOTM	-
ABF	RIJNH	BOTCGW	RP10	LOBI	RP10	RP10	BOTCGO	BOTCGO	-
ACM	RIJNH	BOTCGW	RP10	LOBI	RP10	RP10	BOTCGO	BOTCGO	-
AFO	RIJNH	BOTCGW	RP10	LOBI	RP10	RP10	BOTCGO	BOTCGO	-

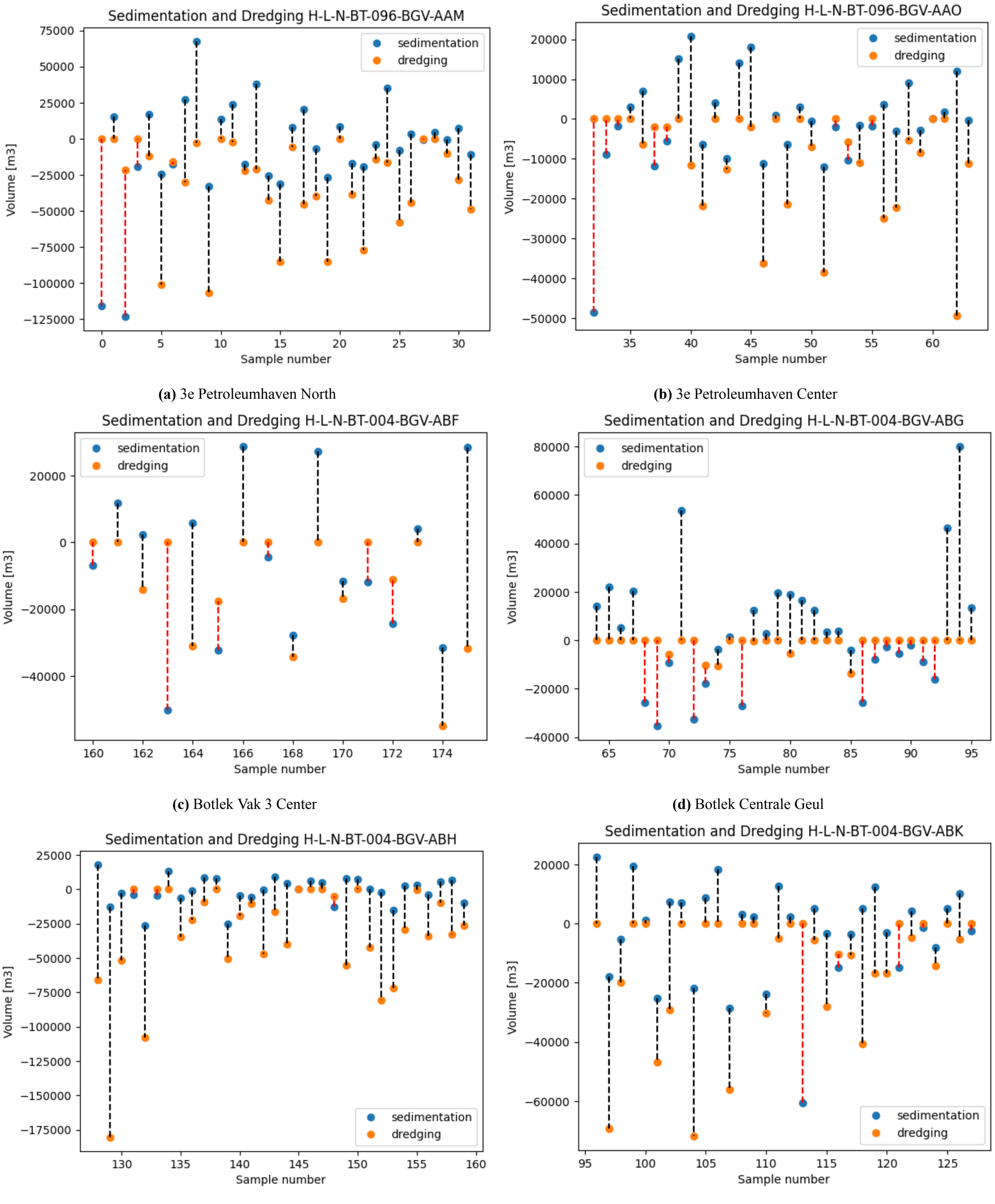
B.3. Dredging data



(a) Botlek Mond 1

(b) 3e Petroleumhaven North

Figure B.21: Examples of sample outliers



(a) 3e Petroleumhaven North

(b) 3e Petroleumhaven Center

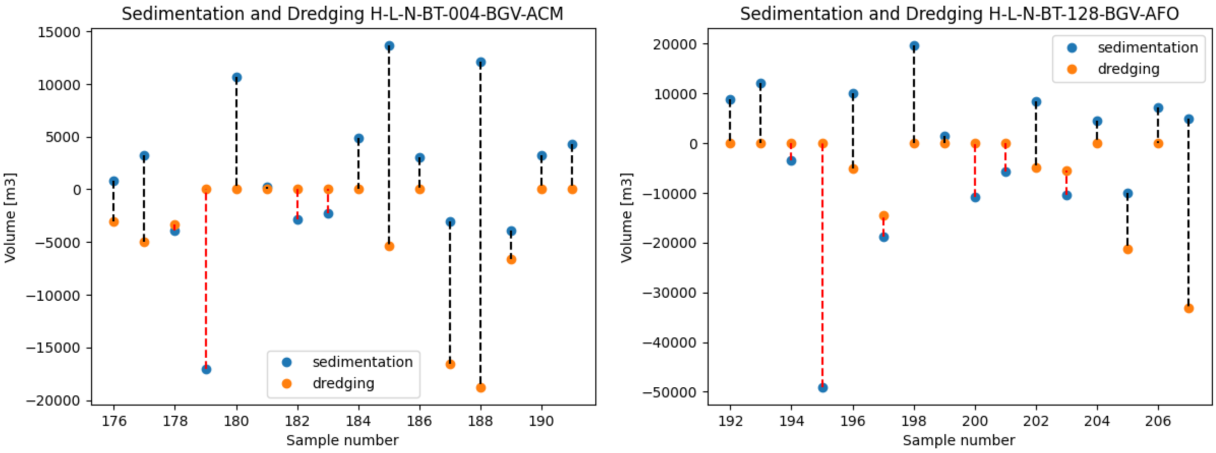
(c) Botlek Vak 3 Center

(d) Botlek Centrale Geul

(e) Botlek Mond 1

(f) Botlek Mond 2

Figure B.22: Difference between the amount of sedimentation and dredging in a sample (part 1). The red line indicates the samples where the sedimentation was lower than the volume that was dredged.



(a) Botlek Vak 3 Left

(b) Botlek Vak 3 North

Figure B.23: Difference between the amount of sedimentation and dredging during a survey period (part 2). The red line indicates the samples where the sedimentation was lower than the volume that was dredged.

C
Appendix

C.1. Modelling phase 1.0

Table C.1: Parameter grid RFR

Hyperparameter	Values
n_estimators	50, 100, 200, 300, 400, 450, 500, 600
max_depth	None, 10, 20, 30, 40, 50
min_samples_split	2, 4, 5, 8, 10
min_samples_leaf	1, 2, 4, 6
max_features	None

Table C.2: Paramater grid SVR

Hyperparameters	Values
kernel	’rbf’
C	0.1, 0.3, 0.6, 1, 10, 100, 1000
epsilon	0.001, 0.01, 0.1, 0.5, 1
gamma	’scale’, ’auto’, 0.001, 0.01, 0.1, 1

C.1.1. Phase 1.0: training on dredging data and SR in m^3

LR with SR in m^3

Table C.3: Phase 1.0 LR with SR in m^3 with erosion samples.

Random state	R2	RMSE	MSE
0	0.20	16008.95	2.56E+08
20	-0.02	12311.89	1.52E+08
42	-0.02	16373.87	2.68E+08

Table C.4: Phase 1.0 LR with SR in m^3 without erosion samples.

Random state	R2	RMSE	MSE
0	0.02	13015.40	1.69E+08
20	0.12	10527.12	1.11E+08
42	0.28	12915.75	1.67E+08

SVR with SR in m^3

Table C.5: Phase 1.0 SVR with SR in m^3 with erosion samples.

Random state	C	epsilon	gamma	CV	R2	RMSE	MSE
0	10	0.01	0.01	0.05	0.18	16186.44	2.62E+08
20	500	0.3	0.0005	0.16	0.05	11891.11	1.41E+08
42	100	0.5	0.001	0.16	-0.01	16299.35	2.66E+08

Table C.6: Phase 1.0 SVR with SR in m^3 without erosion samples.

Random state	C	epsilon	gamma	CV	R2	RMSE	MSE
0	500	0.5	0.0005	0.37	0.05	12789.59	1.64E+08
20	500	0.5	0.0005	0.37	0.10	10652.23	1.13E+08
42	500	0.3	0.0005	0.34	0.35	12273.71	1.51E+08

C.1.2. Phase 1.0: training on dredging data and SR in m^3/day

LR with SR in m^3/day

Table C.7: Phase 1.0 LR with SR in m^3/day with erosion samples.

Random state	R2	RMSE	MSE
0	0.57	533.63	2.85E+05
20	0.60	410.40	1.68E+05
42	0.25	545.80	2.98E+05

Table C.8: Phase 1.0 LR with SR in m^3/day with erosion samples.

Random state	R2	RMSE	MSE
0	0.65	433.85	1.88E+05
20	0.73	350.90	1.23E+05
42	0.57	430.53	1.85E+05

SVR with SR m^3/day

Table C.9: Phase 1.0 SVR with SR in m^3/day with erosion samples.

Random state	C	epsilon	gamma	CV	R2	RMSE	MSE
0	1000	0.3	0.001	0.46	0.57	534.83	2.86E+05
20	500	0.3	0.0005	0.47	0.61	402.07	1.62E+05
42	100	0.3	0.0005	0.55	0.23	553.11	3.06E+05

Table C.10: Phase 1.0 SVR with SR in m^3/day without erosion samples.

Random state	C	epsilon	gamma	CV	R2	RMSE	MSE
0	1000	0.5	0.001	0.46	0.64	437.78	1.92E+05
20	100	0.5	0.01	0.49	0.68	379.81	1.44E+05
42	500	0.3	0.0005	0.50	0.62	404.27	1.63E+05

RFR with SR in m^3/day

Table C.11: Phase 1.0 RFR with SR in m^3/day without erosion samples.

Random state	**n_est**	**min_split**	**min_leaf**	**max_depth**	CV	R^2	$RMSE$	MSE
0	100	4	2	None	0.42	0.59	467.31	2.18E+05
20	10	5	4	10	0.44	0.58	437.64	1.92E+05
42	100	6	6	None	0.48	0.48	475.29	2.26E+05

C.2. Results Phase 1.1: SR in m^3

Individual runs LR Phase 1.1

Table C.12: Phase 1.1 LR runs with daily sample types

Random state	**R2**	**RMSE**	**MSE**
0	-0.5787	22423.44	5.03E+08
20	-1.8185	20449.93	4.18E+08
42	-0.7098	21181.63	4.49E+08

Table C.13: Phase 1.1 LR runs with weekly sample types

Random state	**R2**	**RMSE**	**MSE**
0	-0.2946	18430.79	3.40E+08
20	-0.4435	14634.92	2.14E+08
42	-0.2946	18430.79	3.40E+08

Table C.14: Phase 1.1 LR runs with monthly sample types

Random state	**R2**	**RMSE**	**MSE**
0	0.3870	13972.57	1.95E+08
20	-0.1474	13047.90	1.70E+08
42	-0.0511	16607.83	2.76E+08

Individual runs RFR Phase 1.1

Table C.15: Phase 1.1 RFR runs with daily sample types

Random state	**n_est**	**min_split**	**min_leaf**	**max_depth**	CV	R^2	$RMSE$	MSE
0	500	5	6	30	-0.01	0.30	14928.37	2.23E+08
20	50	5	4	30	0.19	-0.15	13082.237	1.71E+08
42	400	5	2	10	0.19	0.24	14114.31	1.99E+08

Table C.16: Phase 1.1 RFR runs with weekly sample types

Random state	**n_est**	**min_split**	**min_leaf**	**max_depth**	CV	R^2	$RMSE$	MSE
0	450	10	6	20	0.032	0.36	14303.24	2.05E+08
20	300	2	6	50	0.22	-0.02	12272.13	1.51E+08
42	50	4	1	10	0.25	0.20	14476.35	2.10E+08

Table C.17: Phase 1.1 RFR runs with monthly sample types

Random State	**n_est**	**min_split**	**min_leaf**	**max_depth**	CV	R^2	$RMSE$	MSE
0	300	2	6	50	0.030	0.38	14039.13	1.97E+08
20	450	2	1	10	0.24	-0.23	13505.99	1.82E+08
42	50	8	6	20	0.28	0.12	15214.85	2.31E+08

Individual runs SVR Phase 1.1

Table C.18: Phase 1.1 SVR runs with daily sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	1	0.3	0.001	0.11	0.22	15751.09	2.48E+08
20	1	0.3	0.001	0.15	0.14	11328.11	1.28E+08
42	1	0.5	0.001	0.17	0.15	14924.21	2.23E+08

Table C.19: Phase 1.1 SVR runs with weekly sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	5	0.3	0.001	0.21	0.27	15237.17	2.32E+08
20	10	0.3	0.001	0.27	0.18	11029.29	1.21E+08
42	10	0.3	0.001	0.29	0.22	14336.48	2.23E+08

Table C.20: Phase 1.1 SVR runs with monthly sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	5	0.1	’auto’	0.21	0.25	15466.48	2.39E+08
20	1	0.001	’auto’	0.31	0.18	11000.83	1.21E+08
42	5	0.3	0.01	0.32	0.12	15190.18	2.23E+08

C.2.1. Results Phase 1.2: SR in m^3/day

Individual runs LR Phase 1.2

Table C.21: Phase 1.2 LR runs with daily sample types

Random state	R2	RMSE	MSE
0	0.1525	747.45	558678.42
20	-0.1087	681.66	464666.07
42	-0.2513	706.05	498512.76

Table C.22: Phase 1.2 LR runs with weekly sample types

Random state	R2	RMSE	MSE
0	0.5388	551.43	304074.97
20	0.4322	487.83	237978.77
42	0.0526	614.36	377437.71

Table C.23: Phase 1.2 LR runs with monthly sample types

Random state	R2	RMSE	MSE
0	0.6709	465.75	216925.27
20	0.5487	434.93	189164.19
42	0.2307	553.59	306466.73

Individual runs RFR Phase 1.2

Table C.24: Phase 1.2 RFR runs with daily sample types

Random state	n_est	min_split	min_leaf	max_depth	CV	R^2	$RMSE$	MSE
0	450	4	1	30	0.40	0.54	548.46	3.01E+05
20	400	10	6	10	0.45	0.50	456.73	2.09E+05
42	100	8	4	10	0.53	0.13	587.74	3.45E+05

Table C.25: Phase 1.2 RFR runs with weekly sample types

Random state	n_est	min_split	min_leaf	max_depth	CV	R^2	$RMSE$	MSE
0	400	8	2	40	0.45	0.57	532.27	2.83E+05
20	400	10	6	10	0.46	0.55	434.69	1.89E+05
42	100	4	2	30	0.54	0.16	577.86	3.34E+05

Table C.26: Phase 1.2 RFR runs with monthly sample types

Random state	n_est	min_split	min_leaf	max_depth	CV	R^2	$RMSE$	MSE
0	100	5	4	50	0.47	0.57	534.68	2.86E+05
20	400	10	4	50	0.48	0.47	473.35	2.24E+05
42	100	5	4	50	0.58	0.24	549.25	3.02E+05

Individual runs SVR Phase 1.2

Table C.27: Phase 1.2 SVR runs with daily sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	100	0.3	0.001	0.40	0.63	492.96	2.43E+05
20	100	0.3	0.001	0.50	0.34	524.64	2.75E+05
42	100	0.3	0.001	0.51	0.33	516.29	2.67E+05

Table C.28: Phase 1.2 SVR runs with weekly sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	10	0.1	0.001	0.45	0.62	498.40	2.48E+05
20	50	0.001	0.001	0.55	0.52	449.78	2.02E+05
42	100	0.3	0.001	0.55	0.31	524.76	2.75E+05

Table C.29: Phase 1.2 SVR runs with monthly sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	50	0.001	0.001	0.54	0.67	468.21	2.19E+05
20	5	0.1	0.01	0.59	0.62	400.36	1.60E+05
42	50	0.001	0.001	0.63	0.23	553.95	3.07E+05

C.3. Results Phase 1.3: SR in m^3/day and no erosion samples

Table C.30: Expanded SVR parameter grid

Hyperparameters	Values
kernel	’rbf’
C	0.1, 0.3, 0.6, 1, 5, 10, 25, 50, 100, 200, 1000
epsilon	0.0005, 0.001, 0.0015, 0.01, 0.1, 0.3, 0.5, 0.7, 1, 2
gamma	’scale’, ’auto’, 0.0005, 0.001, 0.0015, 0.01, 0.1, 0.5, 1, 1.5

individual runs LR Phase 1.3

Table C.31: Phase 1.3 LR runs with daily sample types

Random state	R^2	RMSE	MSE
0	0.48	529.20	2.80E+05
20	0.19	591.66	3.5E+05
42	-0.01	677.11	4.58E+05
60	0.45	555.56	3.09E+05

Table C.32: Phase 1.3 LR runs with weekly sample types

Random state	R^2	RMSE	MSE
0	0.57	482.63	2.33E+05
20	0.34	548.35	3.01E+05
42	0.42	569.87	3.25E+05
60	0.42	569.87	3.25E+05

Table C.33: Phase 1.3 LR runs with monthly sample types

Random state	R^2	RMSE	MSE
0	0.62	448.85	2.01E+05
20	0.56	446.21	1.99E+05
42	0.53	451.94	2.04E+05
60	0.69	419.34	1.76E+05

Individual runs SVR Phase 1.3

Table C.34: Phase 1.3 SVR runs with daily sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	100	0.1	0.001	0.489	0.614	455.11	2.07E+05
20	100	0.3	0.001	0.508	0.542	455.71	2.08E+05
42	100	0.1	0.001	0.539	0.498	467.58	2.19E+05
60	100	0.3	0.001	0.542	0.649	443.52	1.97E+05

Table C.35: Phase 1.3 SVR runs with weekly sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	50	0.1	0.001	0.582	0.705	397.57	1.58E+05
20	50	0.1	0.001	0.544	0.720	356.31	1.27E+05
42	100	0.1	0.001	0.612	0.471	479.85	2.30E+05
60	100	0.001	0.001	0.601	0.603	471.99	2.23E+05

Table C.36: Phase 1.3 SVR runs with monthly sample types

Random state	C	epsilon	gamma	CV	R^2	$RMSE$	MSE
0	5	0.1	0.01	0.564	0.685	411.12	1.69E+05
20	50	0.001	0.001	0.591	0.731	349.11	1.22E+05
42	10	0.1	0.01	0.665	0.610	411.96	1.70E+05
60	1000	0.1	0.001	0.586	0.739	382.91	1.47E+05

Individual runs RFR Phase 1.3

Table C.37: Phase 1.3 RFR runs with weekly sample types

Random state	n_est	min_split	min_leaf	max_depth	CV	R^2	$RMSE$	MSE
0	50	5	4	30	0.43	0.60	462.29	2.14E+05
20	100	8	6	40	0.45	0.45	500.48	2.50E+05
42	50	4	1	10	0.54	0.42	503.27	2.53E+05
60	50	10	1	10	0.45	0.57	491.40	2.41E+05

Table C.38: Phase 1.3 RFR runs with monthly sample types

Random state	n_est	min_split	min_leaf	max_depth	CV	R^2	$RMSE$	MSE
0	300	10	4	50	0.47	0.62	449.92	2.02E+05
20	50	5	4	30	0.44	0.56	447.36	2.00E+05
42	500	2	1	30	0.50	0.57	430.44	1.85E+05
60	300	2	6	50	0.43	0.57	490.89	2.41E+05

C.3.1. Feature importance scores Phase 1

Scores Phase 1.1

The average feature importance scores per hydro-meteo variable are shown in Figure C.1. Each bar represent the average of the three random states for that sample type. The importance score is an aggregated score, meaning that for the daily and weekly runs, the scores of all lagged features of a specific hydro-meteo variable are aggregated.

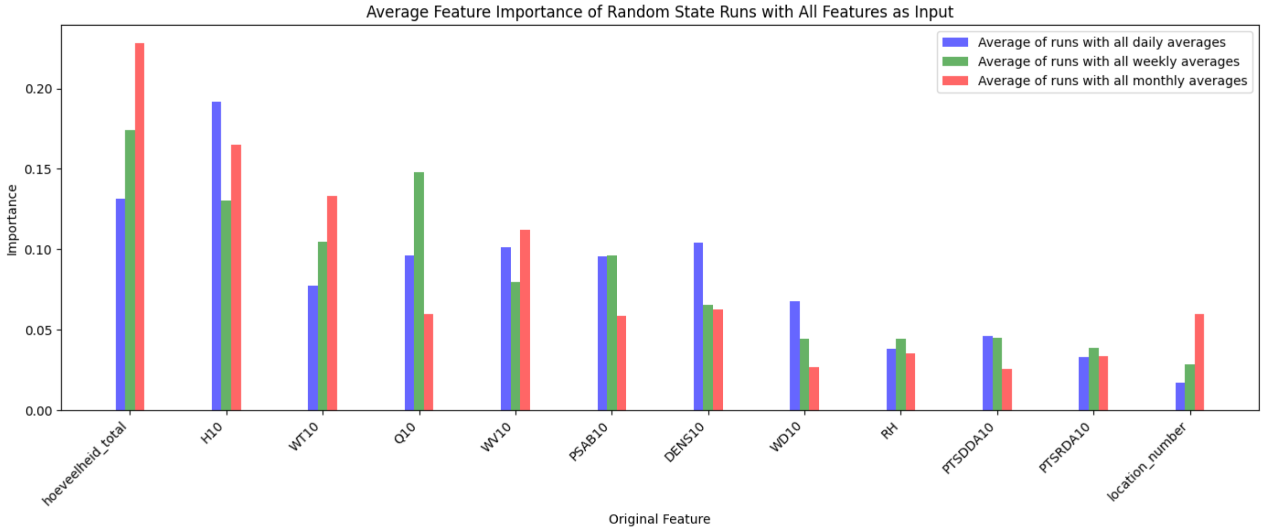


Figure C.1: Feature importance scores of all RFR runs in Phase 1.1. The 'hoeveelheid_total' is the dredged volume in between surveys. The rest of the abbreviations can be found in Table 3.2.

C.3.2. Runs with reduced feature set

Individual runs RFR

Table C.39: RFR runs with reduced feature set with monthly samples

Random state	n_estimators	min_split	min_leaf	max_depth	R2	RMSE	MSE
0	50	5	4	20	0.63	442.71	1.96E+05
20	500	4	2	60	0.65	400.96	1.61E+05
42	600	2	2	40	0.63	403.56	1.63E+05

Individual runs SVR

Table C.40: SVR runs with reduced feature set with monthly samples

Random state	C	epsilon	gamma	R2	RMSE	MSE
0	50	0.1	0.001	0.66	423.91	1.80E+05
20	100	0.1	0.0005	0.72	357.67	1.28E+05
42	50	0.1	0.001	0.67	381.70	1.46E+05

Individual runs LR

Table C.41: LR runs with reduced feature set with monthly samples

Random state	R2	RMSE	MSE
0	0.63	447.68	2.00E+05
20	0.56	445.64	1.99E+05
42	0.60	418.43	1.75E+05

Scores Phase 1.2

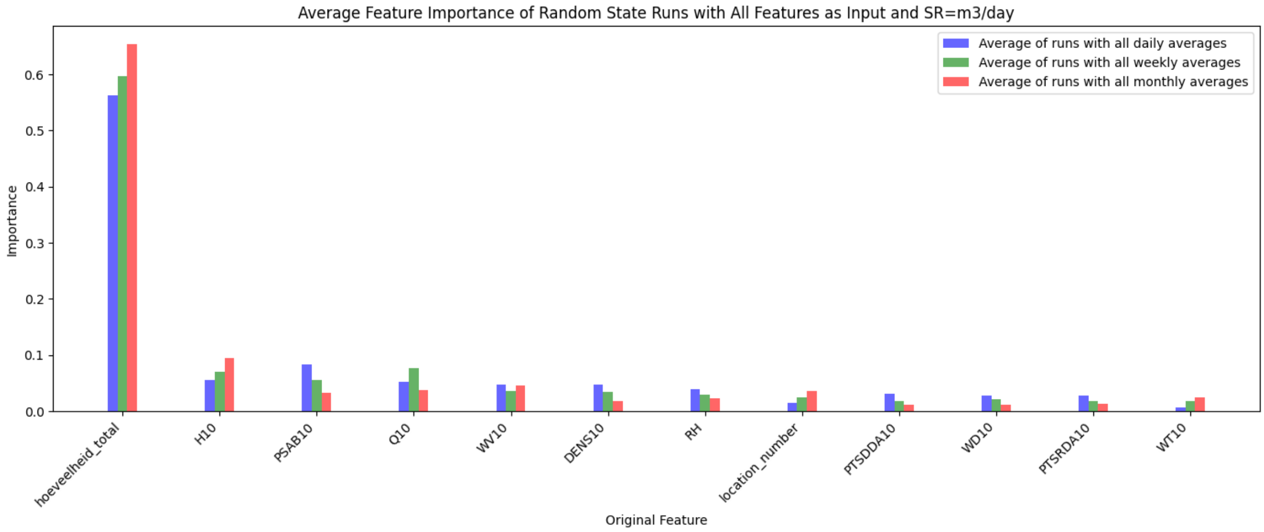


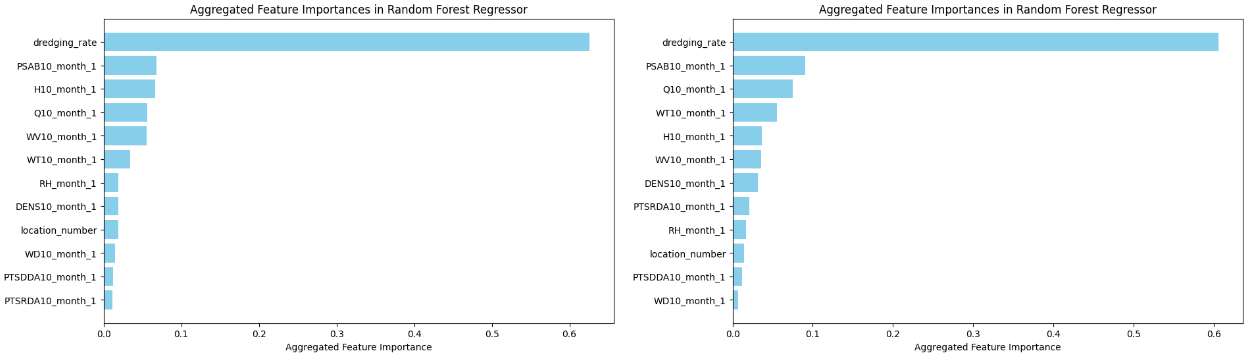
Figure C.2: Average scorse Phase 1.2

D

Appendix

D.1. Result Analysis

Feature importance score analysis



(a) RFR run 1 ($R^2 = 0.61$)

(b) RFR run 2 ($R^2 = 0.57$)

Figure D.2: Scores individual runs

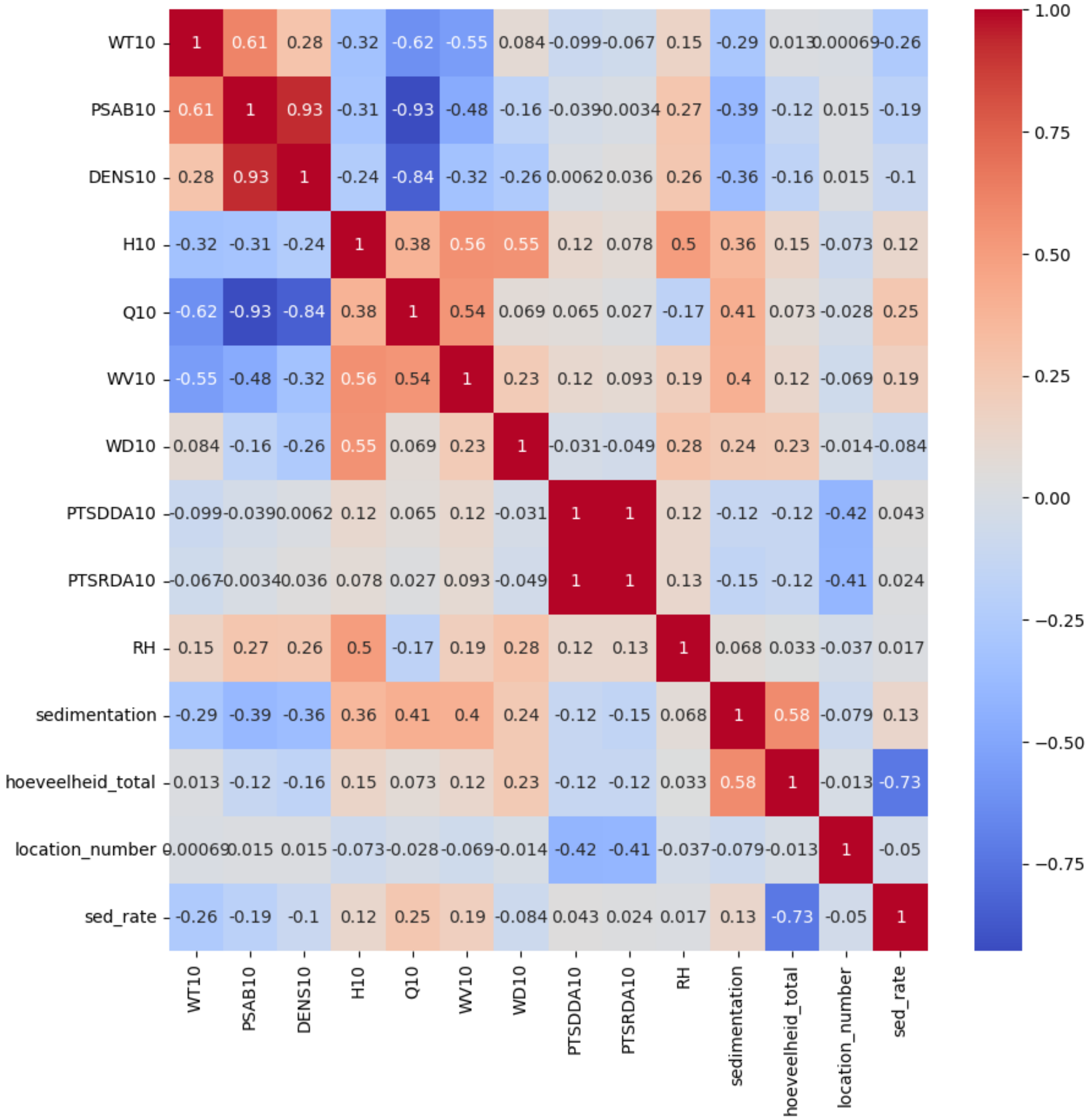


Figure D.1: Pearson correlation matrix of all input variables