NTNU
Norwegian University of
Science and Technology

Analysis of the Boil-Off Phenomenon in Relation to Ambient Conditions

Erling Singstad Paulsen

Master of Science in Engineering and ICT
Submission date: June 2017
Supervisor: Eilif Pedersen, IMT
Co-supervisor: Nicolas Lefebvre, IMT

Norwegian University of Science and Technology
Department of Marine Technology

NTNU Trondheim
Norwegian University of Science and Technology
Department of Marine Technology

MASTER'S THESIS IN MARINE CYBERNETICS SPRING 2017

ERLING SINGSTAD PAULSEN

Analysis of the boil-off phenomenon in relation to ambient conditions

There are both economic and environmental incentives for exploring how ambient conditions, such as wave height and atmospheric temperature affect the production of boil-off during marine transportation of liquefied natural gas (LNG).

Using global atmospheric reanalysis data providing global measurements of ambient conditions, coupled with sensory data from an LNG carrier, the relationship between the boil-off phenomenon and ambient conditions is explored. The data considered spans a period of three years, despite periods of missing data. As the process of boil-off causes the cargo levels to decrease, the relationship is investigated indirectly through the computed change in cargo level by means of statistical learning methods.

Objectives The main objective is to analyze the relationship between boil-off and ambient conditions through the use of real-world sensory data, and assess the relative importance of the ambient conditions.

Work Description The main tasks can be summarized by the following steps:

1. To perform a literature review on the boil-off phenomenon in general and relevant methods for data exploitation.
2. To describe a set of methods that are relevant regarding the data exploitation.
3. To achieve a benchmark of programs and programming languages that can be used for the data analysis.
4. To extract and preprocess the relevant data.
5. To simulate a virtual cargo level.
6. To analyze both the simulated and real-world data using the relevant methods and paying attention to visualization and interpretation of the results.

The report shall be written in English and edited as a research report including literature survey, description of mathematical models, description of algorithms, simulations results, real data analysis, discussion and a conclusion including a proposal for further work. The thesis should be submitted within June 11.

Co-supervisor: Nicolas Lefebvre, IMT

Eilif Pedersen, IMT
Main supervisor

Preface

The work described in this Master's thesis is part of the study program Engineering and ICT with a specialization within Marine Cybernetics at NTNU. The thesis is carried out the spring semester of 2017. With a background in both ICT and marine technology, I wanted to explore the use of statistical learning methods for data analysis using real-world data from the maritime industry. The thesis is a continuation of previous work carried out the fall semester of 2016.

The thesis is written for NTNU in cooperation with DNV GL, who initially provided the data together with relevant background information and interesting problems to explore.

The reader is assumed to have basic knowledge of marine vessels, linear algebra, and statistical analysis.

Acknowledgments

I would like to thank the employees at DNV GL who provided the sensory data and assisted me throughout the project by providing necessary documentation. I would especially like to thank Christos Chryssakis for providing valuable resources and for continuously being involved in the process through Skype meetings.

It has been a great pleasure to have Nicolas Lefebvre as my co-supervisor. He supported me through weekly meetings to discuss the work done and to how to best proceed. I would also like to thank Eilif Pedersen for being my main supervisor. Finally, I would like to thank Christian de Jonge, who has been working with the same data, for our collaboration and numerous discussions throughout the semester.

Trondheim, June 2017
Erling Singstad Paulsen

Summary and Conclusions

The relationship between ambient conditions and the boil-off of liquefied natural gas during marine transportation has been investigated using a data-driven approach. The data consisted of sensory data collected from a vessel over a three year period coupled with global measurements of ambient conditions such as temperature and wave height. Statistical models were trained to predict the change in cargo level from ambient conditions, as the change in cargo level is a direct effect of boil-off. The models were then analyzed to assess the relative importance of the different ambient conditions.

The relevant data were extracted and put on a suitable format. The preprocessing consisted of outlier removal, noise reduction and synchronization of the data. Five datasets were constructed; one for each of the four cargo tanks and one for the cargo levels combined. Preprocessing of the cargo levels reduced the amount of data by about 95 %.

A crude polynomial model was used to simulate a virtual cargo level as a function of the real ambient conditions. The simulated data were inspected using statistical learning methods to verify that the employed methodology would uncover the relationships in the data. The models learned from the data showed strong similarities with the true model and improved the prediction error rate by 87.29 %. Both linear and nonlinear models were trained.

The same methodology was employed on the real-world data. The best linear and nonlinear model reduced the prediction error rate by 45.20 % and 68.45 % respectively. The results verified that ambient conditions can be used to predict boil-off, although it is unlikely that true relationship is linear in nature. The parameter sensitivity of the best linear models was analyzed to assess the importance of the different ambient conditions. The results consistently ranked the ambient temperature and waves as most important, as expected due to heat leakage and sloshing in the tanks. From the best linear model, the change in cargo level was found to be varying linearly with ambient temperature, as shown in the literature, and nonlinearly with waves and wind.

Sammendrag og Konklusjoner

Sammenhengen mellom vær- og sjøforhold og avdampningen av flytende naturgass under marin transportasjon ble analysert med en datadrevet fremgangsmetode. Dataen bestod av sensordata samlet fra et skip over en treårsperiode i kombinasjon med beregninger hentet fra en atmosfærisk reanalyse. Statistiske modeller ble trent opp til å predikere forandringen i lastnivået, ettersom avdampning direkte fører til at lasten avtar. Modellene ble så analysert for å vurdere viktigheten til de forskjellige variablene fra omgivelsene.

Den relevante dataen ble utvunnet og konvertert til et passende format. Forbehandlingen bestod i å fjerne utenforliggere, redusere støy og synkronisere dataen. Fem datasett ble opprettet; ett for hver av de fire tankene og ett for lastnivåene kombinert. Forbehandlingen av lastnivåene reduserte datamengden med omkring 95 %.

En enkel polynomisk modell ble brukt til å simulere et virtuelt lastnivå som en funksjon av forholdene i omgivelsene. Den simulerte dataen ble analysert ved hjelp av statisiske metoder for å verifisere at fremgangsmåten ville avdekke sammenhengene i dataen. Modellene trent opp på dataen viste store likheter med den ekte modellen og forbedret prediksjonsfeilen med 87.29 %. Både lineære og ulineære modeller ble trent opp på dataen.

Den samme fremgangsmåten ble brukt på den ekte dataen. Den beste lineære og ulineære modellen reduserte prediksjonsfeilen med henholdsvis 45.20 % og 68.45 %. Resultatene verifiserte at målinger av omgivelsesforholdene kan brukes til å predikere avdampning, selv om de underliggende sammenhengene trolig ikke er lineære. Parametersensitiviteten til de beste lineære modellene ble analysert for å vurdere viktigheten til de forskjellige variablene. Resultatene viste konsekvent at avdampningen var mest sensitiv til temperatur og bølger, som forventet på grunn av varmelekkasje og skvulping i tankene. I den beste lineære modellen ble det avdekket at avdampningen varierer lineært med temperatur, som tidligere vist i litteraturen, og ulineært med bølger og vind.

Contents

List of Figures

List of Tables

Glossaries and Acronyms

AIS	Automatic Identification System.
BBOG	Ballast Boil-Off Gas.
BOG	Boil-Off Gas.
BOR	Boil-Off Rate.
BPT	Bubble Point Temperature.
BTU	British Thermal Unit.
CBOG	Cargo Boil-Off Gas.
CMA	Central Moving Average.
CV	Cross-Validation.
DOF	Degree of Freedom.
GC	Gas Chromatograph.
GCU	Gas Combustion Unit.
IDE	Interactive Development Environment.
JBOG	Jetty Boil-Off Gas.
kNN	k-Nearest Neighbors.
LNG	Liquefied Natural Gas.
LOO	Leave One Out.
LOWESS	Locally Weighted Scatterplot Smoothing.
MSE	Mean Squared Error.
NaN	Not a Number.
NG	Natural Gas.

PC	Principal Component.
PCA	Principal Component Analysis.
Python	High-level, general-purpose, interpreted programming language.
RMSE	Root Mean Squared Error.
RSS	Residual Sum of Squares.
SAS	Statistical Analysis System.
SVD	Singular Value Decomposition.
TBOG	Tankage Boil-Off Gas.
TPES	Total Primary Energy Supply.

Notation

In order of appearence

L_{OA}	Length overall
L_{PP}	Length between perpendiculars
B_M	Moulded breadth
D_M	Moulded depth
d_D	Design draft
N	Number of samples
p	Number of variables
$\mathbb{R}^i$	An i-dimensional vector of real numbers
$\mathbb{R}^{N \times p}$	A N-by-p matrix of real numbers
X	Generic aspect of a variable
$\mathbf{X}$	Data matrix
x_i	Observation i
y_i	Target value i
$\mathbf{x}_j$	Variable j
$\bar{x}_j$	Mean value of $\mathbf{x}_j$

σ_j^2	Variance of $\mathbf{x}_j$
k	Window size in central moving average
λ	Polynomial order in local regression
α	Smoothing factor in local regression Level of significance Model complexity
$\mathbf{x}'_j$	Transformation of $\mathbf{x}_j$
$f(X)$	A function of X
β	Regression coefficient
$\hat{X}$	Prediction or estimation of X
Var[X]	Variance of X
ϵ	Additive Gaussian noise
z_j	Z-score of regression coefficient j
v_j	jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$
α	Level of significance Model complexity
H_0	Null hypothesis
H_1	Alternative hypothesis
E[X]	Expected value of X
$\mathcal{T}$	Training set
k	Number of neighbors in nearest-neighbors regression
$N_k(x)$	Neighborhood of x of size k in nearest-neighbors regression
$\mathbf{S}$	Covariance matrix between observations x_i and x_k
$L(Y, \hat{f}(X))$	Loss function

$\overline{\text{err}}$	Training error for a statistical model
$\text{Err}_{\mathcal{T}}$	Test error over an independent test set for a statistical model
Err	Expected test error for a statistical model
$\widehat{\text{Err}}$	Estimate of the expected test error for a statistical model
k	Subset size in variable selection
α	Model complexity
$\hat{f}_\alpha$	Model denoted by a model complexity α
K	Number of folds in K-fold cross-validation
M	Number of principal components in principal component analysis
Z_m	Principal component m
$\mathbf{H}_M$	Projection matrix in principal component analysis
$\mathbf{E}$	Residual matrix in principal component analysis
$\mathbf{Z}$	Scores matrix in principal component analysis
$\mathbf{V}_M$	Loadings matrix in principal component analysis
$\mathbf{U}$	Orthogonal matrix whose columns are left singular vectors in singular value decomposition
$\mathbf{V}$	Orthogonal matrix whose columns are right singular vectors in singular value decomposition
$\mathbf{D}$	Diagonal matrix whose elements are singular values in singular value decomposition
RH	Relative humidity
P_w	Water vapour pressure
P_{ws}	Water vapour saturation pressure
A	Costant in equation (4.2)

m	Costant in equation (4.2)
T_n	Temperature costant in equation (4.2)
T	Temperature in equation (4.3)
T_d	Dew point temperature in equation (4.3)
r	Length of vector defined by eastward and northward wind components
α	Angle of vector defined by eastward and northward wind components
r	Pearson correlation coefficient
d_i	Distance vector
t	Threshold in nearest-neighbors outlier detection
Δy	Computed change in cargo level
y_{prev}	Previous cargo level measurement
P_{atm}	Atmospheric pressure
T_{atm}	Atmospheric temperature
$H_{1/3}$	Significant wave height
T_1	Mean wave period
v_{wind}	Wind speed
$S(\omega)$	Wave spectrum
α, β	Parameters for a Beta distribution
y_0	Initial cargo level
Δy_{tot}	Total change in cargo level
y_{tot}	Total cargo level

Chapter 1

Introduction

Natural gas (NG) accounts for almost a quarter of the global energy demand, of which 9.8 % is supplied in its liquid form (LNG) by means of marine transportation[1,2]. LNG is stored in heavily insulated tanks as a cryogenic liquid during marine transportation, kept at −162°C through the process of auto-refrigeration. Due to heat leakage and sloshing in the tanks, boil-off gas (BOG) is created through the evaporation of LNG at the surface. The excessive BOG needs to be removed from the tanks to maintain a safe operating pressure. Several methods exist to take care of the BOG, as it can be (1) released directly into the atmosphere, (2) used for propulsion by dual-fuel engines, (3) burnt in a gas combustion unit, and (4) reliquefied and returned to the cargo tanks.

The total LNG trade in 2015 was at 244.8 million tons[2]. Due to boil-off, as much as 2-6 % of the total cargo is lost during a typical voyage[3]. With an average price close to 5 USD per million BTU (British thermal unit) worldwide as of April 2017[4], the cost of boil-off exceeds 1.2 billion USD yearly. Furthermore, if (1) or (3) is used for BOG disposal, a significant amount of greenhouse gases are released into the atmosphere. Thus there are both economic and environmental incentives for exploring how ambient conditions, such as wave height and atmospheric temperature affect the production of BOG.

The relationship between the boil-off phenomenon and ambient conditions is explored using global atmospheric reanalysis data coupled with sensory data from an LNG tanker. The considered data spans a period of three years. As the process of boil-off causes the cargo levels to decrease, the relationship is investigated indirectly through the computed change in cargo level by means of statistical learning methods such as principal component analysis (PCA) and linear and nonlinear regression.

1.1 Background

This thesis is a continuation of a project thesis carried out the fall semester of 2016 where sensory data from the vessel in consideration were analyzed using several statistical learning methods. The problem of boil-off was not analyzed specifically, as the project mainly focused on the understanding, demonstration and visualization of the methods applied. However, it provided the necessary background for the methodology in this thesis, as well as a good understanding of the sensory data and vessel in consideration.

To my knowledge, real-world data has not been used to investigate the effect of ambient conditions on the production of BOG in the way presented here. A literature review and a general background covering the problem of BOG in the LNG supply chain is presented in depth in Chapter 2.

1.2 Objectives

The main objective is to analyze the relationship between BOG and ambient conditions through the use of real-world sensory data. The main objective can be split into two parts: (1) assess the predictive capabilities and relative importance of the ambient conditions on the change of cargo level, and (2) compare the uncovered relationships with previous work done.

1.3 Limitations

There are several limitations of importance in the presented work: (1) sensory data from only one vessel is considered, (2) the exact geometry and measures of the cargo tanks is unknown, (3) the specific composition and quality of the LNG for each voyage is unknown, (4) boil-off is investigated through its effect on the cargo levels, (5) the change in cargo level may not only be contributed to natural boil-off, since the vessel supports the use of forced BOG for propulsion, and finally (6) real-world sensory data is always to some degree contaminated by noise and faulty values. Limitations will also be addressed throughout the thesis.

Due to limitations (1)-(6), the results presented here will at best be suggestive of the effect of ambient conditions on boil-off, but will still provide a solid foundation for how one can

proceed in future work.

1.4 Approach

The approach is data-driven and can be summarized by the following steps: (1) the relevant data is extracted, preprocessed and put on a suitable format, (2) a virtual cargo level is simulated using ambient conditions and a crude model for cargo level change, and (3) both the simulated and real-world data are analyzed in an unsupervised and a supervised framework using statistical learning methods. This approach is used to verify the applied methodology with the simulated data since the underlying model is known. It also allows for comparison of results between the two cases.

1.5 Structure of the Report

The remainder of the thesis is organized as follows: Chapter 2 gives an overview of the vessel and provides the necessary background for understanding the boil-off phenomenon in the LNG supply chain. Chapter 3 gives an introduction to methods in statistical learning and covers the core methods used in the exploratory data analysis. Both supervised and unsupervised learning will be considered, as well as data preprocessing. Model selection and validation will also be discussed briefly. Chapter 4 covers the necessary preprocessing of the data, from raw data to a synchronized, preprocessed dataset ready for analysis. Chapter 5 explores the underlying relationship between the individual cargo levels and the ambient conditions using methods from Chapter 2. The methods are applied to both ideal, simulated data and real-world data. In Chapter 6 the results are summarized and discussed, and recommendations for further work are presented.

Chapter 2

Vessel and Data Description

This chapter gives an overview of the vessel and provides the necessary background for understanding the boil-off phenomenon in the LNG supply chain. A brief description of the different datasets is also given. Section 2.1 covers some key aspects of LNG, its properties and its market, while Section 2.2 presents a literature review of the boil-off phenomenon in particular. Sections 2.3 and 2.4 describe the vessel and the datasets that will be used. In Section 2.6 we briefly discuss the temporal coverage of the datasets. Section 2.7 explores different relevant software platforms.

2.1 LNG Overview

Natural gas (NG) is a nontoxic, colorless, odorless and noncorrosive fossil fuel. It consists primarily of methane (about 90 %) but commonly contains ethane, propane, butane and trace amounts of nitrogen and carbon dioxide (CO2). NG is often described as the cleanest fossil fuel and compared to coal, gas is nearly half as emission intensive on average. In 2014 coal represented 29 % of the total primary energy supply (TPES) and accounted for 46 % of the global CO2 emissions. Similarly in 2014 NG represented 21 % of the world TPES and accounted for 19 % of the emissions[1].

While NG accounts for almost a quarter of the global energy demand, only 9.8 % of the NG is supplied in its liquid form, LNG as of 2015[2]. Transportation of NG by pipelines is preferred up to distances of 2000 km, after which the costs grow significantly faster than the costs of transporting it as LNG[5]. In its liquid form, the original volume is reduced by a factor of 600, allowing for more economical transportation over long distances. The LNG supply chain consists of extraction and production of NG, liquefaction, marine transportation and

storage of LNG, re-gasification and delivery of NG to consumers[5]. This thesis will primarily focus on the marine transportation of LNG. As of January 2016 the global LNG fleet consisted of 410 vessels, with 17 countries exporting and 33 countries importing LNG in 2015. The global trade of LNG reached 245 million tons in 2015, with an expected growth of 46 % by 2021[2].

The composition of the LNG depends on the NG source and the liquefaction pre-treatment and liquefaction process. Typical thermo-physical properties of LNG are presented in Table 2.1. Since the price of LNG depends on its energy content, it is important to determine the quality and composition of the LNG at the port during unloading and loading. Density is commonly used for classification of LNG, and we can differentiate between heavy, medium or light LNG[6]. The typical composition and density of heavy, medium and light LNG is presented in Table 2.2.

Parameter	**Value**
Boiling point	-160°C to -162°C
Density	$425-485$ kg/m^3
Specific heat capacity	$2.2-3.7$ kJ/kg°C
Higher heat value	$38-44$ MJ/m^3

Table 2.1: Thermo-physical properties of LNG[5].

Composition [%]	**LNG Light**	**LNG Medium**	**LNG Heavy**
Methane	98.00	92.00	87.00
Ethane	1.40	6.00	9.50
Propane	0.40	2.00	2.50
Butane	0.10	0.00	0.50
Nitrogen	0.10	1.00	0.50
Density [kg/m^3]	427.74	445.69	464.83
Higher heat value [MJ/m^3]	40.64	41.94	44.42

Table 2.2: Classification of LNG by density[6].

In general, there exist several types of LNG containment systems. In the International Gas Carrier Code (IGC Code), Chapter 4 (resolution MSC.370(93)) the International Maritime Organization (IMO) gives the following classification[7]:

- *Independent tanks*: Self-supporting tanks that do not form part of the ship's hull and

are not essential to the hull strength.

- *Membrane tanks*: Non-self-supporting tanks that consist of a thin liquid and gastight layer (membrane) supported through insulation by the adjacent hull structure.
- *Integral tanks*: Tanks that form a structural part of the hull and are influenced in the same manner by the loads that stress the adjacent hull structure.
- *Semi-membrane tanks*: non-self-supporting tanks in the loaded condition and consist of a layer, parts of which are supported through insulation by the adjacent hull structure.

When carrying liquid low temperature cargo one has to ensure that the cargo is protected by a partial or complete secondary barrier (except independent tank type C), that proper insulation is installed to minimize heat flux into the tanks and to prevent cold spots in the hull structure, and that the vapour pressure P_0 in the tanks are kept below a certain limit (0.025 - 0.07 MPa)[7].

During loading, IMO requires a default filling limit of 98 % of the total tank volume at the reference temperature. This is to prevent the entry of LNG into the ventilation pipeline and from spilling into the surrounding hull structure. In no case should the filling limit exceed 99.5 % at the reference temperature. During unloading the industry practice is to retain 5 % of the total tank volume as a heel to maintain the pressure and temperature in the tanks. It is also normal to use the heel to spray the tanks to keep them cool. Without heel, the tanks would get warm and excess boil-off would occur at the start of next loading[3].

2.2 The Boil-Off Phenomenon

As previously mentioned, LNG is stored as a cryogenic liquid at −162°C in heavily insulated tanks. The liquid remains at its bubble point temperature (BPT), the temperature at a given pressure where the first bubble of vapor is formed, through a process known as auto-refrigeration. Due to imperfect insulation and the large temperature differential between the tank and the ambient, the LNG continuously absorbs heat. The absorbed heat evaporates liquid at the surface with no visible bubble formation[3]. This is known as boil-off. The excess BOG is withdrawn from the tank, causing the temperature and pressure to fall. This reduction again increases the temperature differential and the heat influx. Auto-refrigeration occurs when the gas is withdrawn at a rate so that cooling exceeds the heat available from

ambient sources[8].

BOG alters the quality of the LNG through a process known as aging or weathering. Table 2.3 presents the boiling point (at atmospheric pressure), higher heat value and molar mass for the different LNG constituents. From this table, one can see that nitrogen evaporates first if present in the LNG. Since nitrogen is an inert gas, it causes the higher heating value of the LNG to increase. The next component to evaporate is methane, which has the lowest molar mass and highest higher heating value and accounts for roughly 90 % of the volume. As the lightest constituents of the LNG evaporates first, the LNG becomes heavier over time. Aging or weathering refers to this process of change in the LNG due to boil-off. Aging is therefore important in the LNG trade, as it directly affects the quality and thus the pricing of the LNG.

Constituent	**Boiling point [°C]**	**Higher heat value [MJ/kg]**	**Molar mass [g/mol]**
Methane	-161.5	55.39	16.04
Ethane	-88.8	51.63	30.07
Propane	-42.04	50.16	44.10
Butane	-0.5	50.34	58.12
Nitrogen	-196	N/A (inert gas)	28.01

Table 2.3: Boiling point, higher heat value and molar mass of different LNG constituents[9–11].

BOG is present throughout the LNG supply chain. First, LNG is stored in cryogenic tanks with imperfect insulation both at production plants and receiving terminals. The BOG produced during storage is called tankage BOG (TBOG). From the production plants, the LNG is loaded into tankers, and at the other end unloaded to the receiving terminals. BOG produced during loading and unloading is called jetty BOG (JBOG). BOG is produced in both laden and ballast conditions during a voyage, called cargo BOG (CBOG) and ballast BOG (BBOG) respectively. In this thesis, we look specifically into the CBOG and how it relates to ambient conditions.

Most of the BOG in the supply chain is produced during the marine transportation of LNG. For a 21-day voyage, it is typical with a boil-off rate (BOR) of $0.1 - 0.15$ % of the full cargo content per day. While the BOR varies significantly with different voyages, the amount of BOG produced can be as high as $2 - 6$ % of the total cargo in a typical voyage[3]. Dobrota et al.[5] and Hasan et al.[3] list the following important factors for the production of BOG:

- Heat ingress into cargo tanks due to the large temperature differential.
- Cooling of a ship's tanks during ballast voyages, achieved by spraying LNG in the upper

part of the tanks.

- Sloshing of cargo in partially filled tanks due to the action of waves.
- LNG composition and quality, in particular nitrogen content.
- The overall thermal transmittance of the tanks.
- Operating pressure in the tanks.

In particular, Hasan et al.[3] look into the boil-off losses in LNG transportation using an extensive dynamic simulation of boil-off during loading, unloading, and transportation. By varying ambient temperature, tank pressure, overall thermal transmittance, LNG composition and voyage length they provide insightful data on the boil-off dynamics. They show that nitrogen content has a significant effect on CBOG, with room for optimizing the nitrogen content for minimized boil-off. Generally, the CBOG reduces as nitrogen content increases. Furthermore, they show that CBOG increases nonlinearly with operating pressure if nitrogen is present in the LNG, with stronger nonlinearities for longer voyages. They also show that CBOG increases linearly with ambient temperature. Sea conditions were neglected in the simulations due to the complex and stochastic nature of tank sloshing. To our knowledge, the effect of different sea conditions on boil-off has not been analyzed in the literature.

2.3 Vessel Description

The vessel in consideration is approximately 300 meter long and transports LNG from port to port. Some common measurements of the vessel are presented in Table 2.4. Here L_{OA} is the overall length, L_{PP} the length between perpendiculars, B_M the molded breadth, D_M the molded depth and d_D the design draft. It has four cargo tanks of the membrane type supported by the adjacent hull structure. The containment system has a total capacity of 162574 m^3 and a total primary area of 27526 m^2. This roughly translates into a capacity of around 73000 tons. The capacities and areas of the individual tanks are specified in Table 2.5. Cargo tank no. 2, 3 and 4 are identical, while tank no. 1 is smaller and has less than half the capacity of the others. Tank no. 1 is located in the forward part of the ship and tank no. 4 is located in the aft part adjacent to the engine room, as illustrated in Figure 2.1. Figure 2.2 shows a cross-section of the vessel, displaying the octagonal shape of the tank cross-sections. Unfortunately, we do not have data on the exact geometric measures of the tanks.

Parameter	Length [m]
L_{OA}	≈ 295.0 m
L_{PP}	284.0 m
B_M	43.4 m
D_M	26.0 m
d_D	11.5 m

Table 2.4: Common measurements of the vessel.

Cargo tank no.	Capacity	Primary area
1	21395 m^3	4529 m^2
2	46880m^3	7666 m^2
3	46880 m^3	7666 m^2
4	46880 m^3	7666 m^2

Table 2.5: Cargo tanks capacity and primary area.

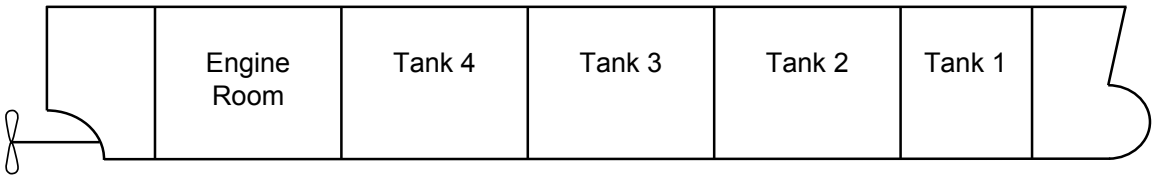

Figure 2.1: Illustration of cargo tank placement on the vessel.

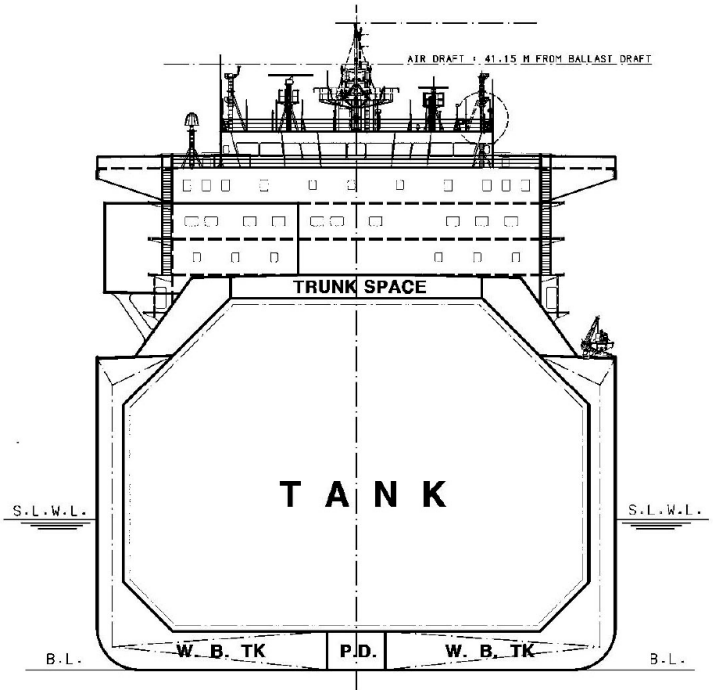

Figure 2.2: Vessel cross-section.

The tanks are protected by the twin-membrane system NO96 developed by GTT[12]. The NO96 membrane system is a cryogenic liner directly supported by the vessel's inner hull. The liner includes two identical metallic membranes and two independent insulation layers:

1. A 0,7 mm thick Invar (36 % nickel-steel alloy) membrane in direct contact with the LNG.
2. A 230 mm insulation layer made of prefabricated plywood boxes filled with expanded perlite.
3. A second 0,7 mm thick Invar membrane, causing redundancy in case of leakage.
4. A second 300 mm insulation layer made of prefabricated plywood boxes filled with expanded perlite.

The insulation space is occupied by nitrogen gas and has a total volume of 10102 m^3. The overall thermal transmittance is $U = 0.105$ W/m^2K. In the capacity calculations for the cargo handling equipment, it is specified a nominal BOR of 0.150 %/day of the total initial cargo volume, with a nominal production of 4275 kg/h for 98.5 % tank filling. Furthermore, it is specified that the LNG filling level should always be between 10 % and 80 % of the cargo tank height with a normal operating pressure of 1060 mBar. A summary of the LNG containment system is presented in Table 2.6.

Parameter	**Value**
Tank type	Membrane (NO96 by GTT)
Total capacity	162574 m^3
Total insulation thickness	0.53 m
Overall thermal transmittance	0.105 W/m^2K
Nominal BOR	0.150 %/day
Nominal BOG production	4275 kg/h
Maximum allowable filling	98.5 % of total capacity
Operating pressure	1060 mBar

Table 2.6: Summary of LNG containment system.

The vessel is equipped with a dual-fuel propulsion system with four main generator engines from Wärtsilä. Two of the engines have an output of 11000 kW and the other two have an output of 5500 kW, with a total of 33000 kW. The engines can run either on natural gas, light fuel oil or heavy fuel oil and are designed to provide the same output regardless of the fuel. This allows the engines to use the excessive BOG for propulsion. To analyze the properties of the BOG entering the engines, a gas chromatograph (GC) has been installed. The GC measures various properties of the BOG such as composition, density, and heating value.

When the ship operates at low speeds it is not able to use all the BOG for propulsion and the

remaining BOG is handled by a gas combustion unit (GCU). The GCU burns the excessive BOG and releases the by-products into the atmosphere. For this particular vessel, during a laden voyage, as much as 4 tons of BOG is burnt and released into the atmosphere every hour. In cases where the need for propulsion exceeds that of available BOG, forced BOG can be taken from the tanks.

2.4 Vessel Data

The vessel data provided consist of two different datasets,

1. Automatic identification system (AIS) data, providing the geographical position of the vessel in decimal degrees roughly three times per hour.
2. General vessel data, providing various parameters such as cargo levels, speed over ground and atmospheric temperature with varying logging frequency.

Table 2.7 shows the variables extracted from the datasets. By limiting the analysis to a subset of the available data, the preprocessing and analysis of the data can be carried out more efficiently. The different variables have been selected based on the nature of the boil-off phenomenon and the quality of the data together with insights into the datasets given by DNV GL.

Dataset	**Variable**	**Unit**
AIS data	Latitude	Decimal degrees
	Longitude	Decimal degrees
Vessel data	Atmospheric temperature	C
	Atmospheric pressure	mbar A
	Cargo level, tank 1	m
	Cargo level, tank 2	m
	Cargo level, tank 3	m
	Cargo level, tank 4	m

Table 2.7: Selected variables from the general vessel data.

2.5 Atmospheric and Ocean Wave Reanalysis Data

The wave and weather data used throughout the thesis is extracted from the ERA-Interim dataset. ERA-Interim is the latest global atmospheric and ocean wave reanalysis produced

by the European Centre for Medium-Range Weather Forecasts (ECMWF). The dataset covers the period from 1979 until today and is continuously updated in real time[13].

Reanalyses are created with an unchanging data assimilation scheme and models which are fed all available observations at each time step. With an unchanging framework, the reanalysis provides dynamically consistent states at each time step[14]. Since one need to know the initial conditions of a model to perform forecasting, data assimilation is used to estimate the initial conditions from observations. This is usually a sequential process where the previous model forecast is compared with new observations, which allow for an update of the model state[15]. As described by Dee et al.[14] the key strengths of reanalyses are that they provide global datasets with consistent spatial and temporal resolution over long periods of time while incorporating millions of observations into a stable data assimilation system. On the other hand, one should be aware that observational constraints, either temporal or spatial, can affect the reanalysis reliability.

The ERA-Interim reanalysis dataset has a temporal resolution of six hours with data given at 00:00, 06:00, 12:00 and 18:00 every day. The data is represented on a $0.125° \times 0.125°$ latitude-longitude grid of the Earth's surface. Figure 2.3 illustrates the resolution of the spatial grid. The first coordinate represents the longitudinal position in decimal degrees, while the second coordinate represents the latitudinal position in decimal degrees. Positive latitudes are north of the equator and positive longitudes are east of the Greenwich Meridian. At the equator one degree longitude corresponds to a longer distance than near the poles, leading to larger grid cells near the equator, and smaller grid cells near the poles. If the vessel is located between four vertices of the grid near the equator, the maximum spatial error between the position of the vessel and a data point will be approximately 10 km.

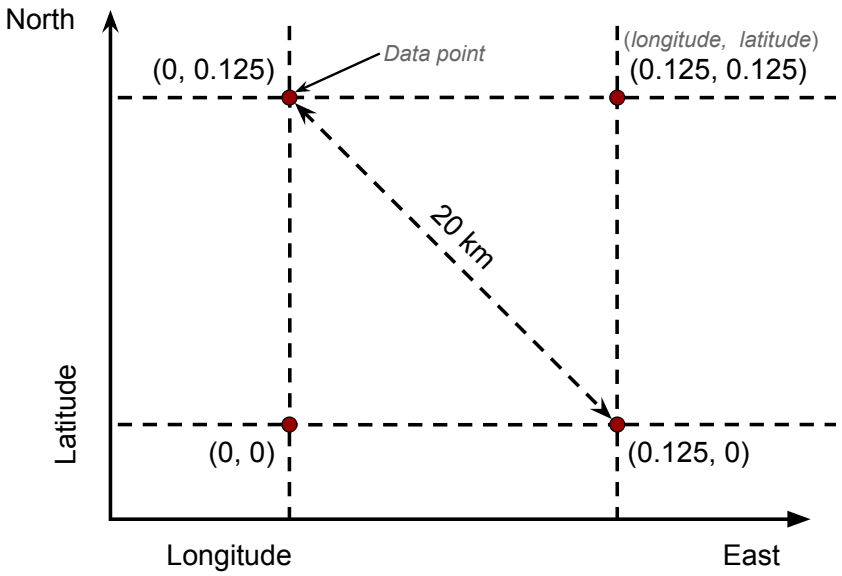

Figure 2.3: Longitude latitude grid for ERA-Interim.

Eight useful variables from both the atmospheric and ocean-wave model of the ERA-Interim dataset have been extracted as shown in Table 2.8. The ocean-wave model uses a total of 30 wave frequencies and 24 wave direction at each grid vertex[16]. All variables are produced by the analysis and not the forecast model.

Variable	Model	Unit
2 metre temperature	Atmospheric	K
2 metre dew point temperature	Atmospheric	K
10 metre U wind component (northward)	Atmospheric	m/s
10 metre V wind component (eastward)	Atmospheric	m/s
Mean sea level pressure	Atmospheric	Pa
Mean wave direction	Ocean-wave	degrees
Mean wave period	Ocean-wave	s
Significant height of combined wind waves and swell	Ocean-wave	m

Table 2.8: Selected variables from the ERA-Interim dataset.

2.6 Data Availability

As we are combining data from three different sources, the availability of the different datasets will determine the size of our combined dataset. The data availability for the sets is visualized in Figure 2.4. The reanalysis data are available every day during the full period, while

the AIS data are missing some days. The vessel data are available from the end of May 2014 to December 2016 with several periods of missing data, most notably January and February 2016. Figure 2.5 shows the amount of vessel data stored for each day.

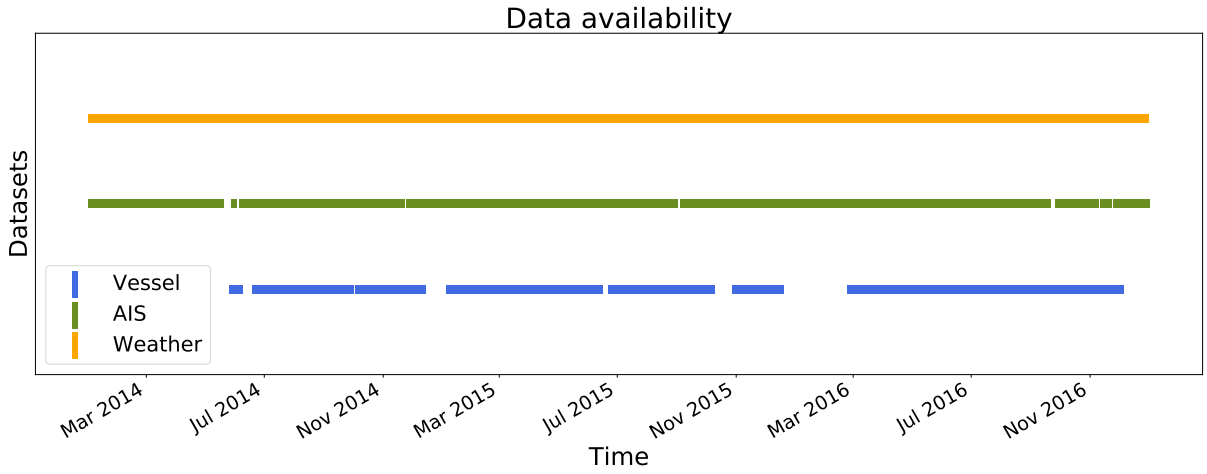

Figure 2.4: Data availability for the different datasets over the full period.

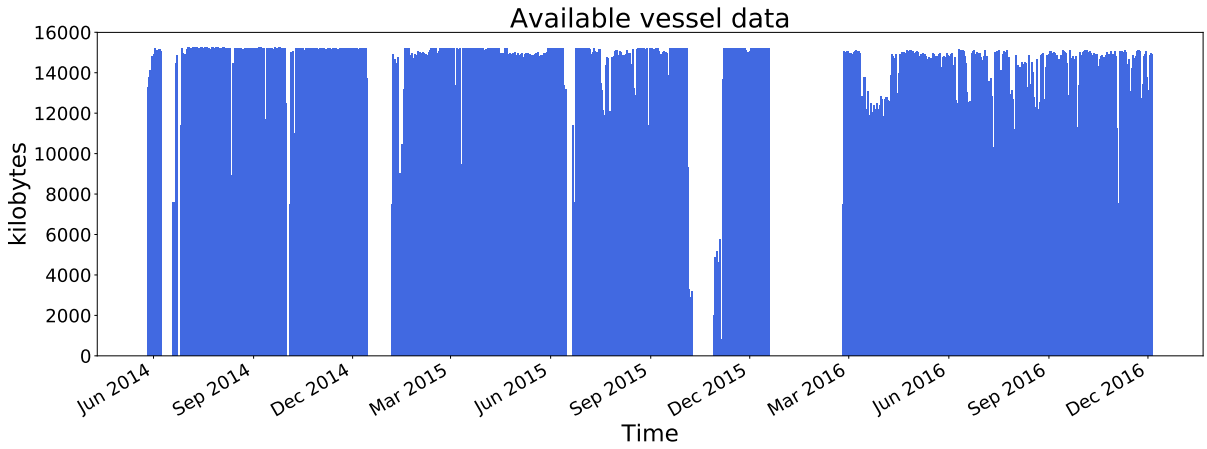

Figure 2.5: Available vessel data in kilobytes for each day.

2.7 Software Platforms

Processing and analyzing large amounts of data using statistical methods can be performed using a variety of programming languages, Interactive Development Environments (IDEs) and software suites. Some popular languages include R, Python, C/C++, Java and Matlab and a commonly used software suite is Statistical Analysis System (SAS). Especially Python and R are popular programming languages for data science, with large communities providing open-source packages and helpful discussion forums online. In this work, we will utilize

Python as the main platform for implementation and data analysis. This is due to several factors:

- Python is a free, interpreted, open-source platform with a large community[17].
- Python can be augmented by a huge variety of free, open-source libraries and packages such as:
    - **Pandas**, an open-source Python library with powerful tools for handling and manipulating large amounts of data in an efficient manner[18].
    - **NumPy**, a fundamental package for scientific computing with Python. It implements n-dimensional array objects, linear algebra, the Fourier transform and random number capabilities among other things[19].
    - **SymPy**, a Python library for symbolic mathematics, including features such as symbolic integration and differentiation[20].
    - **SciPy**, a collection of numerical algorithms and domain-specific toolboxes, including signal processing, optimization, statistics and much more[21].
    - **SciKit-Learn**, an open-source Python library for machine learning and data mining[22].
    - **Matplotlib**, an extensive plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms[23].

When working with Python we will utilize Spyder, a free, open-source IDE for scientific programming in the Python language[24].

There are of course several disadvantages with Python compared to other languages and tools. Since Python is a high-level interpreted language it is much slower than compiled languages like C and C++, but Python programs are in general shorter and more compact[25]. R was specifically developed for statistical use and has a richer set of libraries and packages for data science and more novel visualization possibilities than Python[26]. However, some commands in R display poor memory management, while the language is known for its steep learning curve and at times impenetrable documentation[27].

Chapter 3

Statistical Learning Methods

This chapter presents the theory and interpretation of several methods from the broad field of statistical learning. Statistical learning is heavily rooted in theoretical statistics and only a small subset of important methods are selected, with a focus on interpretation and practical use. Section 3.1 will give a brief introduction, while Section 3.2 covers the most important tasks in data preprocessing. Learning in the supervised framework is discussed in Section 3.3, with main a focus on regression analysis and model assessment and validation. Learning in the unsupervised framework is discussed in Section 3.4, covering principal component analysis.

Some parts of this chapter are directly taken from my previous project thesis carried out the fall of 2016 written about multivariate analysis of ship data[28]. Most of the mathematical derivations and explanations are based on *The Elements of Statistical Learning* by Hastie et al.[29]. References are made in the text where other sources have been used.

3.1 Introduction

Throughout the thesis, we will let uppercase letters, such as X and Y refer to generic aspects of variables. Observed values are written in lowercase, such that the ith observation of X is denoted x_i, which can either be a scalar or a vector. Bold uppercase letters represent matrices, such that a set of N samples of p variables is denoted $\mathbf{X} \in \mathbb{R}^{N \times p}$. In general, vectors will not be bold, except when they have N components. This makes a clear distinction between the p-vector of variables x_i for the ith observation and the N-vector of observations $\mathbf{x}_j$ for variable X_j.

Thus, we have a set of N samples, also called observations or objects,

$$\mathbf{X} = [x_1, x_2, \cdots, x_N]^T, \tag{3.1}$$

where each observation x_i is a vector of p variables, also called features or attributes,

$$x_i = [x_{i1}, x_{i2}, \cdots, x_{ip}]^T. \tag{3.2}$$

Alternatively, we can say that the set consists of p variables, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p]$, where each variable is a column vector of N samples, $\mathbf{x}_j = [x_{j1}, x_{j2}, \cdots, x_{jN}]^T$. We want to be able to extract information from the dataset $\mathbf{X} \in \mathbb{R}^{N \times p}$, either by inspecting the structure of $\mathbf{X}$ alone or by exploring its relationship to some other dataset $\mathbf{Y} \in \mathbb{R}^{N \times q}$.

When performing exploratory data analysis on a large set of complex, real-world data, it is common to first explore the data in the *unsupervised* framework. This can reveal structures and patterns of the data itself, often providing some important preliminary insight into the problem at hand. Then a *supervised* approach can be taken to construct and train models for prediction and classification. The distinction between the two frameworks will be explained later.

3.2 Preprocessing

Preprocessing is a necessary step in data mining that involves transforming the raw, real-world data into an understandable, consistent format. Most statistical learning methods require some sort of preprocessing to effectively learn from the data. In general, the real world data is often incomplete, noisy and inconsistent and likely to contain many errors. Low-quality data will lead to low-quality results[30].

For our purpose, we separate the preprocessing steps into data cleaning (Section 3.2.1) and data transformation (Section 3.2.2). Note that these categories are not mutually exclusive and that some methods may be seen as a form of data cleaning, as well as data transformation.

Before covering the preprocessing steps we introduce some summary statistics to describe a data matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$. The central tendency of a variable $\mathbf{x}_j$ can be measured by its arithmetic

mean over the N observations,

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \quad j = 1, \ldots, p. \tag{3.3}$$

Similarly the variance of a variable $\mathbf{x}_j$ over N observations can be written as

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2, \quad j = 1, \ldots, p, \tag{3.4}$$

where the standard deviation σ_j of the variable is the square root of the variance.

3.2.1 Data Cleaning

Data gathered by sensor systems are likely to contain noise, outliers, erroneous data and periods of missing values. Data cleaning methods try to fill in missing values, reduce noise, remove outliers and correct inconsistencies in the data.

3.2.1.1 Missing Values

Missing values in time series data is a common problem; values may not be measured, values may be measured but get lost or values may be measured but are considered unusable. Samples with missing values can either be removed from the dataset, or the missing values can be imputed using simple univariate techniques[31]:

- **Constant insertion**: The missing values are replaced by some predefined constant.
- **Mean, mode or median insertion**: The missing values are replaced by the variable mean, mode or median.
- **Interpolation**: The missing values are replaced by interpolation. Either linear or higher order interpolation can be used.

With few unstructured missing values, one can simply remove samples or impute the values by one of the techniques above. However, if the missing values structured in time over a significant period it can hardly be justified to impute the data using any of these methods. In such cases, the time period with missing data could be removed entirely, or imputed by more advanced methods taking into account inter-variable correlations, such as imputation by regression[32].

3.2.1.2 Noise

Noise is a random error or variance in the data, often assumed to be Gaussian by nature. Given a time series of a variable one often wants to smooth out the time series by removing noise. For this purpose we will mention two methods:

Moving Average

A moving average is commonly used with time series data to smooth out high-frequency variations and highlight lower frequency trends and cycles in the data. It is similar to a low-pass filter. In a central moving average, a window of odd length k is centered around a data point, and the value is replaced by the unweighted average of the points within the window. This is repeated for every data point. By using a central window, instead of only a backward looking window, we do not introduce a phase lag in the time series.

If we let $n = (k-1)/2$ and $y(t)$ be the time series value at time t, then the central moving average $\mathrm{CMA}(t)$ can be defined as

$$\mathrm{CMA}(t) = \frac{1}{2n+1} \sum_{i=-n}^{n} y(t+i). \tag{3.5}$$

At the tails of the time series, where values are not defined for the full window, $\mathrm{CMA}(t)$ is either not computed, or it is computed using the available values within the window. Figure 3.1 illustrates the central moving average on a sine curve with $N = 300$ samples and noise, for $k = 5, 15, 51$, and 99 respectively.

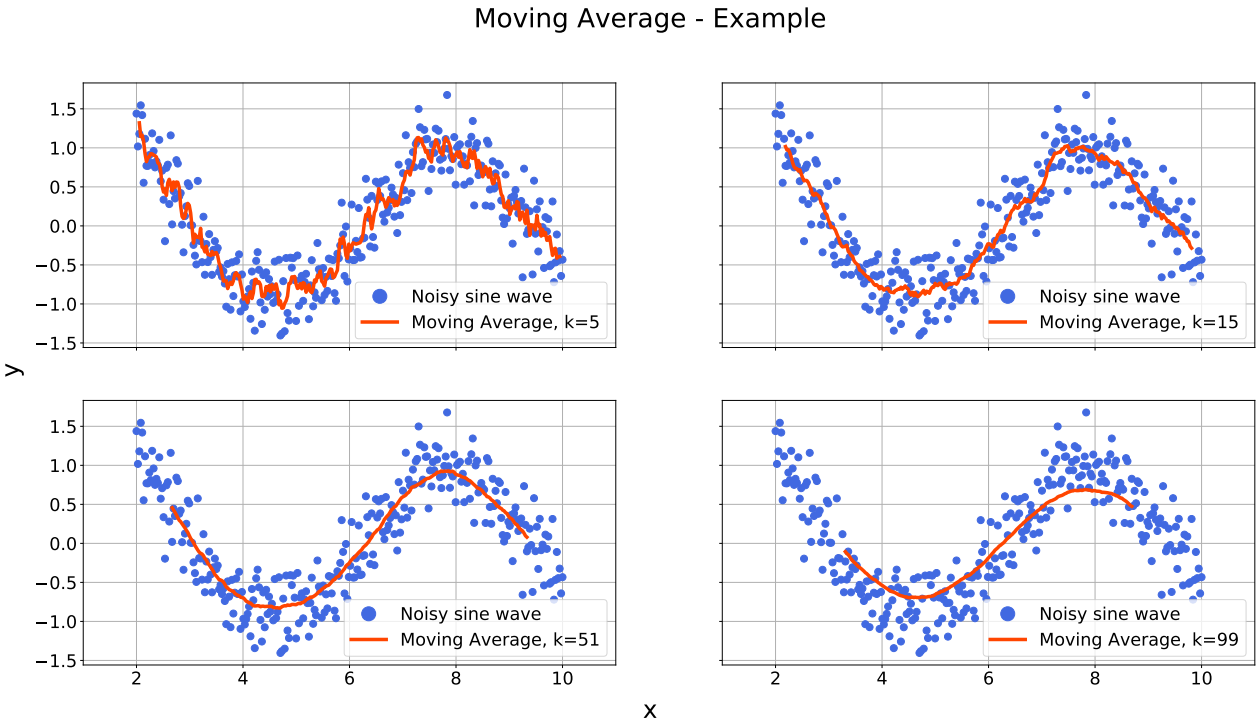

Figure 3.1: Example of moving average smoothing for different values of k.

LOWESS

Locally weighted scatterplot smoothing (LOWESS) is a method of local regression that utilizes least squares fitting for local segments of the data[29]. In general, LOWESS is a non-parametric method that for each point locally fits a polynomial of order λ to a fraction α of the N data points. The data points used for the local fit are weighted, such that the points closest to the point of estimation are given the most weight. α is known as the smoothing factor and has a typical range of $0.25 - 0.5$. Usually, it is sufficient to use a local linear fit, *i.e.* $\lambda = 1$. Using $\lambda = 0$ would turn the method into a weighted moving average. Figure 3.2 shows LOWESS smoothing on a sine curve with $N = 300$ samples and noise, using $\lambda = 1$ and $\alpha = 0.02, 0.25, 0.5$, and 0.9 respectively.

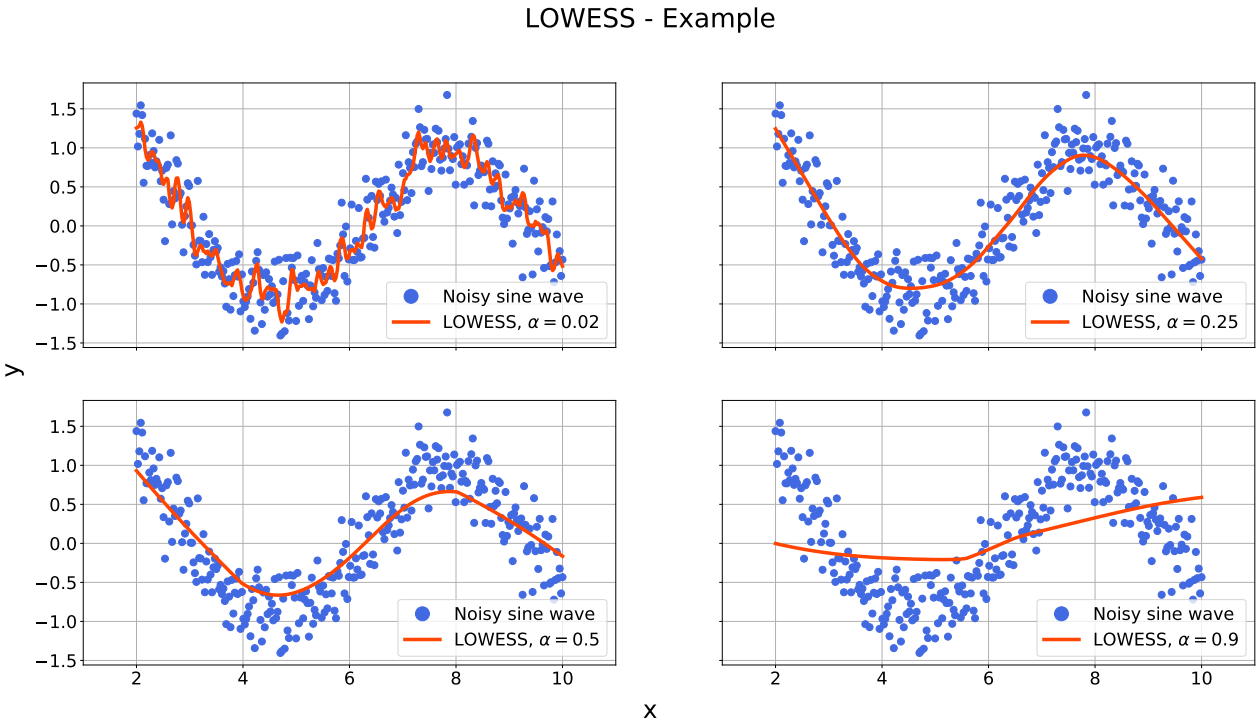

Figure 3.2: Example of LOWESS smoothing for $\lambda = 1$ and different values of α.

3.2.1.3 Outliers

Often the data are contaminated by outliers, *i.e.* observations or clusters of observations that are distant from other observations. Outliers can occur by chance in any distribution, but may also indicate erroneous data due to experimental errors or faulty values inserted during data handling and storage. In some cases, one may be interested in analyzing the outliers specifically, but often we seek to remove such observations.

Outlier detection is no straightforward task and various methods exist. In this thesis, outlier detection will not be considered in any detail and we propose two simple methods of detection:

- **Range check**: A variable is constrained to an allowable range where all observations outside the range are marked as outliers. This is ideal when working with sensory data where sometimes non-physical values appear, such as negative cargo levels.
- **Nearest-neighbors**: The distances from an observation x_i to its k-nearest neighbors are computed and used to determine if the observation is marked as an outlier[33].

The range check is applicable for one variable at a time, while the nearest-neighbors algorithm can handle multidimensional observations. Box plots or cluster analyses are also com-

monly used for outlier detection[30]. For a more advanced approach, taking into account the local densities of high-dimensional observations, the local outlier factor is proposed[34].

3.2.2 Data Transformation

The range of different variables often varies considerably, and many statistical learning methods assume mean centered or scaled input variables. Rescaling and standardization can ensure that variables measured in different units and magnitudes have equal importance in the transformed dataset. For linear regression models, it makes the regression coefficients directly comparable. We let $\mathbf{x}'_j$ denote the transformed variable $\mathbf{x}_j$.

3.2.2.1 Rescaling

Rescaling simply refers to redefining the scale of a variable, typically within $[0,1]$ or $[-1,1]$. The general formula is given as

$$\mathbf{x}'_j = (b-a)\frac{\mathbf{x}_j - \min(\mathbf{x}_j)}{\max(\mathbf{x}_j) - \min(\mathbf{x}_j)} + a, \tag{3.6}$$

where the desired new range is $[a,b]$.

3.2.2.2 Standardization

While rescaling can be useful in many situations it does not mean center the data. An effective method for both mean centering and scaling the data is standardization. The variables are transformed into z-scores, ensuring zero mean and unit variance. Standardization is done by

$$\mathbf{x}'_j = \frac{\mathbf{x}_j - \bar{x}_j}{\sigma_j}. \tag{3.7}$$

3.2.2.3 Basis Expansions

As will be explained in Section 3.3.1, the linear regression models assume a linear relationship between the inputs X and outputs Y. Often this is not the case, and through a basis expansion of X one can express more complicated regression relationships. The idea is to

augment the vector of inputs X with additional variables, which are transformations of X[29]. Some common transformations are:

(a) Polynomial transformations such as X^2, $X_j X_k$ *etc.*

(b) Nonlinear transformations such as $\log(X)$ and $\sqrt{X}$.

One can then create a more flexible regression model which is linear in the new basis expansion of X.

3.3 Supervised Learning

In supervised learning, we are concerned with finding a model that best describes the relationship between some set of independent variables $\mathbf{X} \in \mathbb{R}^{N \times p}$ and some set of dependent variables $\mathbf{Y} \in \mathbb{R}^{N \times q}$. The dependent variables represent outcomes whose variations are being studied, and the independent variables represent causes or potential reasons for variation. This leads to models used for prediction or classification. To construct such a model one has to use a set of training data $\mathcal{T}$, where each sample is said to be a pair consisting of an input observation x_i and a desired output value y_i. This is often called labeled data. The model is then trained to fit the data and can be used to perform classification or prediction on new unlabeled samples that were not used to train the model. In this thesis, we will explore models that predict quantitative outcome labels y_i, namely regression models.

3.3.1 Linear Methods for Regression

A linear regression model assumes that the regression function $\mathrm{E}[Y|X]$ is linear in the inputs $X_1, \ldots, X_p$. Linear models are simple and it is often easy to interpret how the inputs affect the outputs. For prediction, they can sometimes outperform fancier nonlinear models, especially in situations with few training samples or low signal-to-noise ratio[29].

3.3.1.1 Ordinary Least Squares Regression

Given the input vector $X^T = [X_1, X_2, \ldots, X_n]$, we want to predict the real-valued output $Y \in \mathbb{R}$. The linear regression model can be written as

$$f(X) = \beta_0 + \sum_{j=1}^{n} X_j \beta_j, \tag{3.8}$$

where β_0 is the intercept of the model, while the β_j's are the unknown parameters of coefficients to be determined. This model assumes that the regression function $\mathrm{E}[Y|X]$ is linear, or that a linear model is a reasonable approximation. The input variables X_j can be quantitative inputs, or basis expansions of quantitative inputs, as explained above. No matter the source of X, the model is still linear in its parameters β_j.

With a set of training data $(x_1, y_1) \dots (x_N, y_N)$ we can estimate the coefficients $\beta = [\beta_0, \beta_1, \dots, \beta_p]^T$ using the least squares method. That is, choose the coefficients β such that the residual sum of squares (RSS),

$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \end{aligned} \tag{3.9}$$

is minimized. (3.9) makes no assumptions about the validity of model (3.8), but simply finds the best linear fit to the data by measuring the average lack of fit. If we let $\mathbf{X} \in \mathbb{R}^{N \times (p+1)}$ be a matrix where each row is an input vector with a 1 in first position, and similarly let $\mathbf{y} \in \mathbb{R}^N$ be the vector of outputs in the training set, then the RSS can be rewritten as

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta). \tag{3.10}$$

If we assume that $\mathbf{X}$ has full column rank, we can differentiate (3.10) and set it equal to zero to obtain the unique solution,

$$\begin{aligned} \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) &= 0 \\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned} \tag{3.11}$$

Thus the predicted values at the training inputs are given as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{3.12}$$

where $\hat{y}_i = \hat{f}(x_i)$. In the case where $\mathbf{X}$ not has full column rank, *i.e.* the columns of $\mathbf{X}$ are not linearly independent, then $\mathbf{X}^T\mathbf{X}$ is singular and the least squares coefficients $\hat{\beta}$ are not uniquely defined. This occurs if two of the inputs are perfectly correlated.

Let us now assume that the observations y_i are uncorrelated with a constant variance σ^2, and that the x_i are nonrandom. The covariance matrix of the least squares parameter estimates can be derived from (3.11) as

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2. \tag{3.13}$$

An unbiased estimate of the variance σ^2, *i.e.* $\text{E}(\hat{\sigma}^2) = \sigma^2$, is typically given by

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{3.14}$$

Furthermore, by assuming that (3.8) is the correct model for the mean, and that the deviations of Y around its expected value are additive and Gaussian, we can write

$$f(X) = \beta_0 + \sum_{j=1}^{n} X_j\beta_j + \epsilon, \tag{3.15}$$

where $\epsilon \sim N(0,\sigma^2)$. From (3.15) we get that $\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$, a multivariate normal distribution. Under these assumptions we can use the distributional properties to form hypothesis tests and confidence intervals for β_j. By using the Z-score

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}, \tag{3.16}$$

where v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$, we can test the hypothesis that a particular coefficient $\beta_j = 0$. For a sufficiently large sample size, $N \geq 100$ we can use normal quantiles for z_j. By forming a null hypothesis and an alternative hypothesis

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0, \end{aligned} \tag{3.17}$$

we can use z_j to reject H_0 if the corresponding p-value is less than a given threshold α (typically 0.05 or 0.01). That is, if the probability of attaining a particular β_j is very low under the assumption of H_0, say 5 %, we reject H_0 and assume H_1. We then say that the result is

significant. Note that this also corresponds to a false rejection of the null hypothesis in 5 % of the cases.

Furthermore, by assuming α to be the threshold at each tail of the distribution we can obtain a $1-2\alpha$ confidence interval for β_j,

$$[\hat{\beta}_j - z^{(1-\alpha)}\sqrt{v_j}\hat{\sigma},\ \hat{\beta}_j + z^{(1-\alpha)}\sqrt{v_j}\hat{\sigma}], \tag{3.18}$$

where $z^{(1-\alpha)} = 1.96$ amounts to a 95 % confidence interval under the assumption of additive and Gaussian residuals.

3.3.2 Nonlinear Methods for Regression

In the previous section, we relied on the assumption that the regression function $\mathrm{E}[Y|X]$ was linear. However, in regression problems, it is highly unlikely that the true function $f(X) = \mathrm{E}[Y|X]$ is linear in X. It will typically be nonlinear and nonadditive in X. Representing $f(X)$ by a linear model is usually a convenient approximation, allowing for easier interpretation of the model. Extending our model beyond linearity can give better predictions, but at the cost of interpretability[29]. Some nonlinear models, such as neural nets can also be difficult to tune correctly due to overparameterization[29]. We will focus on a simple nearest-neighbor method for regression, a highly unstructured and model-free method.

3.3.2.1 Nearest-Neighbors Regression

In nearest-neighbors regression, we use those observations in the training set $\mathcal{T}$ that are closest to an unseen observation x in input space to estimate an output y. A nearest-neighbors model will not give you an explicit model but relies on the training set for making predictions of unseen data. Using the k-nearest neighbor fit, $\hat{Y}$ can be defined as

$$\hat{Y}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i, \tag{3.19}$$

where $N_k(x)$ denotes the neighborhood of x of size k. As the method relies on some measure of distance between pairs of observations we need to define a metric. Table 3.1 lists some frequently used distance metrics, where $\mathbf{S}$ denote the covariance matrix between observations x_i and x_k.

Names	Function
Euclidean distance	$\|x_i - x_k\|_2 = \sqrt{\sum_j (x_{ij} - x_{kj})^2}$
Manhattan distance	$\|x_i - x_k\|_1 = \sum_j \|x_{ij} - x_{kj}\|$
Maximum distance	$\|x_i - x_k\|_\infty = \max_j \|x_{ij} - x_{kj}\|$
Mahalanobis distance	$\sqrt{(x_i - x_k)^T \mathbf{S}^-1 (x_i - x_k)}$

Table 3.1: Frequently used distance metrics.

While the model appears to have a single parameter k, the effective number of parameters can be estimated as N/k and is generally bigger than the p parameters in least-squares models[29]. Thus, model complexity decreases with k. With the choice of k and a given metric, (3.19) is simplistic and easy to understand, but not useful for understanding the nature of the relationship between X and Y. Nearest-neighbors regression works reasonably well for low-dimensional features, while it should be avoided for high-dimensional features due to the bias-variance tradeoff[35] (see Section 3.3.3.1).

To choose an optimal k one can utilize cross-validation to find the k that minimizes the expected test error. This is an efficient and reliable method for selecting k. Cross-validation is defined in Section 3.3.3.2. Generally, a small k will yield low bias and high variance predictions, while a large k will do the opposite, as one is averaging over larger neighborhoods.

3.3.3 Model Assessment and Selection

The performance of a model relates to its prediction capability on independent test data. In practice, it is important to assess this performance, as it guides the choice of model and gives a measure of the quality of the chosen model. To better understand the problem of model selection we investigate the tradeoff between bias and variance as explained in Chapter 7 in Hastie et al.[29].

3.3.3.1 Bias-Variance Tradeoff

Let us consider an output variable Y, a vector of input variables X, and a prediction model $\hat{f}(X)$ estimated from a training set $\mathcal{T}$. To measure the error between Y and $\hat{f}(X)$ we introduce the loss function denoted by $L(Y, \hat{f}(X))$. A typical choice is the squared error

$$L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2. \tag{3.20}$$

The performance of the model can be assessed by the test error, which is the prediction error over an independent test set where both X and Y are drawn randomly from their population. The test error is conditional on the specific training set $\mathcal{T}$ that was used to estimate the model and can be written as

$$\text{Err}_{\mathcal{T}} = \text{E}[L(Y, \hat{f}(X))|\mathcal{T}]. \tag{3.21}$$

The test error is also called prediction error or generalization error, as it describes how well the model generalizes to new, unseen data. While we want to estimate $\text{Err}_{\mathcal{T}}$, in reality, most methods effectively estimate the expected test error $\text{E}[\text{Err}_{\mathcal{T}}]$, which we will call Err. Additionally the training error, a measure of how well our model fits the training data, is the average loss over the training set

$$\overline{\text{err}} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i)). \tag{3.22}$$

Let us now define error due to bias and error due to variance[36]:

- **Error due to bias** is the difference between the expected value of our prediction and the true target value. Bias in general measures how far off our prediction is from the correct value. A high bias can cause a model to miss relevant relationships between inputs and target outputs, which is called underfitting.
- **Error due to variance** is the expected squared deviation of a prediction around its mean, *i.e.* the variability of a model prediction. A high variance can lead to the modeling of random fluctuations in the training data, which is called overfitting.

Typically, as the model complexity of $\hat{f}$ increases, the bias decreases while the variance increases. Similarly, for a less complex model, the variance typically decreases while the bias increases. If we assume $Y = f(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, we can mathematically decompose the expected test error at an input point $X = x_0$ as

$$\begin{aligned}\text{Err}(x_0) &= \text{E}[(Y - \hat{f}(x_0))^2] \\ &= \sigma_\epsilon^2 + [\text{E}[\hat{f}(x_0)] - f(x_0)]^2 + \text{E}[\hat{f}(x_0) - \text{E}[\hat{f}(x_0)]]^2 \\ &= \sigma_\epsilon^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) \\ &= \text{Irreducible Error} \ + \ \text{Bias}^2 + \ \text{Variance},\end{aligned} \tag{3.23}$$

where the squared error loss function has been used. The first term is the variance of the target around its true mean and cannot be avoided. The second and third terms are due to the previously explained bias and variance. Thus when selecting a model $\hat{f}_\alpha$, where α denotes the model complexity, there exists an optimal α that minimizes the expected test error Err. This is the problem of the bias-variance tradeoff.

Throughout the thesis, we use cross-validation to estimate Err and perform model selection, while an independent test set is used to compute $\text{Err}_\mathcal{T}$ to assess the final chosen model. $\text{Err}_\mathcal{T}$ is then compared to the *base error rate, i.e.* the error over the independent test set when the mean target value over the training set is used for prediction. This allows us to assess the relative improvement of a statistical learning method over some baseline error.

3.3.3.2 Cross-Validation

In a data-rich situation, one would typically split the data intro three parts: a training set, a validation set, and a test set. The training set is used to fit the models, the validation set is used to estimate the prediction error for model selection, and the test set is used to assess the test error over an independent test set for the final chosen model[29].

However, data is often sparse and instead of splitting the data into three parts one can utilize cross-validation. Cross-validation is a simple and widely used method to estimate the expected prediction error Err. We consider K-fold cross-validation. The idea is simple: instead of setting aside an independent validation set to assess the performance of our model, we pseudo-randomly split the data in K roughly equal-sized parts and use each of the parts as a validation set once. This implies that for each part k we fit a model on the remaining $K-1$ parts and validate it using part k. This process is repeated K times until all the data has been used for training and all the data has been used for validation. This will give us K estimates of the prediction error, which allows us to compute the mean value and standard error of the

estimates. In other words, we obtain an estimate of Err, $\widehat{\text{Err}}$. The difference between splitting the data into three parts, and splitting the data into two parts while using cross-validation on the training set is illustrated in Figure 3.3

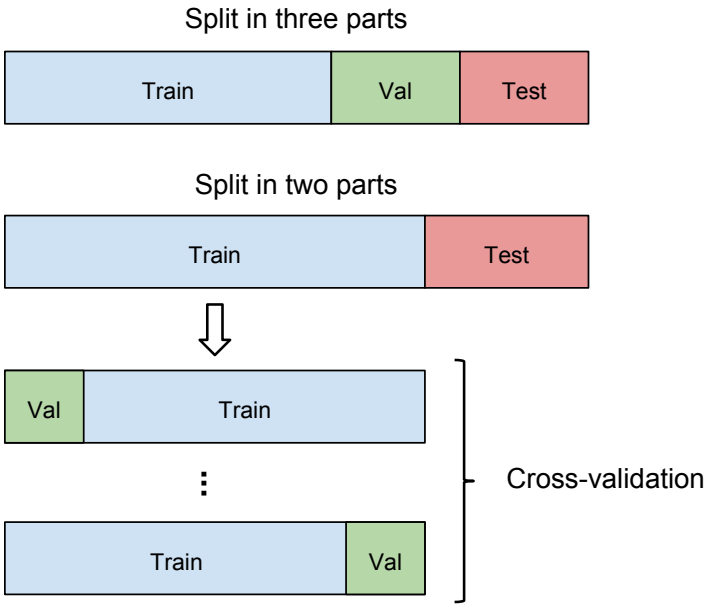

Figure 3.3: Cross-validation illustration.

Cross-validation is often used to perform model selection. It can be used to determine an optimal tuning parameter α, such as k in nearest-neighbors methods. Or it can be used to determine the optimal subset of input variables to be used in the model. The latter will be explained in Section 3.3.3.3. In both cases the selection is based on the minimization of cross-validation error, *i.e.* the model that results in the lowest $\widehat{\text{Err}}$. Often the one-standard-error rule, advocated by Breiman [37] and Hastie et al. [29], is used to choose the most *parsimonious* model. That is, one chooses the simplest model in terms of model complexity within one standard error of the best model. The standard error can be computed as

$$\text{SE}(\alpha) = \frac{\sigma_\alpha}{\sqrt{K}}, \tag{3.24}$$

where σ_α is the standard deviation of the cross-validation errors over all K folds for a given model complexity α.

The choice of K is not a straightforward problem. With $K = N$, also known as leave-one-out (LOO) cross-validation, the estimation of Err is approximately unbiased but can have high variance. Additionally, the computational costs increase as K increase. With a lower value

for K, *e.g.* $K = 2$, cross-validation has lower variance, but bias can be a problem depending on how the performance of the model varies with training size. Both $K = 5$ and $K = 10$ are common practices in the community and is recommended by Kohavi[38] and Breiman and Spector[39].

3.3.3.3 Variable Selection

In learning tasks, we are often faced with the problem of variable selection. Especially in linear regression one may want a simple model with only the most important variables for ease of interpretation. Again we are facing the bias-variance tradeoff: we seek a good model in terms of generalization and goodness of fit while penalizing model complexity to avoid overfitting.

In reality, variable selection is no different from model selection based on parameter tuning. However, when the number of input variables p gets sufficiently large, the number of possible subsets $2^p - 1$ gets so large that it is computational exhaustive to check them all. Rather than searching through all possible subsets of p, which become unfeasible for p much larger than 40, we can use a greedy algorithm to find a good path through them. This will give us a set of models indexed by the subset size k. Typically we choose the subset size k that minimizes $\widehat{\text{Err}}$.

Forward-stepwise selection is a greedy algorithm that starts with no input variables, only the intercept, and adds to the model the variable that best improves the fit. In other words, the predictor that reduces the residual sum of squares the most will be added. This leads to a sub-optimal, constrained search through all possible subsets, but can be preferred due to computational considerations, since only $1/2p(p+1)$ models are evaluated. As the residual sum of squares will decrease for each new variable added, some criterion to choose the optimal subset size k has to be used. Hastie et al.[29] criticizes the traditional use of the F-statistics to select or remove variables based on their significance, and proposes to rather use cross-validation as a way to choose the best subset size k. Similarly, statistician William Briggs in his blog post[40] and Flom and Cassell[41] address the known problems with stepwise regression, both discouraging the use of level of significance to add or drop variables, while stressing the importance of using cross-validation and independent test sets. A more novel approach for best subset selection is the branch and bound algorithm proposed by Narendra

and Fukunaga [42], which guarantees the optimal solution without the need for an exhaustive search.

Example

Cross-validation, variable selection and the interplay between model complexity, bias and variance is best illustrated by an example. Let our true model be $f(X_1) = \sin(X_1) + \epsilon$ with $\epsilon \sim N(0, 0.25)$ and 100 equidistant points on $x \in [-\pi, \pi]$. The observations are seen in Figure 3.4. Let us now assume a linear regression model as in (3.8), where the input variables X_j are polynomial transformations of X_1. If we let α denote the highest order transformation we include in our model, it will correspond to our model complexity. We want to find an optimal model $\hat{f}_\alpha$ minimizing $\widehat{\text{Err}}$.

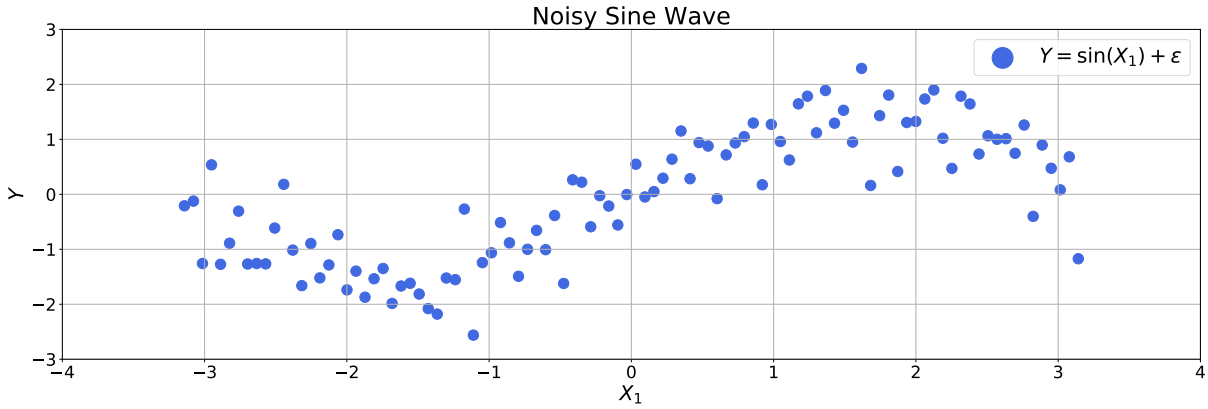

Figure 3.4: Sine wave with random noise, 100 observations.

To estimate the expected test error we use 5-fold cross-validation for each $\hat{f}_\alpha$. We let α vary from 0 (constant model) to 18 (polynomial model of order 18), and use the squared error loss function. Figure 3.5 shows $\widehat{\text{Err}}$ and $\text{E}[\overline{\text{err}}]$ and their standard errors as a function of α. For $\alpha = 0$, a constant model with one parameter β_0, we have no variability in the model predictions and the expected values of our predictions are far off from the target values. In other words we have low variance and high bias, resulting in underfitting and a high $\widehat{\text{Err}}$. On the opposite side, for $\alpha = 18$, a high order polynomial model with 19 parameters, we have high variability in our model predictions around their mean, but the expected value of a prediction is much closer to its target value. Thus we have low bias and high variance, resulting in overfitting and a high $\widehat{\text{Err}}$. We note that $\text{E}[\overline{\text{err}}]$ steadily decreases as a function of α, as expected. For some $0 \leq \alpha \leq 18$ there exists an optimal tradeoff between the bias and

variance, resulting in a minimum $\widehat{\text{Err}}$, found to be $\alpha = 3$, a polynomial model of 3rd degree.

In Figure 3.6 we have plotted the regression models fitted to all the data for $\alpha = 0$, $\alpha = 3$, and $\alpha = 18$ respectively to illustrate the differences between underfitting, overfitting and the choice of best model. We clearly see that for $\alpha = 0$ the model fails to capture important relationships between X_1 and Y, while for $\alpha = 18$ the model has started to fit the variability in the target values caused by noise. The model with $\alpha = 3$ represents the optimal tradeoff.

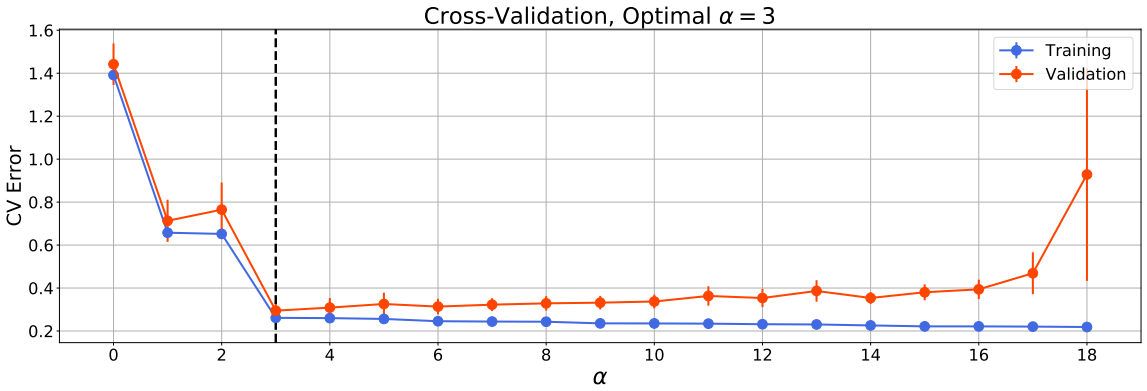

Figure 3.5: Cross-validation curve as a function of α, showing the bias-variance tradeoff.

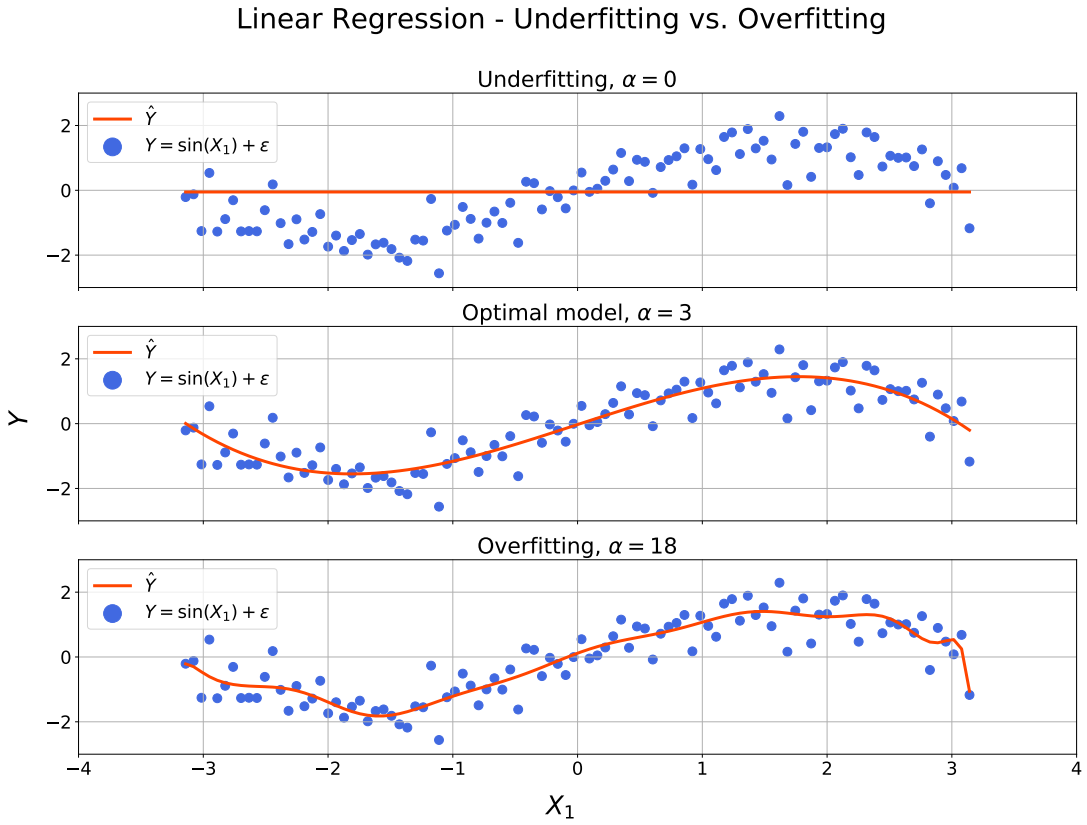

Figure 3.6: Three regression models for $\alpha = 0$, $\alpha = 3$ and $\alpha = 18$ respectively.

3.4 Unsupervised Learning

Unsupervised learning is concerned with revealing internal structures in unlabeled data **X**. In the unsupervised framework, we observe only the data itself with no relation to a measured outcome. Without an outcome variable to guide the learning process we call it unsupervised learning. Our task is rather to describe how the data are organized or clustered. In the literature, the unsupervised problem is less developed than the supervised one. With supervised learning, there is a clear measure of success and you can more easily compare the effectiveness of different methods. With unsupervised learning, there is no such direct measure of success[29].

3.4.1 Principal Component Analysis

A powerful unsupervised method is principal component analysis (PCA), which goes as far back as 1901, invented by Karl Pearson[43], and later independently developed and named by Harold Hotelling in the 1930s[44]. In situations where we have a set of p possibly correlated variables X_j, we can transform the variables into a set of linearly uncorrelated variables Z_m for $m = 1, \ldots, M$ and $M \leq p$. Each Z_m is a linear combination of the variables X_j and is called a *principal component* (PC). The PCs are a sequence of projections of the data, mutually uncorrelated and ordered in variance. In other words, the transformation is defined such that each PC accounts for as much of the variability in the data as possible, under the constraint that it is orthogonal to the preceding PCs.

If we consider a mean centered data matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, the best linear approximation of rank M of the data can be written as

$$\mathbf{X} = \mathbf{X}\mathbf{H}_M + \mathbf{E}, \tag{3.25}$$

where $\mathbf{H}_M \in \mathbb{R}^{p \times p}$ is a projection matrix mapping each observation x_i onto its rank-M reconstruction $\mathbf{H}_M x_i$. The residual matrix $\mathbf{E}$ is the difference between the data and its rank-M reconstruction. The model can be fitted to the data by minimizing the reconstruction error. Several methods exist for extracting the PCs, where eigenvalue decomposition of the covariance matrix of $\mathbf{X}$ is common. We follow Hastie et al.[29] and use the singular value decomposition (SVD), a standard decomposition used in numerical analysis[45], to define the

PCs. Using SVD

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{3.26}$$

it can be shown that $\mathbf{H}_M = \mathbf{V}_M\mathbf{V}_M^T$ where $\mathbf{V}_M \in \mathbb{R}^{p\times M}$ consists of the first M columns of $\mathbf{V}$[29]. Here $\mathbf{U} \in \mathbb{R}^{N\times p}$ is an orthogonal matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$) whose columns $\mathbf{u}_j$ are called left singular vectors, $\mathbf{V} \in \mathbb{R}^{p\times p}$ is an orthogonal matrix whose columns v_j are called right singular vectors, and $\mathbf{D} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with diagonal elements $d_1 \geq \cdots \geq d_p$ called singular values. The columns of $\mathbf{UD}$ are the principal components of $\mathbf{X}$, that is the projections of the data $\mathbf{X}$ onto the principal directions. This is also known as the PC scores.

As an alternative to (3.25) we can write

$$\mathbf{X} = \mathbf{Z}\mathbf{V}_M^T + \mathbf{E}, \tag{3.27}$$

where $\mathbf{Z} \in \mathbb{R}^{N\times M}$ is the scores matrix and $\mathbf{V}_M^T \in \mathbb{R}^{M\times p}$ is the loading matrix. Here each row of $\mathbf{Z}$, z_i is the observation x_i projected onto the M-dimensional space spanned by the loading vectors v_j. Thus $\mathbf{Z}$ consists of the M linearly uncorrelated principal components Z_m. As we increase M the complexity of our model increase, and for $M = p$ we get back our original data exactly and $\mathbf{E} = 0$.

By projecting the data onto a lower-dimensional subspace, defined by an uncorrelated orthogonal basis set, one can more easily inspect and visualize the latent structures and correlations in the data. For example, by projecting the data onto the first two PCs one might discover natural groupings, or clusters within the data, that otherwise would go unnoticed in the original high-dimensional data. Furthermore, the loading vectors which are of unit length describe the relative importance of each original variable in the given PC direction, with a high absolute value corresponding to high relative importance. To give all input variables X_j equal chance to affect the result, they should be standardized before the PCs are extracted.

PCA assumes that the data can be represented by a linear model, which might be a crude estimation. However, it is easy to interpret and might still provide valuable information. Alternatively, several areas of research have explored how applying a nonlinearity prior to performing PCA could extend the method, which is known as kernel-PCA[46]. Furthermore,

PCA assumes that the mean and variance are sufficient statistics to describe the variables, *i.e.* that they entirely describe the probability distribution of a variable. The only zero-mean probability distribution being described by its variance is the Gaussian distribution. In order for this assumption to hold, the probability distributions of the variables must be Gaussian[47]. At last, since PCA attributes importance to the variability in the data, it assumes that components accounting for most of the variation in the data are the most important components. Thus the first principal components would account for important dynamics in the data, while latter components would be attributed more and more to noise.

Example

To illustrate PCA with an example, we used the Iris dataset, a multivariate dataset introduced by statistician and biologist Ronald Fisher[48]. The dataset consists of 50 samples from each of the three species of Iris, *Iris setosa, Iris virginica* and *Iris versicolor*. For each sample, four features, the length and width of the sepals and petals in centimeters are measured. To get an idea of how the four variables vary with the three species of Iris, one could plot one variable against another and try to find interspecies differences. Instead, we will use PCA to project the 4-dimensional data onto an M-dimensional subset, where $M \leq 4$. The data were standardized to have zero mean and unit variance before extracting the principal components. Figure 3.7 shows the cumulative explained variance as a function of M. By projecting the data onto a two-dimensional subspace spanned by the principal directions, as much as 95.8 % of the variability in the data was still contained.

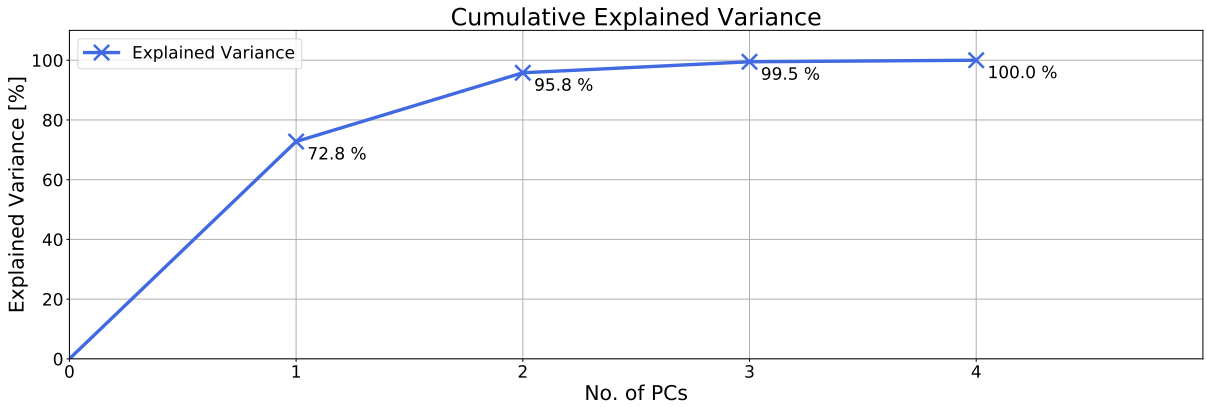

Figure 3.7: Cumulative explained variance as a function of no. principal components.

Figure 3.8 shows the data projected onto the first two principal directions. The samples are

colored by their species, displaying a separation between the species, with an overlap between *versicolor* and *virginica*. Thus, PCA clearly reveals important structures in the data and allows for visualization and interpretation in a low-dimensional subspace.

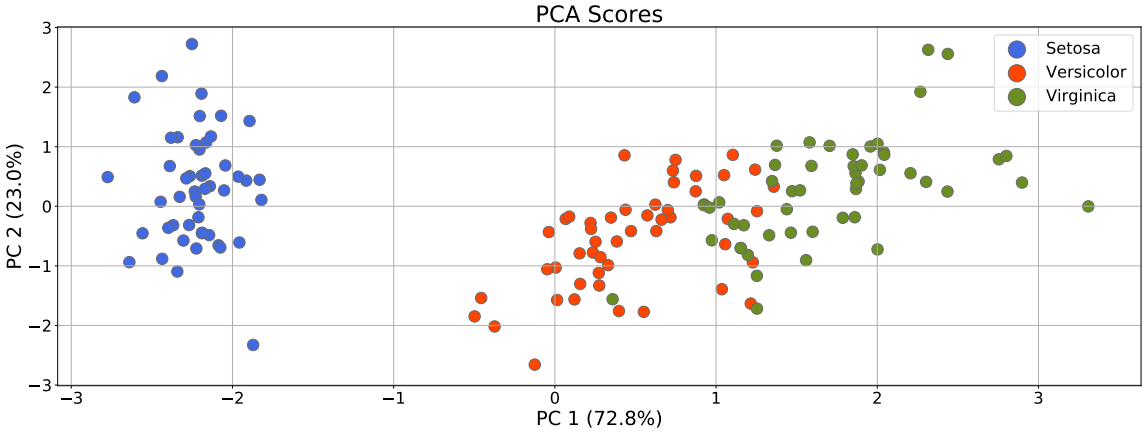

Figure 3.8: PC 1 and PC 2 scores for the Iris dataset.

To further investigate how the PCs are constructed as a linear combination of the original variables, the loading vectors v_j were investigated. In Figure 3.9 each loading vector v_j is represented horizontally as a linear combination of the variables, $v_j = C_1 l_s + C_2 w_s + C_3 l_p + C_4 w_p$, where $|v_j| = 1$. The size of the circles indicates the absolute size of C_i, while the color indicates the sign. Blue represents negative, while red represents positive.

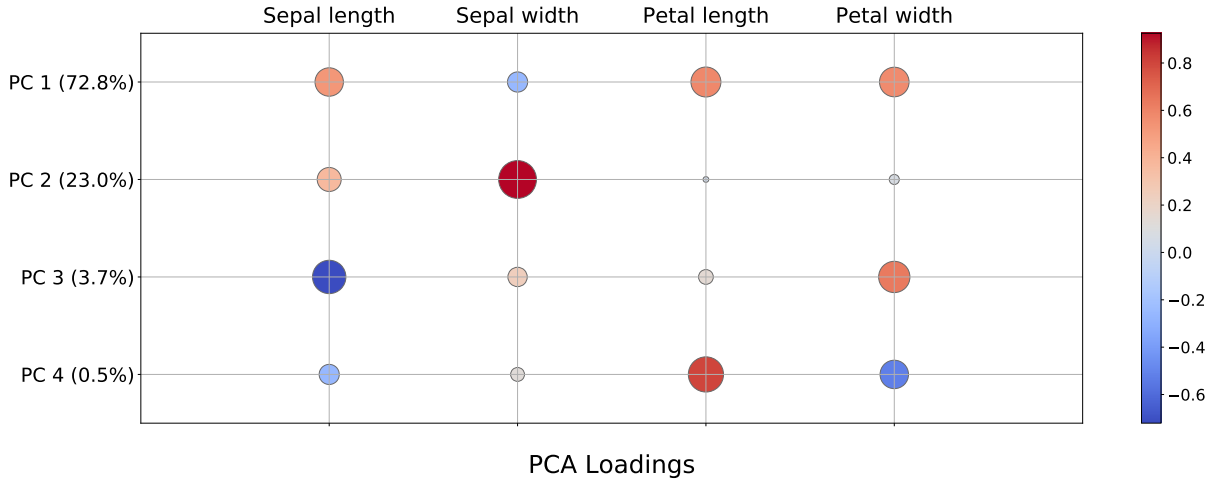

Figure 3.9: Graphical representation of the PC loadings.

Chapter 4

Preprocessing

This chapter summarizes the necessary preprocessing of the data. Section 4.1 covers the AIS data, Section 4.2 covers the atmospheric reanalysis data and Section 4.3 covers the vessel data. Section 4.4 presents the resulting datasets.

4.1 AIS Data

The AIS data were given in CSV-format with a total of 88706 data points during the three-year period. Each row contained the latitudinal and longitudinal position in decimal degrees with an associated timestamp. Timestamps were given in the OLE automation date format as a floating point value counting days since midnight 1899-30-12. Hours and minutes are represented as fractional days. The timestamps were converted to a "YYYY-MM-DD HH:MM:SS"-format and rounded off to the nearest second. The average value was stored when several data points were logged within the same second. This resulted in 88070 data points, or a 0.72 % reduction, with an average of 18 minutes between each logging. The maximum and minimum time difference between two data points were 8.82 days and 1 second respectively. Similarly, the median time difference was 8.3 minutes.

The data were resampled at a 10-minute frequency using a moving average filter placed at fixed positions. Missing periods were replaced using linear interpolation. All values at 00:00, 06:00, 12:00 and 18:00 were extracted, resulting in 4384 data point, or a 95.06 % reduction.

4.2 Atmospheric Reanalysis Data

The atmospheric reanalysis data were downloaded as two separate datasets, one with parameters from the atmospheric model and one with parameters from the ocean-wave model. Monthly, global data were downloaded as NetCDF-4-files with a temporal resolution of 6 hours and a spatial resolution of 0.125×0.125 degrees. In total this amounted to 18.7 GB of data, where only a fraction of it was extracted and used.

4.2.1 Data Extraction

In total for the three year period, 4383 data points were extracted based on the vessel's position at the timestamps specific for the atmospheric reanalysis data.

With a spatial grid of 0.125×0.125 degrees, the maximum longitudinal and latitudinal error between the vessel and an actual data point was 0.06 degrees, as expected. The ocean-wave data in the reanalysis is only available for points in the ocean, resulting in missing values when the vessel is sufficiently close to land due to the finite resolution of the spatial grid. Figure 4.1 illustrates this problem, with the vessel going from the Red Sea to the Mediterranean Sea through the Suez Canal. In total this occurred for 419 data points. To fill in the missing values we used the mean value of the neighboring cells, either using the 8 nearest neighbors in a 3×3 grid, the 24 nearest neighbors in a 5×5 grid or the 48 nearest neighbors in a 7×7 grid. If the full 7×7 grid was empty, NaN was inserted. In total this yielded 95 cells with missing data, corresponding to 2.17% of the data. Table 4.1 summarizes the imputation of the ocean-wave data.

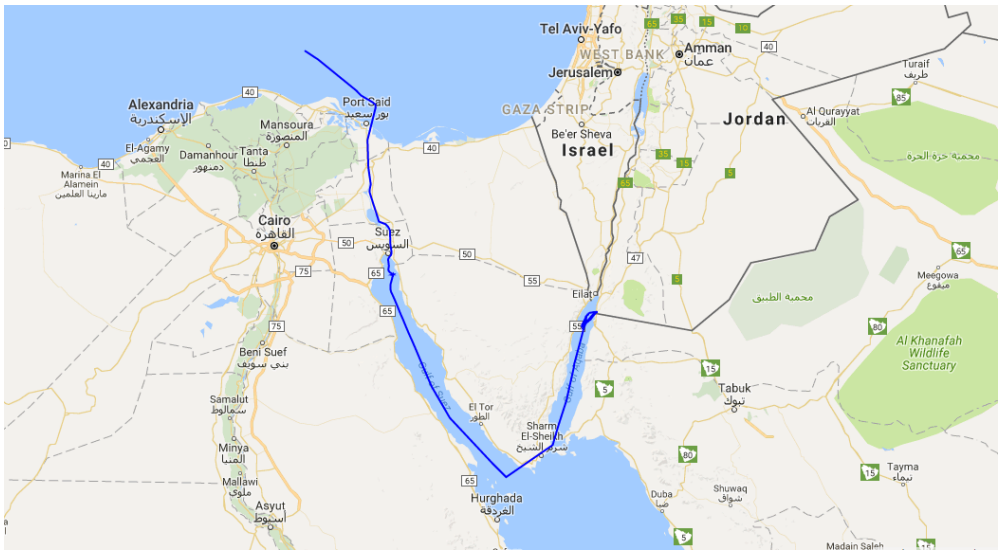

Figure 4.1: Vessel close to land, rendering ocean-wave data unavailable.

Parameter	Value
Total number of data points	4383
Missing values before imputation	419
Imputed with 3 × 3 grid	305
Imputed with 5 × 5 grid	86
Imputed with 7 × 7 grid	28
Missing values after imputation	95
Missing values before imputation	9.56 %
Missing values after imputation	2.17 %

Table 4.1: Summary of ocean-wave data imputation.

4.2.2 Data Transformation

Two transformations were applied on the atmospheric reanalysis data to obtain new variables. Specifically, temperature and dew point temperature were used to compute relative humidity, while the U and V wind components were used to compute the true wind speed and direction.

4.2.2.1 Relative Humidity

From the ambient and dew point temperatures, one can calculate the relative humidity, which for all pressures and temperatures is defined as

$$RH = 100\% \frac{P_w}{P_{ws}}, \tag{4.1}$$

where P_w is the water vapor pressure and P_{ws} the water vapor saturation pressure over water at the gas temperature. The relative humidity cannot reach 100 % when the gas temperature is below $0°C$, or when P_{ws} is greater than the atmospheric pressure in an unpressurized system.

Following[49], the water vapour saturation pressure over water and ice can be calculated as

$$P_{ws} = A \cdot 10^{\left(\frac{mT}{T+T_n}\right)}, \tag{4.2}$$

with a maximum error of 0.083 % when the temperature $-20°C \leq T \leq 50°C$ and $A = 6.116441$, $m = 7.591386$ and $T_n = 240.7263K$. Using (4.2) we can express the relative humidity as

$$RH = 100\% \cdot 10^{m\left[\frac{T_d}{T_d+T_n} - \frac{T}{T+T_n}\right]}, \tag{4.3}$$

where T_d is the dew point temperature and T is the ambient temperature.

4.2.2.2 True Wind

From the northward and eastward wind components, U and V, one can compute the true wind speed and direction by finding the length r of the vector defined by the two components and its angle α relative to the eastward axis.

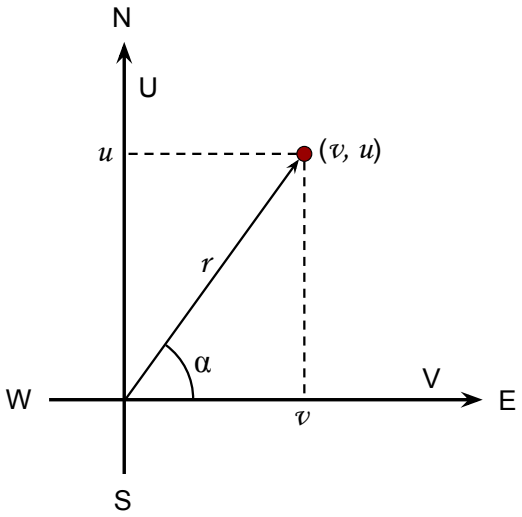

Figure 4.2: True wind speed and direction.

Figure 4.2 illustrates this transformation. The resulting angle α was defined as 0 degrees for eastward wind and 180 degrees for westward wind.

4.2.3 Data Validation

To validate the data, the ambient conditions from the reanalysis were compared with the measurements gathered by the vessel. Figures 4.3 and 4.4 compare the ambient temperature and pressure from the two sources, showing a strong correlation in the variation. The Pearson correlation coefficients are $r = 0.90$ and $r = 0.84$ for the ambient temperature and pressure respectively.

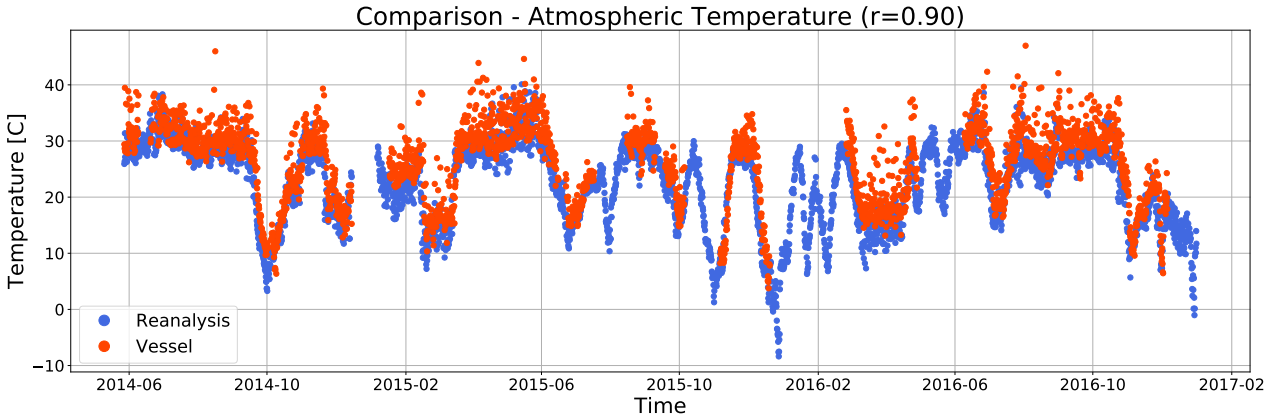

Figure 4.3: Atmospheric temperature from the reanalysis compared with atmospheric temperature measured by the vessel.

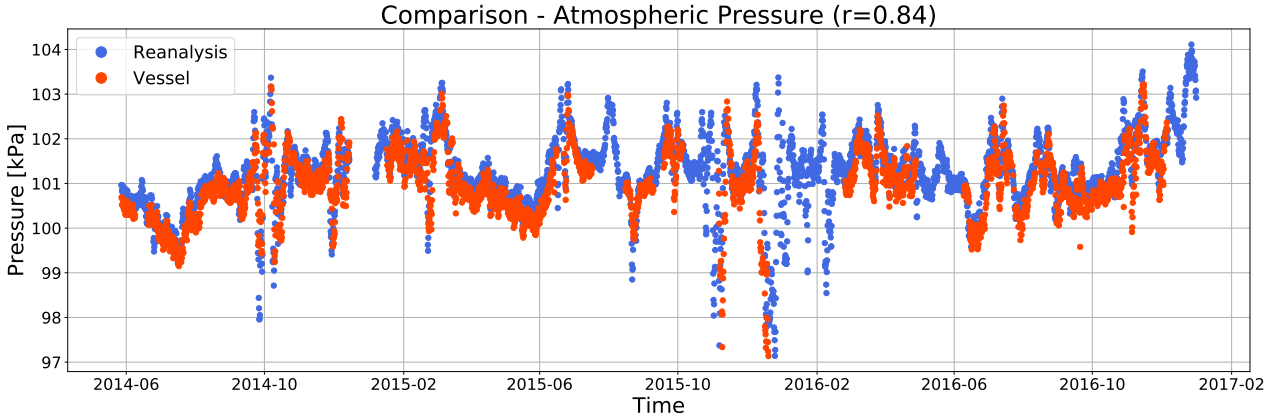

Figure 4.4: Mean sea level pressure from the reanalysis compared with atmospheric pressure measured by the vessel.

In Figure 4.5 the true wind speed is plotted against the significant wave height. It shows a strong correlation with $r = 0.73$ as one would expect since a large portion of the waves is

wind generated. Similarly, Figure 4.6 shows the true wind direction plotted against the mean wave direction, with a correlation of $r = 0.48$. The results indicate consistency within the reanalysis data.

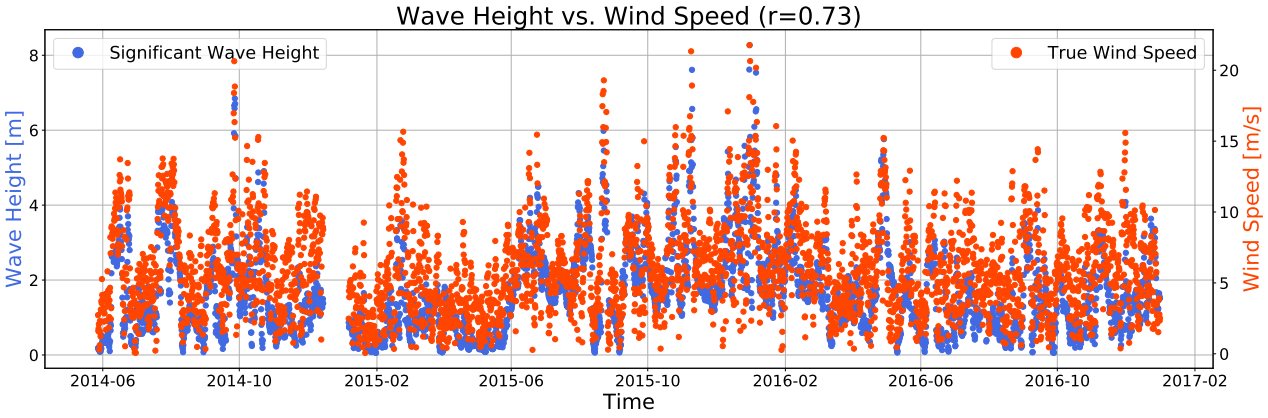

Figure 4.5: True wind speed and significant wave height.

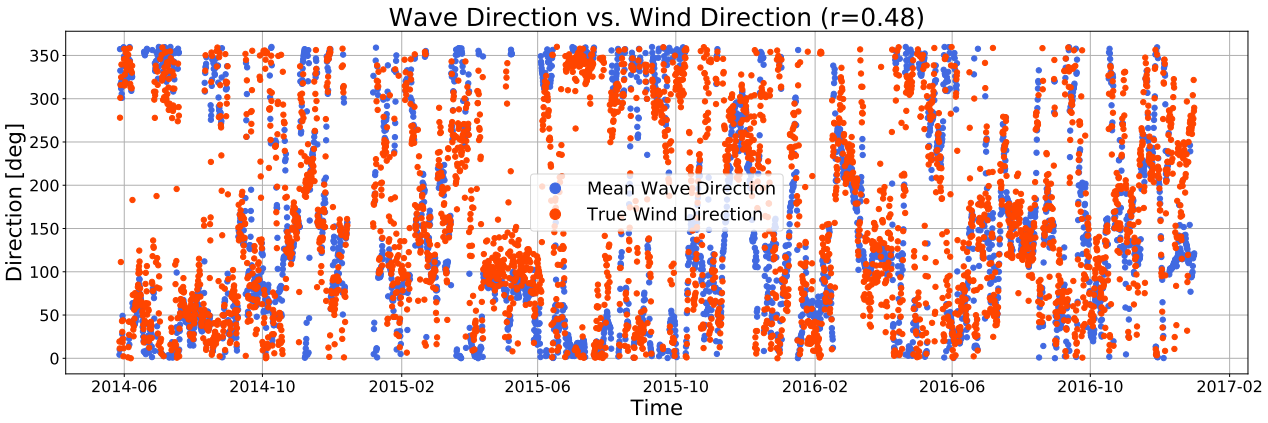

Figure 4.6: True wind direction and mean wave direction.

4.3 Vessel Data

The available vessel data were given in six-hour CSV-files containing over 300 variables logged at varying frequency. The data logging system is constructed such that a variable is logged every time it changes more than a certain threshold, but with no given fixed logging frequency. We have no direct knowledge about the specific sensors, their sampling frequencies or the onboard data handling system. From the vessel data, the variables previously listed in Table 2.7 were extracted with the timestamps rounded off to the nearest second. For multiple values within one second, the average of value was stored. For all the data available

within the three-year period, Table 4.2 lists the number of data points for each variable after the timestamps were rounded off.

Variable	No. of measurements
Atmospheric temperature	25674573
Atmospheric pressure	209027
Cargo level, tank 1	16764
Cargo level, tank 2	20275
Cargo level, tank 3	20436
Cargo level, tank 4	24418

Table 4.2: Number of data points in the given period for the selected variables.

4.3.1 Cargo Levels

To analyze the change of cargo level as a function of ambient conditions, the laden conditions were extracted and cleaned. The outlined process was applied to all four cargo levels. Figure 4.7 shows the cargo level for tank 1. We see a clear separation between laden and ballast conditions with loading and unloading during the transitions. By inspecting the cargo levels, it is evident that the levels are always higher than 20 meters when in laden condition. All points $x < 20$ m were removed. Several points from loading and unloading conditions and obvious outliers were still contained in the data.

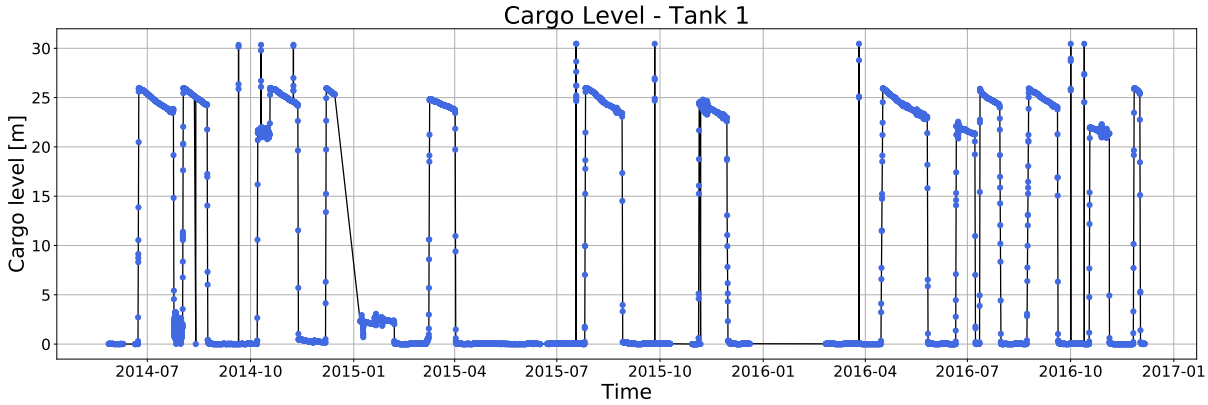

Figure 4.7: Cargo level tank 1, original data.

To isolate the laden conditions an outlier removal using a k-nearest neighbors algorithm was performed. The timestamps of the measurements were transformed from date-time strings to floating point numbers between 0 and 50, resulting in a dataset $\mathbf{X} \in \mathbb{R}^{N\times 2}$, where N is the

number of observations. The distances from any sample $x_i \in \mathbb{R}^2$ to its 8-nearest neighbors were computed, a vector denoted $d_i \in \mathbb{R}^8$. The samples were sorted by the Euclidean norm of the distance vector, $||d_i||$. Samples with a Euclidean norm larger than a tunable threshold t were marked as outliers. With a low t, too many points from the laden conditions were removed, while a high t allowed too many obvious outliers to remain in the dataset. In our case, for all four cargo levels, using $t = 0.23$ provided a sufficiently clean extraction of the laden conditions. Figure 4.8 shows the extraction of voyage 9 from tank 4. Table 4.3 summarizes the number of outliers removed.

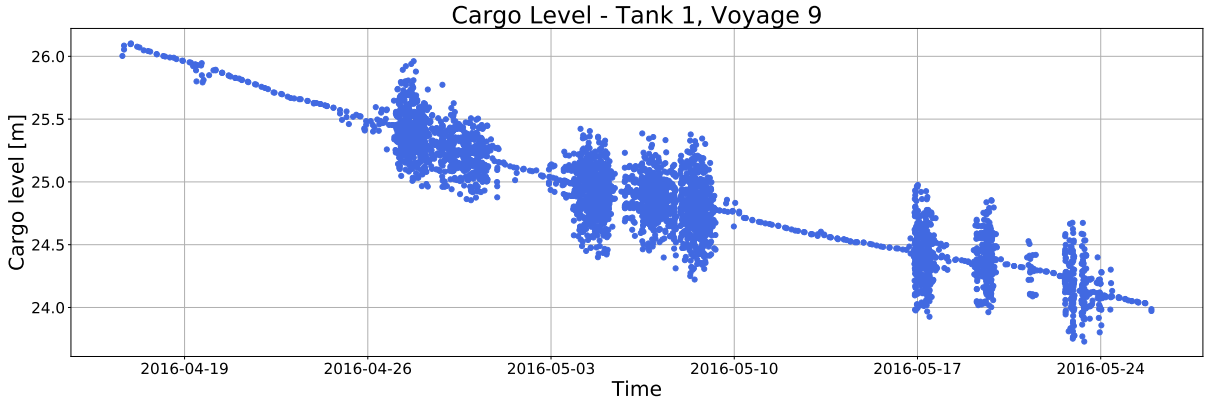

Figure 4.8: Cargo level tank 4, voyage 9.

Tank no.	No. of Outliers	Percentage Removed
1	76	1.41 %
2	81	0.54 %
3	84	0.62 %
4	92	0.71 %

Table 4.3: Summary of cargo level outlier removal.

From the extracted laden conditions the measurements were separated into specific voyages. Inspecting the distinct voyages revealed a lot of noise (especially in tank 2, 3 and 4) and large variations in point density. As can be seen in Figure 4.8 there are several segments of high frequency, high variance data and several segments of low frequency, low variance data. In order to model the change in cargo level using ambient conditions, we would like a smoothed cargo level curve with measurements equally spaced in time. This would give us a time-independent model where each computed change in cargo level is based on the same time interval.

More specifically, since the atmospheric reanalysis data were available every six hours at 00h,

06h, 12h, and 18h, it would be convenient to have cargo level measurements every six hours at 03h, 09h, 15h, and 21h. For every computed change in cargo level Δy_i over six hours, we would then have the ambient conditions x_i at the midpoint during the period of change. This is illustrated in Figure 4.9.

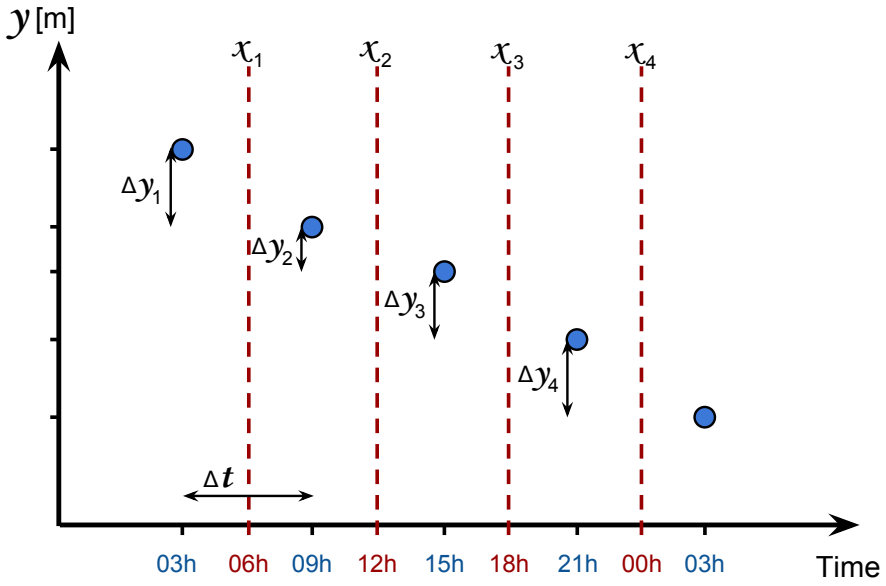

Figure 4.9: Spacing of cargo level and ambient conditions.

The mean time period between two consecutive measurements was 1.07 hours for tank 1, 0.47 hours for tank 2, 0.53 hours for tank 3 and 0.55 hours for tank 4. The difference is due to fewer segments with high frequency and high variance measurements in tank 1 than the rest. The cargo levels are resampled at a six-hour frequency, using an open interval on the left side and a closed interval on the right side. The arithmetic mean of the measurements within the interval was stored at the right side label with a negative offset of three hours. In other words, all observations from 00:00 to and included 06:00 were averaged and stored at 03:00 *etc.*, as shown in Figure 4.10. This works as a central moving average filter placed at fixed positions, resulting in equally spaced measurements and noise reduction. For intervals containing no measurements, NaN was inserted. Table 4.4 summarizes the number of measurements in the individual tanks before and after filtering and the number of missing values after filtering. As there were few missing values, where each missing value indicated more than six hours without a measurement, the segments were removed from the data.

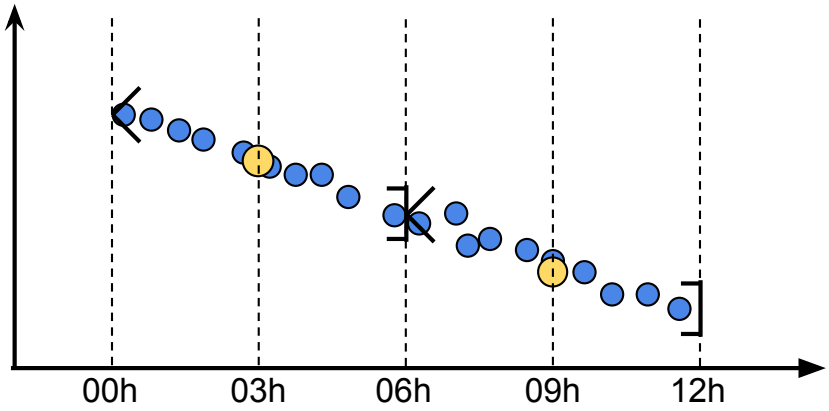

Figure 4.10: Illustration of cargo level resampling using central moving average filtering

No. of measurements	Tank 1	Tank 2	Tank 3	Tank 4
Before filtering	6534	14889	13380	12688
After filtering	1175	1185	1186	1185
Missing values after filtering	14	14	14	14

Table 4.4: Cargo level measurements before and after six hour moving average filtering.

Figure 4.11 shows the distribution of Δy for each tank after resampling and filtering. The distribution curves were estimated using kernel density estimation (KDE) and plotted together with the mean values and standard deviations. The distributions are narrow-banded and long-tailed, indicating the presence of noise. The mean values represent the mean change in cargo level over six hours.

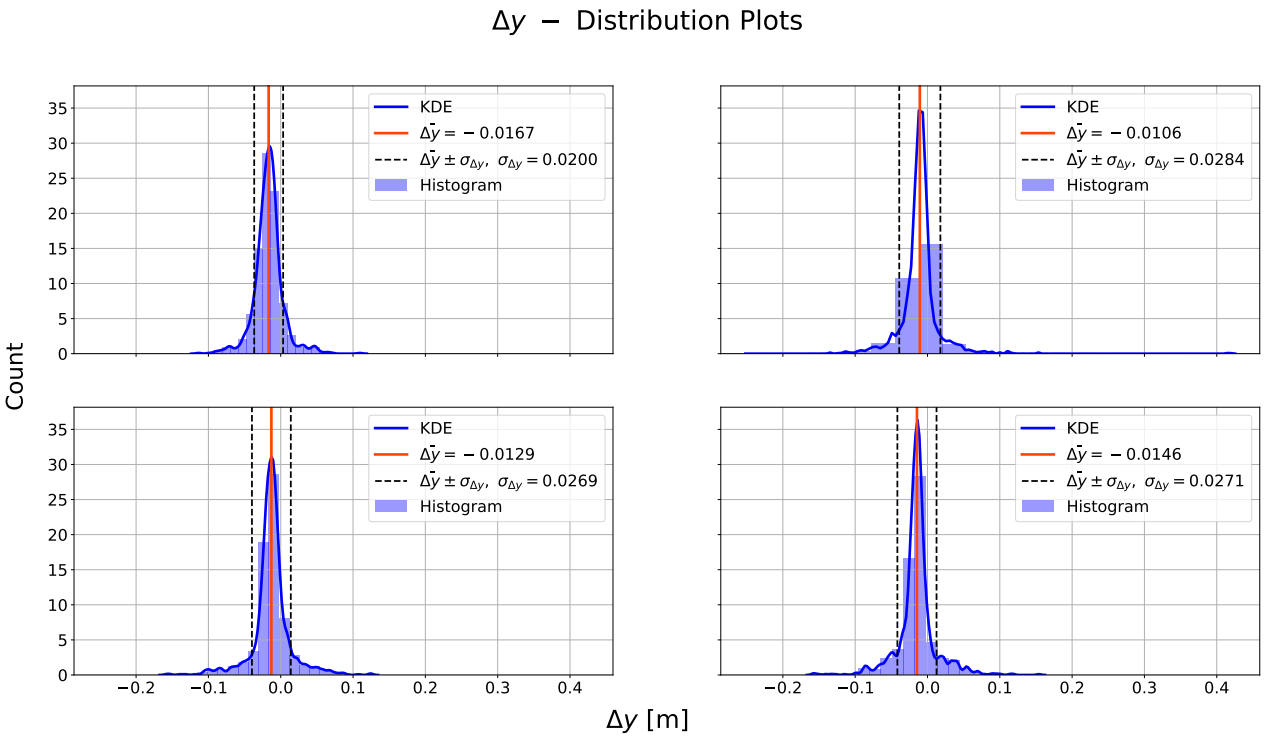

Figure 4.11: Distribution plots of Δy after moving average filtering.

To illustrate, Figure 4.13 and 4.14 show the cargo level in tank 3 for two selected voyages. A clear noise reduction is seen after applying the moving average filter (yellow), but with large variations still remaining in the data.

LOWESS smoothing was applied to capture the local shape of the curves while sufficiently removing noise. With varying voyage lengths, the smoothing parameter α_v, the fraction of the total number of data points N_v used in each local fit for voyage $v = 1, \ldots, 14$, was chosen such that the fraction spans 16 measurements. In some cases the resulting α was too large, and in other cases too low. Due to this α_v was constrained such that $0.15 \leq \alpha_v \leq 0.35$. This yielded a dynamic α_v based on the number of points N_v in each voyage. Figures 4.13 and 4.14 illustrate the LOWESS smoothing (red) for two selected voyages. The distribution of Δy after applying LOWESS smoothing is shown in Figure 4.12. The standard deviations are reduced by an order of one magnitude after smoothing, removing the previously seen long tails.

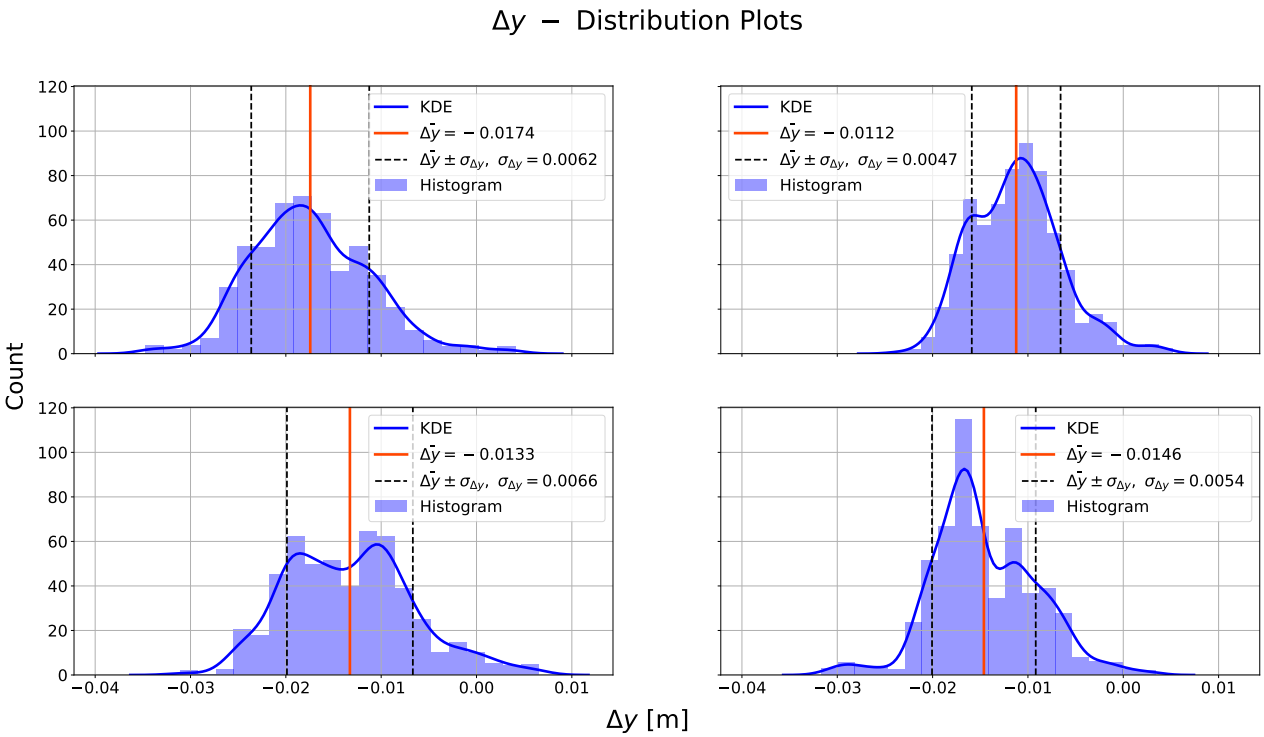

Figure 4.12: Distribution plots of Δy after LOWESS smoothing.

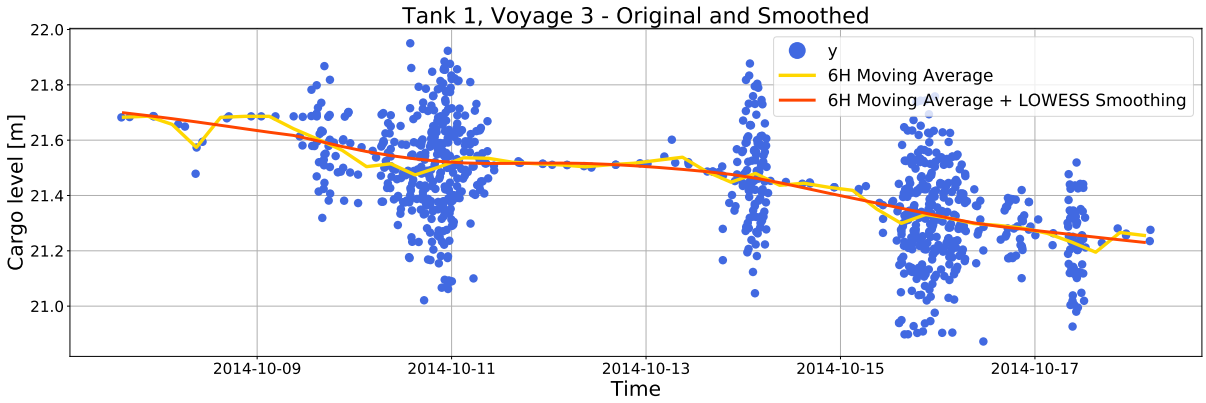

Figure 4.13: Cargo level tank 1, voyage 3, before and after noise reduction.

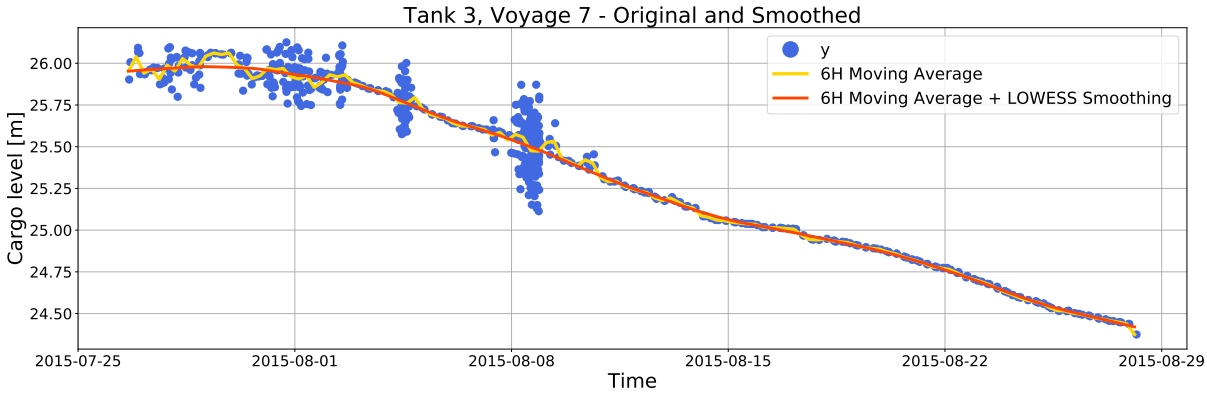

Figure 4.14: Cargo level tank 3, voyage 7, before and after noise reduction.

4.4 Combined Dataset

The extracted atmospheric reanalysis data were combined with the smoothed cargo levels for laden conditions. With cargo level measurements at 03h, 09h, 15h, and 21h the computed difference Δy was combined with the corresponding ambient conditions at the midpoint of the period of change. As $\Delta y = y_{k+1} - y_k$, where y_{k+1} will be a function of y_k due to the geometry of the tanks, y_k (the previous cargo level) was added to the datasets.

Table 4.5 shows the structure of the resulting datasets, where P_{atm} is the atmospheric pressure, T_{atm} the atmospheric temperature, $H_{1/3}$ the significant wave height, T_1 the mean wave period, v_{wind} the wind speed and RH the relative humidity. The voyage number describes which of the 14 voyages the measurements belong to. Each observation is time independent in the sense that they represent the change over equal periods of time.

Δy	y_{prev}	P_{atm}	T_{atm}	$H_{1/3}$	T_1	v_{wind}	RH	Voyage no.
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-

Table 4.5: Structure of combined dataset.

With four different tanks with different conditions and small variations in voyage lengths, five datasets with the given structure were constructed, one for each tank and one for the cargo levels combined. Table 4.6 summarizes the number of observations in each dataset and the percentage reduction compared to the original cargo level measurements. This will constitute the available data when applying unsupervised and supervised learning methods in Chapter 5. In the unsupervised framework, all variables are standardized to have zero mean and unit variance, while in the supervised framework Δy is not transformed as it is considered as a target variable Y.

Dataset	No. of observations	Percentage reduction
Tank 1	1098	93.45 %
Tank 2	1108	94.54 %
Tank 3	1109	94.57 %
Tank 4	1108	95.46 %
Intersection	1089	-

Table 4.6: Number of observations in the combined datasets and their intersection.

Chapter 5

Analysis

In this chapter, various methods from Chapter 3 are used to explore the underlying relationship between the individual cargo levels and the ambient conditions obtained from the atmospheric reanalysis data. The chapter will be divided into two main parts: In the first part, Section 5.1, a polynomial model is constructed to simulate cargo levels from ambient conditions. A methodology for regression is employed to recover the relationships in the original model. In the second part, Section 5.2 the same methodology is applied to real-world data.

5.1 Simulated Data

We assume an ideal situation where the boil-off rate, and thus the observed change in cargo level can be explained by ambient conditions and cargo level height alone. Since we know that the crew can use forced boil-off for propulsion and that the LNG composition affects the boil-off rate, the model will fail to encapsulate the true dynamics. Due to the highly complex nature of the boil-off phenomenon, dependent on computational fluid dynamics, thermodynamics and stochastic tank sloshing, we are forced to consider a crude model displaying the relationships in their simplest terms.

5.1.1 Modeling

We assume that the cargo level at step $k+1$ can be modeled as the cargo level at step k plus some variation as a function of ambient conditions and the cargo level at step k itself. Additive Gaussian noise was assumed, such that

$$\begin{aligned} y_{k+1} &= y_k + f(y_k, x_k) + \epsilon \\ \Delta y_{k+1} &= f(y_k, x_k) + \epsilon, \end{aligned} \tag{5.1}$$

where x_k is a vector of relevant ambient conditions at the midpoint between time steps k and $k+1$ (see Figure 4.9), and $\epsilon \sim N(0, \sigma^2$. The time period between two consecutive values was set to be 6 hours.

Referring to Figure 2.2, we know that the tanks are rectangular in the alongship direction while having an octagonal shape in the cross-section. Thus, when the tanks are filled to 98.5 % of its capacity the cargo levels will decrease nonlinearly as a function of y_k. Furthermore, using the results from Hasan et al.[3], the boil-off rate was assumed to increase linearly with ambient temperature T_{atm}. It was also assumed that the boil-off rate decreases linearly with ambient pressure P_{atm}.

According to linear wave theory, the average energy in a sea state is proportional to the wave height squared, *i.e.* $E \propto H_{1/3}^2$. By assuming a fully developed sea state and linear wave theory, waves can be represented in the frequency domain by a wave spectrum $S(\omega)$. In the statistical description of waves, a common spectrum is the empirical Pierson-Moskowitz (PM) spectrum, proposed by Pierson and Moskowitz[50]. Using the PM spectrum $S(\omega)$, one can find that $H_{1/3}^2$ is proportional v_{wind}^4. Thus, it was assumed that the cargo level decrease as a function of $H_{1/3}^2$ and v_{wind}^4, attributed to the energy in a sea state. Furthermore, it was assumed that to decrease as a function of T_1^2.

Based on the assumptions a polynomial function of 4th order

$$\begin{aligned} f(x_k) = &- 0.0001\, y_k - 0.00002\, y_k^2 + 0.0005 P_{\text{atm}} - 0.002\, T_{\text{atm}} \\ &- 0.001\, H_{1/3} - 0.0002\, H_{1/3}^2 - 0.0008\, T_1 - 0.00015\, T_1^2 \\ &- 0.001\, v_{\text{wind}} - 0.00002\, v_{\text{wind}}^4 \end{aligned} \tag{5.2}$$

was used. No cross-term interactions were included due to simplicity. Relative humidity was not included in the model, allowing for an assessment of the variable selection methodology. The different coefficients were chosen by trial and error to simulate a well-behaved y. Since

the simulated cargo level is dependent on its previous value, y_k as an input can not be standardized to have zero mean and unit variance. Thus the coefficients before the y_k terms are not directly comparable to the other coefficients, whose inputs are standardized.

To assess the relative importance of the input variables in the model we performed a one-at-a-time sensitivity analysis on the input parameters and ranked the inputs by the sensitivity index, local sensitivity and output variance, as explained in Appendix A. Proper probability density functions were assigned to each of the input variables, such that random samples could be drawn. For each ambient condition, an allowable range was set based on physical consideration and observed values. Then, for each variable several distributions were fitted to its histogram, and the distribution yielding the lowest residual sum of squares was chosen. For the cargo level, we assume a uniform distribution within the observed range. The variables, their ranges and their assigned distributions with parameters are summarized in Table 5.1. As beta distributions are defined on $[0,1]$, they were scaled and shifted to fit the histograms. For the half-normal distribution, a scaled and shifted standardized distribution was used. Figure 5.1 shows the histograms of the ambient conditions and their assigned distributions.

Variable	**Variable range**	**Best fit distribution**	**Distribution parameters**
y	$[y_{\text{min}}, y_{\text{max}}]$ [m]	Uniform	$a = y_{\text{min}}, b = y_{\text{max}}$
P_{atm}	[97000, 105000] [Pa]	Beta	$\alpha = 108771.64, \beta = 86.55$
T_{atm}	[-10, 50] [C°]	Beta	$\alpha = 1043.16, \beta = 6.55$
$H_{1/3}$	[0, 9] [m]	Half-normal	$\sigma = 1$
T_1	[0, 16] [s]	Beta	$\alpha = 2.29, \beta = 3.10$
v_{wind}	[0, 25] [m/s]	Beta	$\alpha = 2.41, \beta = 9.76$
RH	[60, 100] [%]	Beta	$\alpha = 14.28, \beta = 2.87$

Table 5.1: Assigned distributions for each input variable.

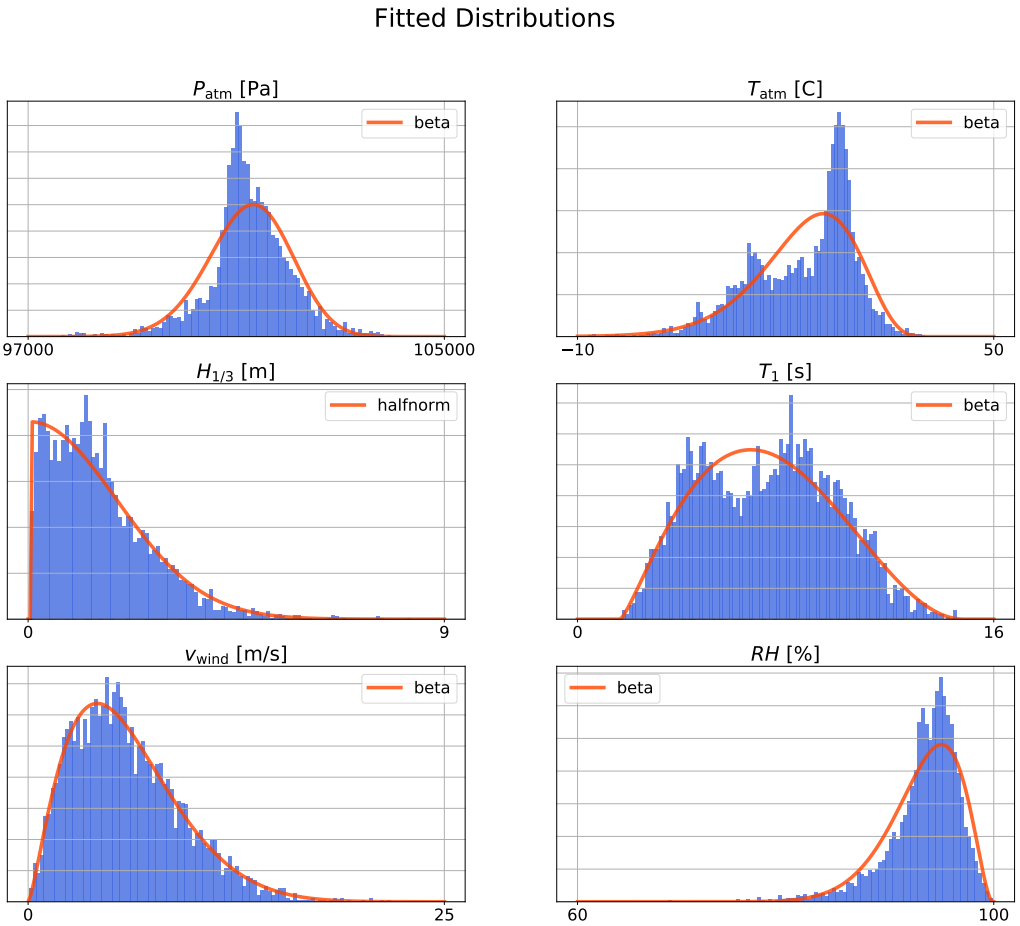

Figure 5.1: Histograms of the ambient conditions and their assigned distributions.

2000 samples were randomly drawn from the assigned distributions and standardized. Partial derivatives of (5.2) with respect to each variable were evaluated at the mean input vector to obtain the local sensitivities. Then model outputs were computed using samples drawn for one variable at a time, while the rest were kept fixed at their mean value. From the computed outputs, the sensitivity indices and output variances were calculated. Figure 5.2 shows the normalized sensitivity measures for each variable for comparison. The model is not sensitive to RH as it is not a model input. The lowest ranking model input was P_{atm}, with T_1 slightly more influential. y, $H_{1/3}$, and v_{wind} displayed a similar degree of influence, except for a higher output variance due to changes in y, possibly accounted for by the fact that it was sampled from a uniform distribution. The overall highest ranking variable was T_{atm}.

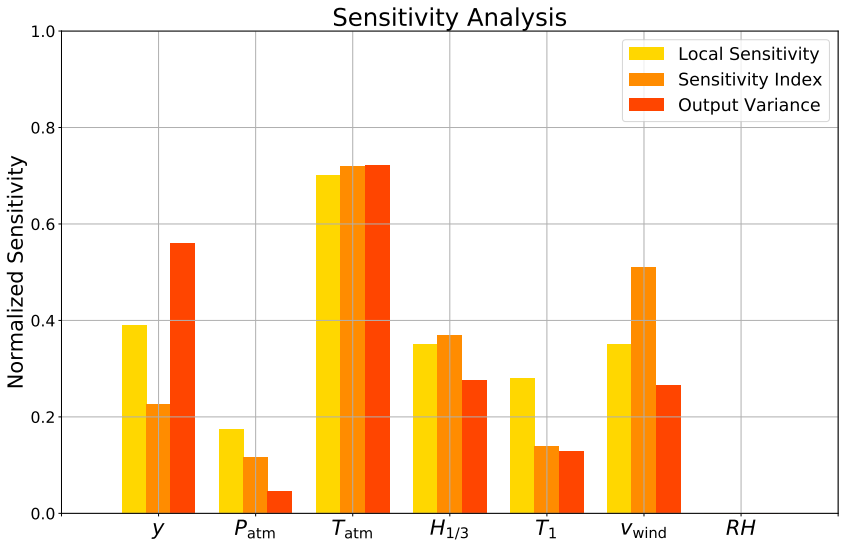

Figure 5.2: Normalized sensitivity measures for each variable in the simulation model.

5.1.2 Simulation

To simulate the data the initial cargo level y_0 was specified for each voyage. Ten voyages were simulated using real-world ambient conditions in the period January 2016 to September 2016 with 23 m $\leq y_0 \leq$ 28 m and varying voyage length. The ambient conditions were standardized before use. Table 5.2 summarizes the initial conditions for the ten voyages.

Voyage no.	**Start Time**	**Start Level [m]**	**Voyage Length [days]**
1	2016-01-01, 03:00:00	27.2	11
2	2016-01-12, 09:00:00	25.6	16
3	2016-01-28, 21:00:00	26	19
4	2016-02-17, 03:00:00	23.8	18
5	2016-03-06, 15:00:00	27.5	32
6	2016-04-08, 03:00:00	24.3	23
7	2016-05-01, 09:00:00	26.6	27
8	2016-05-28, 21:00:00	27.7	27
9	2016-06-28, 03:00:00	24.7	25
10	2016-08-01, 03:00:00	26.3	17

Table 5.2: Initial conditions for the simulated voyages.

The simulated cargo level displayed a clear nonlinear effect as shown in Figure 5.3. By defining the signal-to-noise ratio (SNR) as the ratio between the variance of Δy and the variance of ϵ,

$$\text{SNR} = \frac{\sigma^2_{\Delta y}}{\sigma^2_\epsilon}, \tag{5.3}$$

the standard deviation of the noise could be computed for a given SNR. Figure 5.5 shows the distribution of Δy for SNR $= \infty$ (without noise), SNR = 10, SNR = 7, and SNR = 2. SNR = 7 was chosen for the final model. Figure 5.4 shows the simulated Δy with and without noise and its mean value $\overline{\Delta y} = -0.0154$ m/6 hours.

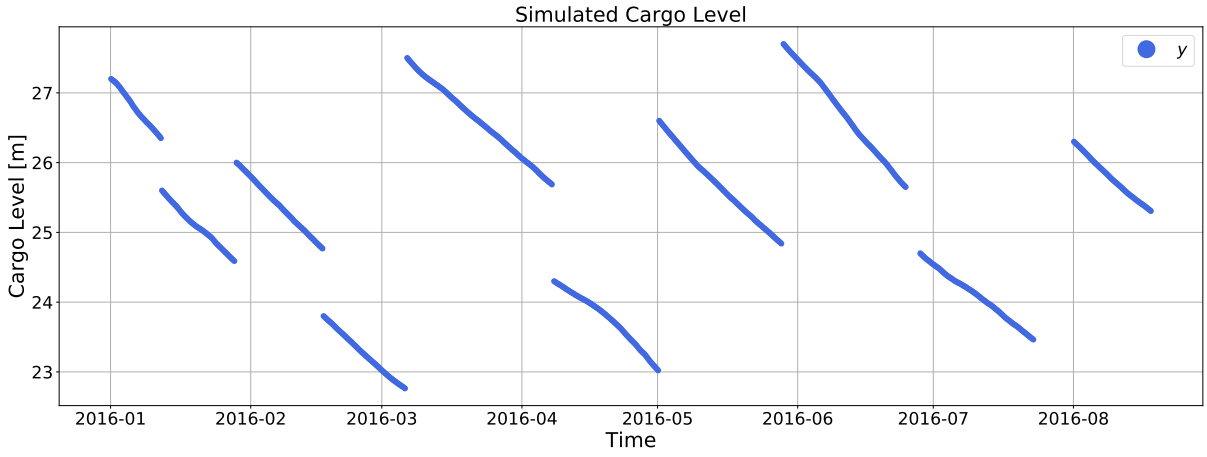

Figure 5.3: Simulated cargo level y.

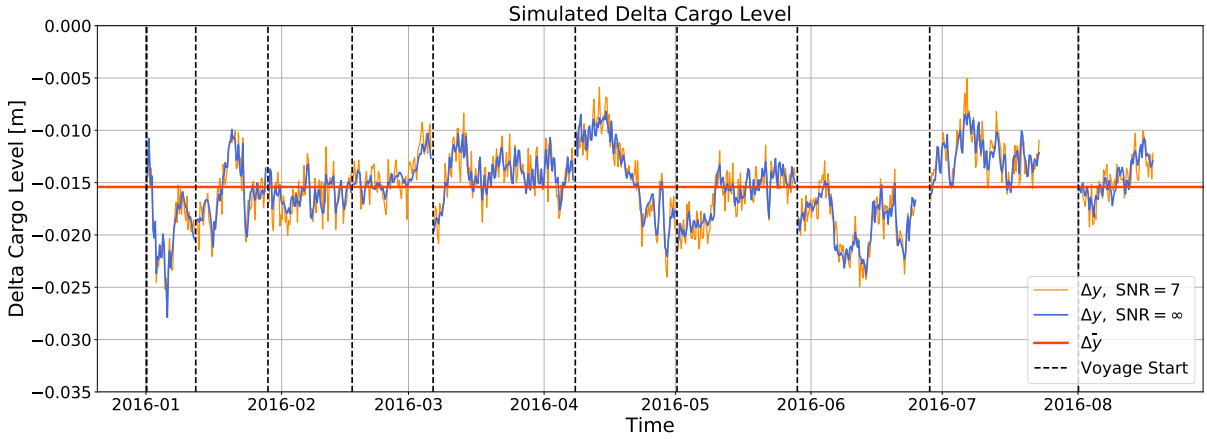

Figure 5.4: Simulated change of cargo level Δy.

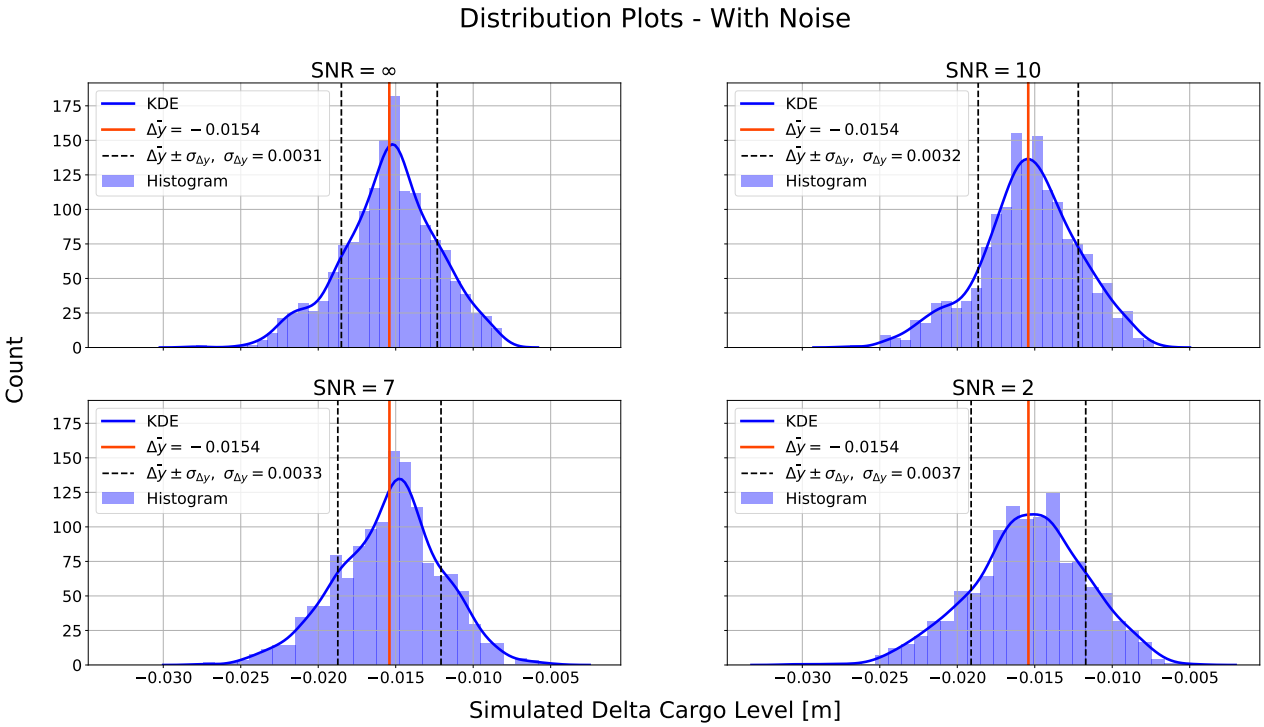

Figure 5.5: Distribution of simulated Δy for different SNR.

5.1.3 Principal Component Analysis

The simulated data were investigated in the unsupervised framework using PCA. A dataset $\mathbf{X}$ containing the ambient conditions and the simulated y and Δy was standardized and decomposed into its principal components. Figure 5.6 shows the cumulative explained variance of the model as a function of principal components included, *i.e.* the model complexity.

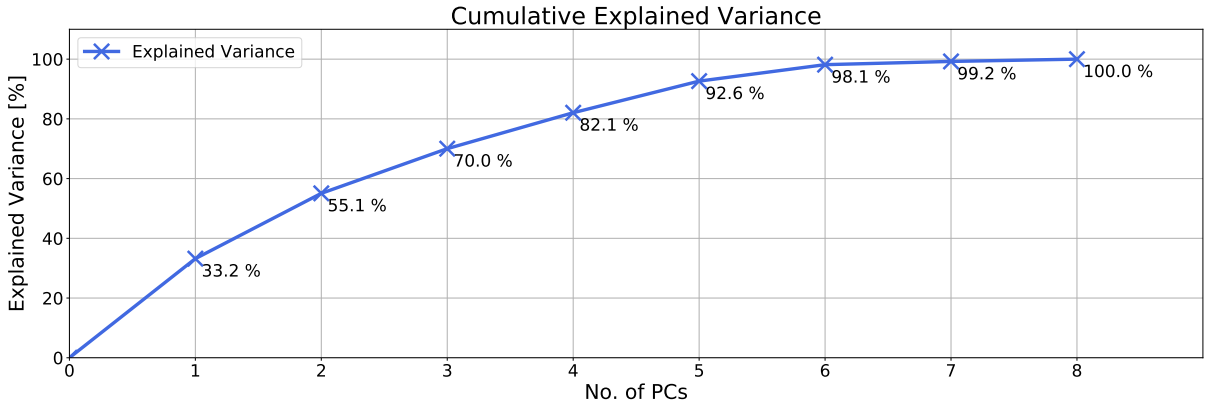

Figure 5.6: Cumulative explained variance as a function of principal components.

Figure 5.7 visualizes the loadings for all variables over all components. In PC 1 we see a cor-

relation between waves and wind, which again is negatively correlated to Δy. Thus rougher weather is correlated with a higher boil-off rate. Similarly, on PC 2, Δy is correlated with P_{atm} and negatively correlated with T_{atm} and y. The results reflect the polynomial model in (5.2). Less obvious relationships intrinsic to the weather data were found in the higher components and not considered in any depth.

Figure 5.7 shows the loadings in the two-dimensional space spanned by PC 1 and PC 2 to the left, and PC 2 and PC 3 to the right. Together these components explained 70 % of the variance in the data. In the left figure, we see a separation of the variables into three groups. The clusters indicate that warmer weather is correlated with calmer sea states, while Δy is negatively correlated with both warmer weather and rougher sea states. This seems to indicate a trade-off between atmospheric conditions and sea conditions in terms of boil-off. On one hand, warm weather and calm sea lead to boil-off due to temperature, while on the other hand cold weather and rough sea lead to boil-off due to sloshing. The figure on the right clearly shows the inverse relationship between P_{atm} and T_{atm} and similarly between Δy and y.

In Figures 5.9 and 5.10, the samples in $\mathbf{X}$ are projected onto the two-dimensional space spanned by two components. Figure 5.9 shows the scores in PC 1 and PC 2 colored by a selection of the variables. Similarly, Figure 5.10 shows the scores in PC 2 and PC 3. This was used to further visualize the relationships explained above. Samples where Δy is most negative, *i.e.* where the BOR is highest, have a large y and low pressure with either high temperature or rough sea conditions.

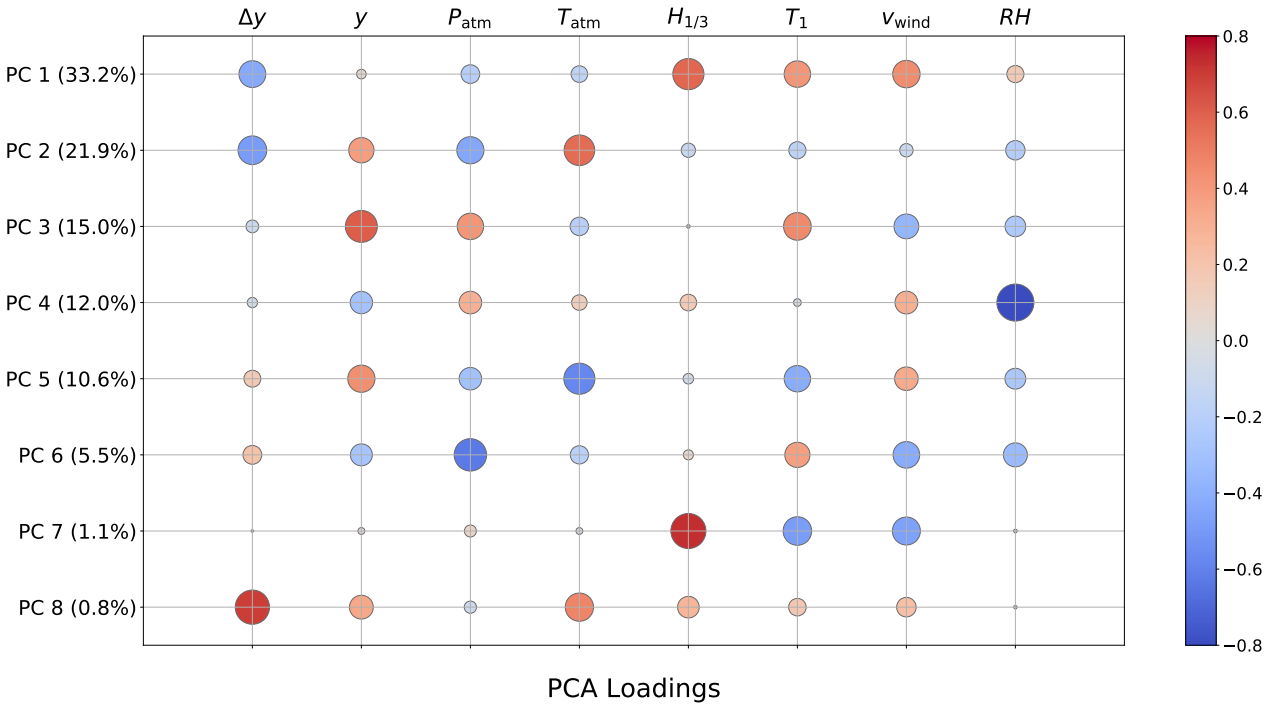

Figure 5.7: PC loadings for all components.

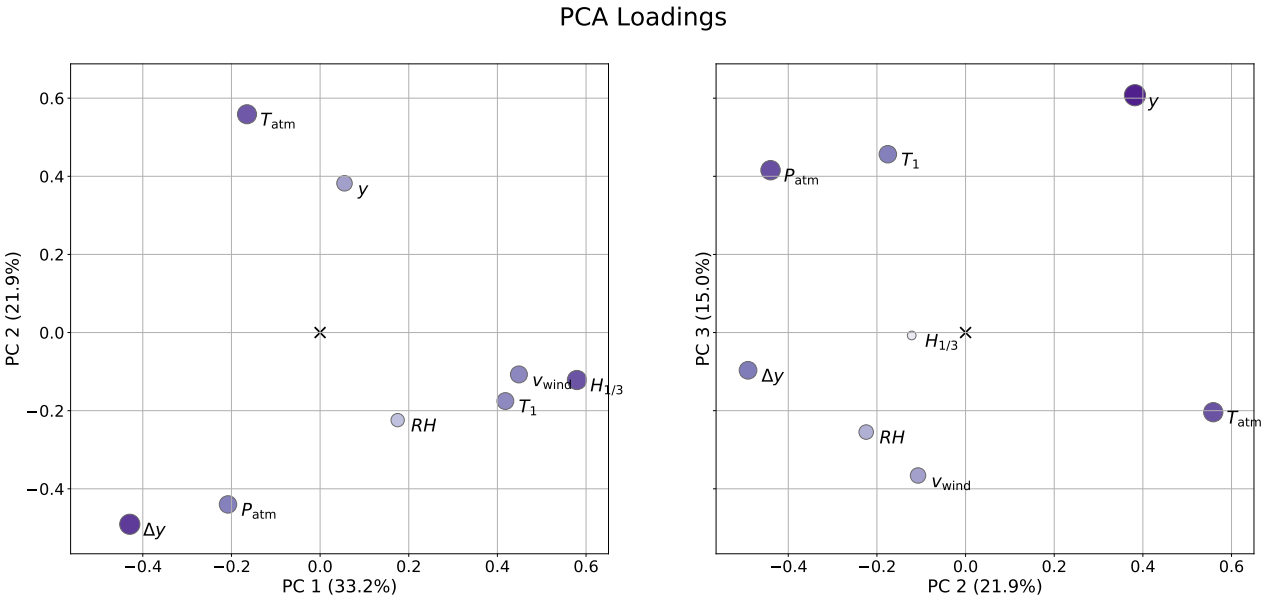

Figure 5.8: PC loadings plot for PC 1, PC 2 and PC 3.

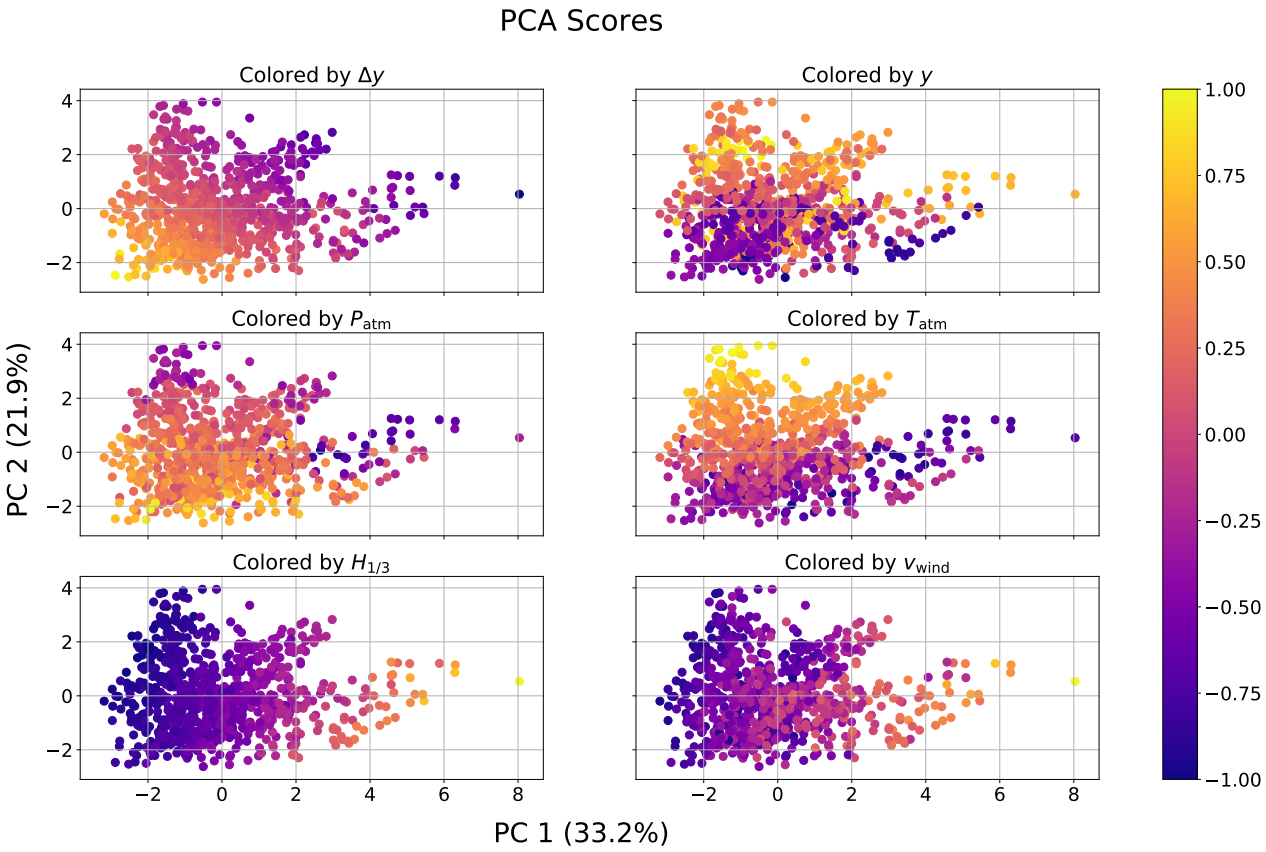

Figure 5.9: PC 1 and PC 2 scores colored by different variables.

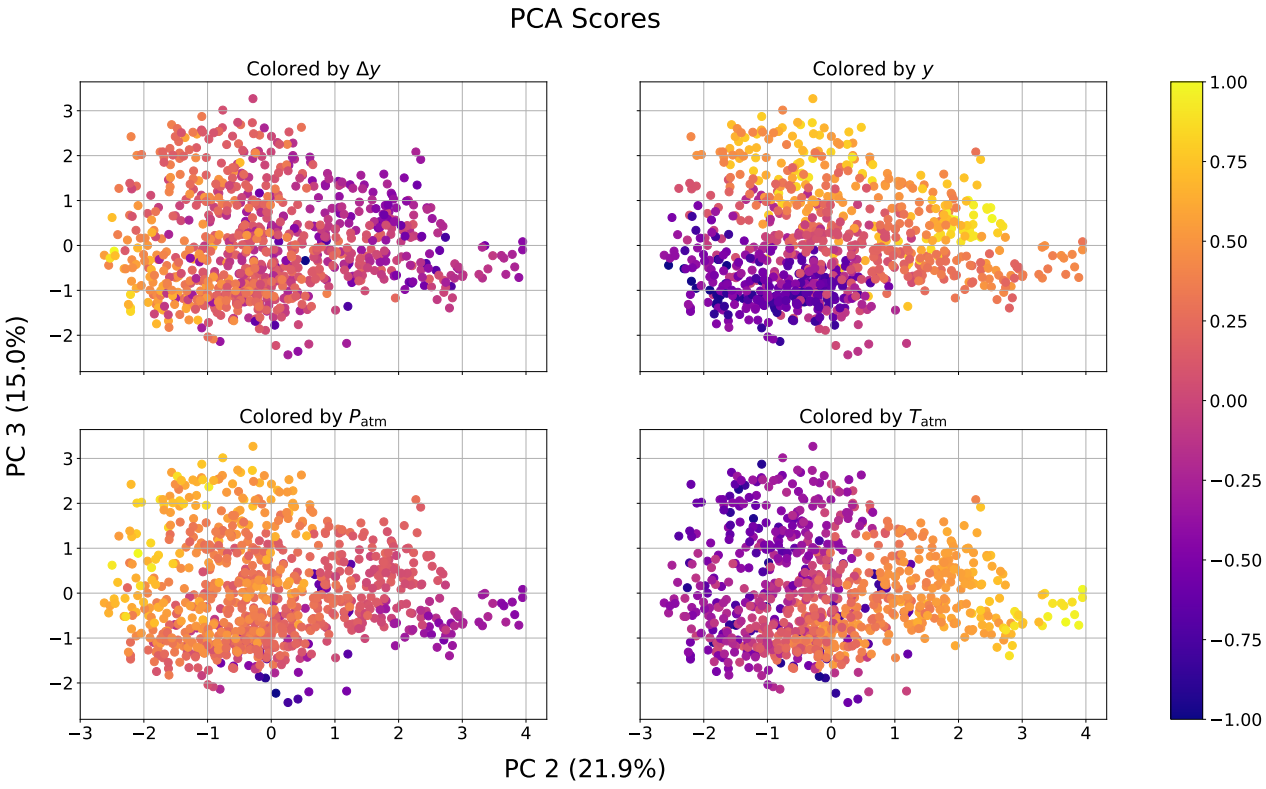

Figure 5.10: PC 2 and PC 3 scores colored by different variables.

5.1.4 Regression Analysis

For the regression analysis on the simulated data we considered four cases, as presented in Table 5.3: In case 1 and 2 the original, untransformed variables were used as inputs, with the ocean-wave data removed in case 2. In case 3 and 4 the first two cases were repeated with a polynomial basis expansion of the input variables. while allowing 2nd, 3rd and 4th polynomial transformations of the input variables. 2nd, 3rd, and 4th-degree polynomial transformations were considered without cross-terms. This resulted in 7 inputs in case 1, 5 in case 2, 28 in case 3 and 20 in case 4.

	All variables	Ocean-wave data missing
Untransformed variables	**Case 1**	**Case 2**
Polynomial variables of 4th degree	**Case 3**	**Case 4**

Table 5.3: Four cases used for regression analysis on the simulated data.

All input variables were standardized. The data were shuffled and split into 70 % training data and 30 % independent test data. Linear regression models were trained for all four cases with forward-stepwise selection. Nearest-neighbors regression models were trained only in case 1 and case 2 without variable selection. The training data were used for model selection by 10-fold cross-validation, and the selected model was trained on the full training set. The test data were used at the end to evaluate the chosen model. All six chosen models were compared in the end based on the reduction of the base error rate.

5.1.4.1 Linear Regression

For each of the four cases, we used the one-standard-error rule to choose the subset size k, as the gain of adding new terms to the model was minimal for $k > 4$. Note that selecting the best model instead of the parsimonious model had minimal effects on the reduction of the base error rate. Moreover, the parsimonious models tended to use fewer variables than the best models. The differences between the best models and the parsimonious models are shown in Table 5.4 and provides justification for using the one-standard-error rule in terms of model performance and simplicity.

	Best model		Parsimonious model	
	Subset size k	**Reduction of base error rate**	**Subset size k**	**Reduction of base error rate**
Case 1	6	87.29 %	6	87.29 %
Case 2	5	64.75 %	4	64.92 %
Case 3	10	87.26 %	6	87.29 %
Case 4	16	69.94 %	15	68.84 %

Table 5.4: Differences between parsimonious and best models in the four cases.

The cross-validation training and test errors as a function of subset size k are shown in Figures 5.11 to 5.14. The prediction error of the best model plus its standard error is plotted as horizontal broken lines, while the simplest model below this limit, *i.e.* the parsimonious model is indicated by the vertical broken lines. The results from the chosen linear regression models are summarized in Table 5.5.

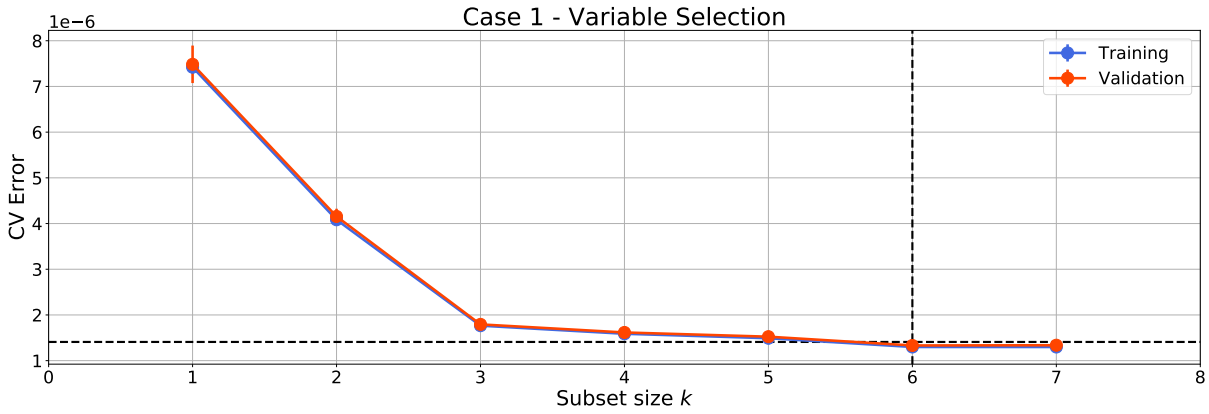

Figure 5.11: Case 1: Cross-validation errors as a function of subset size k with parsimonious model selection.

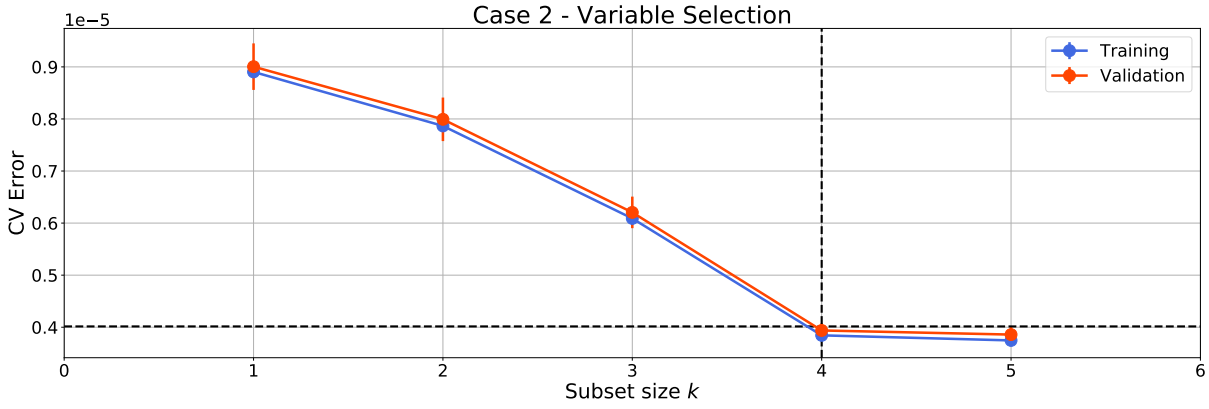

Figure 5.12: Case 2: Cross-validation errors as a function of subset size k with parsimonious model selection.

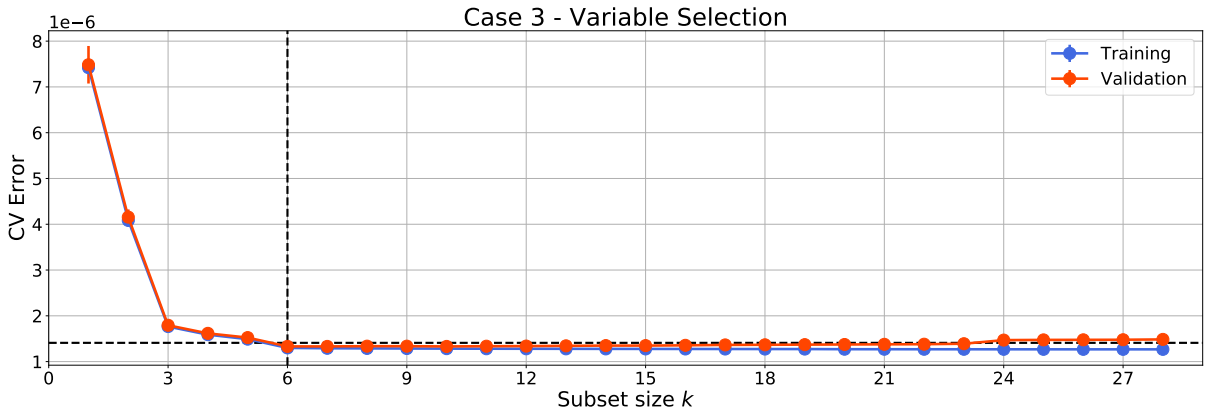

Figure 5.13: Case 3: Cross-validation errors as a function of subset size k with parsimonious model selection.

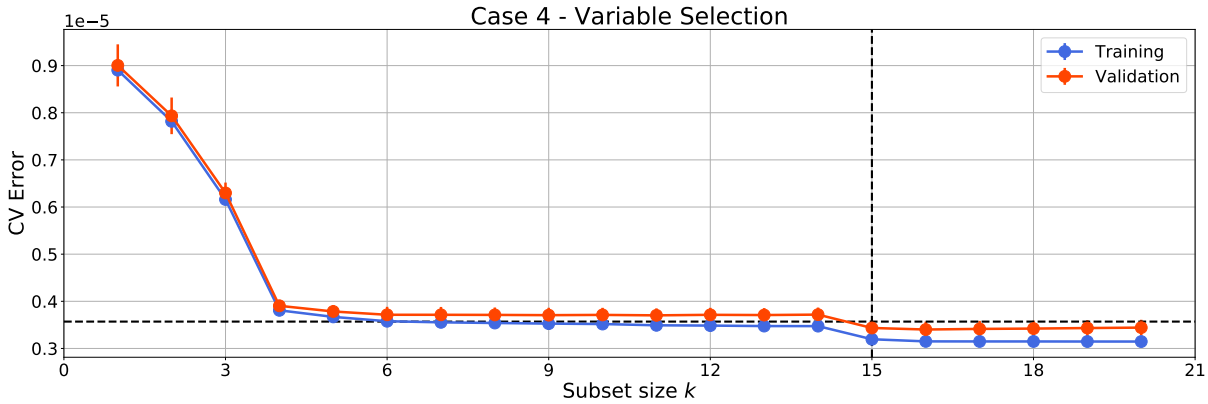

Figure 5.14: Case 4: Cross-validation errors as a function of subset size k with parsimonious model selection.

Case	**Subset size k**	**CV error** $\widehat{\text{Err}} \cdot 10^6$	**Test error** $\text{Err}_{\mathcal{T}} \cdot 10^6$	**Reduction of base error rate**	**Selected variables (in selected order)**
1	6	1.3321 ± 0.0770	1.4333	87.29 %	$H_{1/3}$, T_{atm}, y, P_{atm}, v_{wind}, T_1
2	4	3.9380 ± 0.1381	3.9570	64.92 %	P_{atm}, v_{wind}, T_{atm}, y
3	6	1.3321 ± 0.0770	1.4333	87.29 %	$H_{1/3}$, T_{atm}, y, P_{atm}, v_{wind}, T_1
4	15	3.4344 ± 0.1613	3.5148	68.84 %	See (5.4)

Table 5.5: Summary of the linear regression models selected for each of the four cases.

$$P_{\text{atm}},\ y^4,\ v_{\text{wind}},\ T_{\text{atm}},\ v_{\text{wind}}^2,\ RH,\ P_{\text{atm}}^2,$$
$$T_{\text{atm}}^2,\ T_{\text{atm}}^4,\ RH^2,\ RH^3,\ P_{\text{atm}}^4,\ P_{\text{atm}}^3,\ y,\ y^3 \tag{5.4}$$

The difference between case 1 and 2 and case 3 and 4 was 22.4 % and 18.5 % respectively in terms of the reduction of the base error rate, attributed to the missing ocean-wave data.

Even though polynomial variables were available in case 3, the parsimonious model selection resulted in the same model as in case 1. Furthermore, we see that y, T_{atm}, P_{atm}, and v_{wind} were selected by all models while $H_{1/3}$ was selected first when ocean-wave data were available. Interestingly, P_{atm} was selected first in case 2, despite the low ranking in terms of sensitivity. In case 4 three RH terms were selected, even though it is not a direct input to the true model. We keep in mind that the forward-stepwise selection is a greedy algorithm that does not consider all possible subsets.

For the best model, obtained in case 3, Figure 5.15 shows four diagnostic plots commonly used to investigate the regression results. The true target values were plotted against the predicted values for both the training and test set, as shown in the upper left figure. Points closer to the straight line indicate better predictions.

The upper right figure shows the distribution of the residuals, which are assumed to be Gaussian and additive. However, we can see that the residuals are slightly skewed with a heavy right tail. To graphically check the normality of the residuals normal Q-Q plot was used, as shown in the lower right figure. The residual quantiles were compared against the theoretical quantiles of a normal distribution, illustrated by the red line. The degree of fit between the two distributions is given by the coefficient of determination R^2. In the Q-Q plot, the heavy right-side tail was identified in the upper right corner.

The residuals were also plotted against the predicted values to investigate the structure of the residuals. To check for linearity, additivity, and homoscedasticity, *i.e.* that the variance of the residuals is constant, we checked for a mean residual of zero with equal spreading on either side, which seemed to be the case.

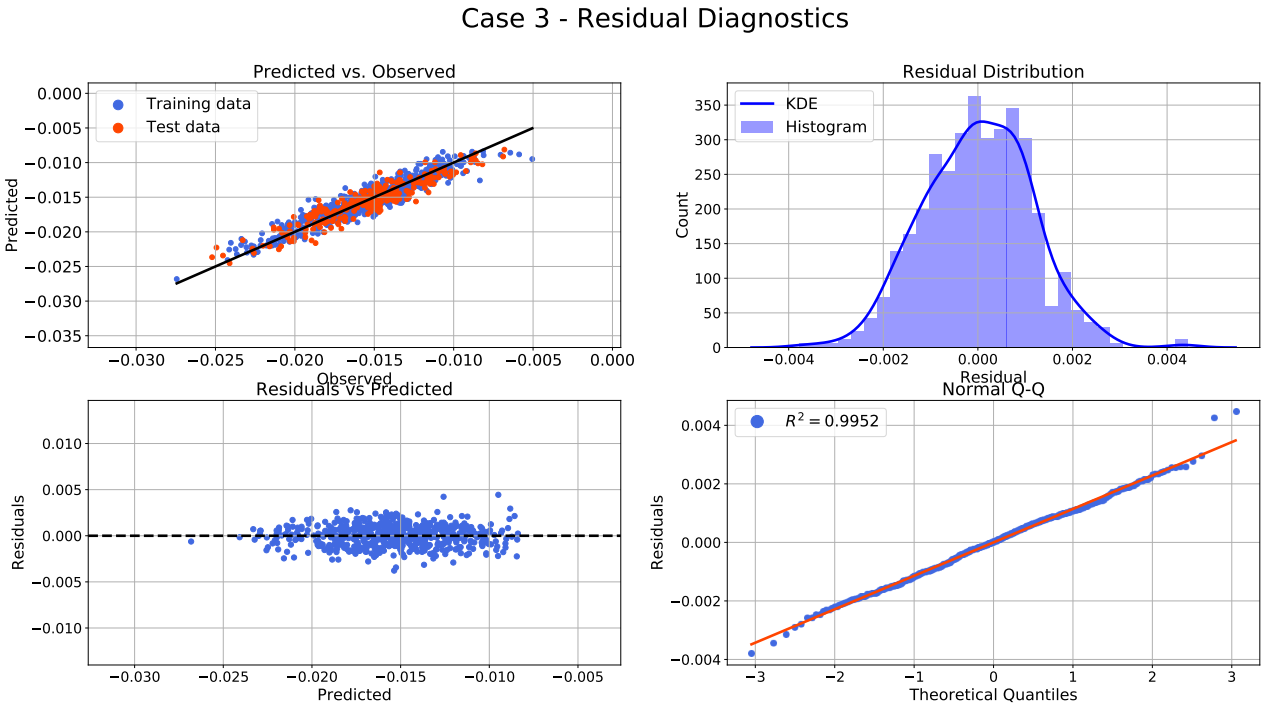

Figure 5.15: Case 3: Residual diagnostic plots.

The best linear model was compared with the original model (5.2). Table 5.6 presents the regression coefficients, their confidence intervals and p-values for the model obtained in case 3. The intercept $\hat{\beta}_0$ represents the expected value of $\widehat{\Delta y}$ when all input variables are set to their means. Thus, the change in cargo level was modeled as a variation around its estimated mean $\widehat{\overline{\Delta y}} = -0.015411$, which was close to the true mean $\overline{\Delta y} = -0.0154$. All coefficients were highly significant at $\alpha = 0.05$ (colored green). $\hat{\beta}_2$ and $\hat{\beta}_5$ were directly comparable to the coefficients for T_{atm} and P_{atm} in (5.2) and showed an error of 1.25 % and 2.8 % respectively.

Variable	**Coefficient**	**Value**	**95 % confidence interval**	**p-value**
-	$\hat{\beta}_0$	-0.015411	$[-0.015502, -0.015319]$	0
$H_{1/3}$	$\hat{\beta}_1$	-0.001316	$[-0.001551, -0.001082]$	$8.27 \cdot 10^{-26}$
T_{atm}	$\hat{\beta}_2$	-0.002025	$[-0.002127, -0.001923]$	$2.20 \cdot 10^{-165}$
y	$\hat{\beta}_3$	-0.001452	$[-0.001550, -0.001354]$	$7.99 \cdot 10^{-117}$
P_{atm}	$\hat{\beta}_4$	0.000514	$[0.000408, 0.000621]$	$4.82 \cdot 10^{-20}$
v_{wind}	$\hat{\beta}_5$	-0.000946	$[-0.001116, -0.000775]$	$2.80 \cdot 10^{-25}$
T_1	$\hat{\beta}_6$	-0.000822	$[-0.000995 - 0.000650]$	$1.60 \cdot 10^{-19}$

Table 5.6: Estimated coefficients with 95 % confidence intervals and p-values for case 3. Significant variables at $\alpha = 0.05$ colored green.

By repeating the same procedure as in Section 5.1.1 we calculated the sensitivity measures for each input variable in the regression model, as shown in Figure 5.16. The results were very similar to that of Figure 5.2, with a small decrease for v_{wind} and a small increase for T_{atm}. The regression model is further compared to the true model by plotting each variable against the computed output while the rest are kept at their mean values, as shown in Figures 5.17 and 5.18. Both models displayed similar plots, except for nonlinearities for large values of $H_{1/3}$ and v_{wind} in the true model. However, the results strongly suggest that the best linear regression model is able to capture the relevant relationships in the data.

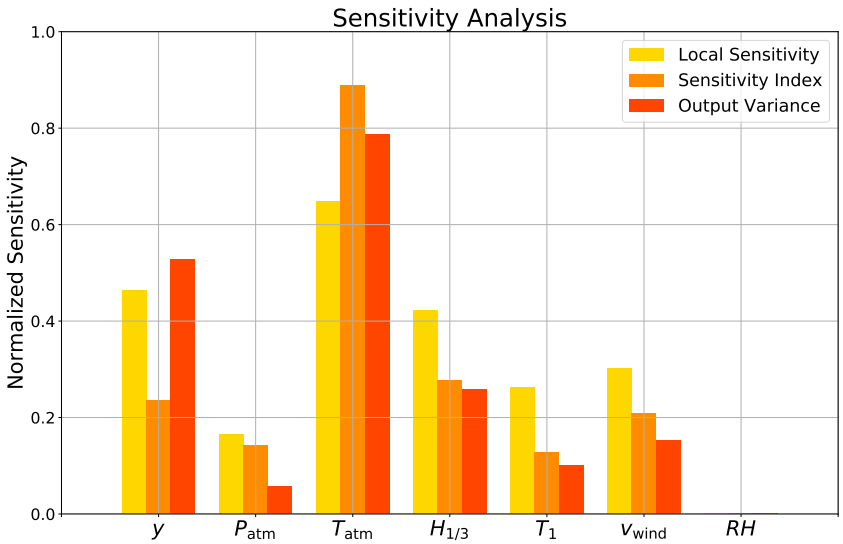

Figure 5.16: Normalized sensitivity measures for each variable in the linear regression model.

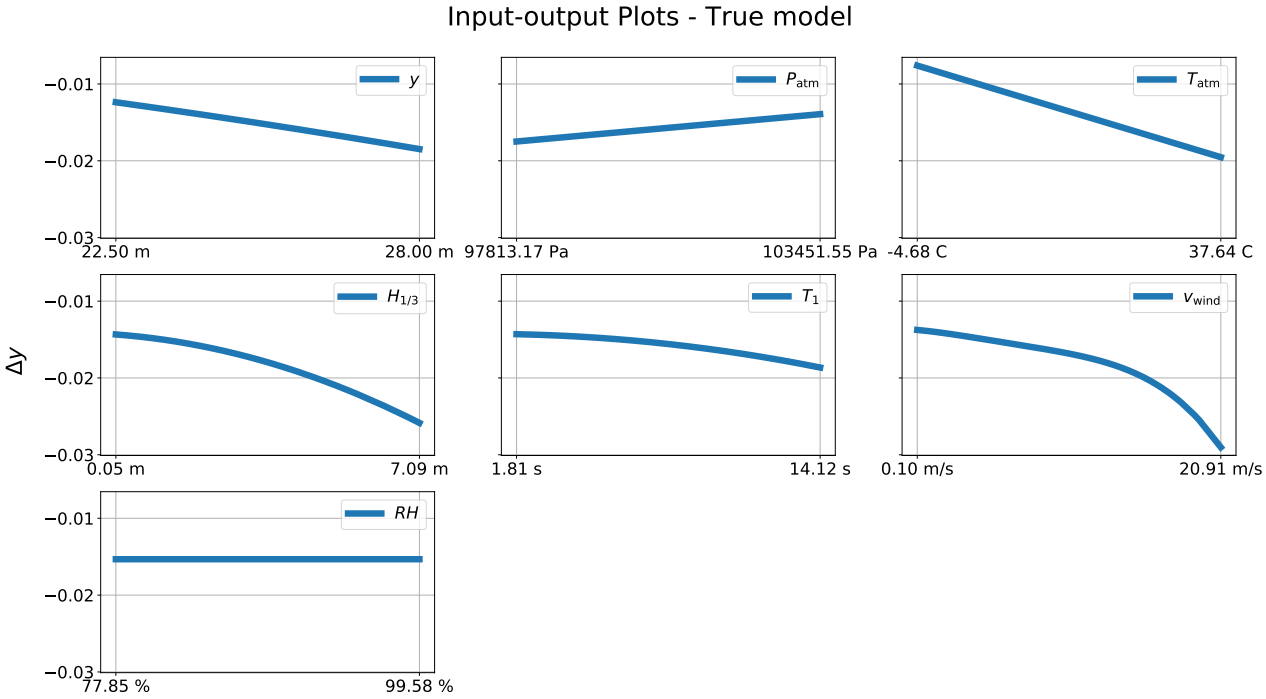

Figure 5.17: Input-output plots for the true model while varying one variable at a time.

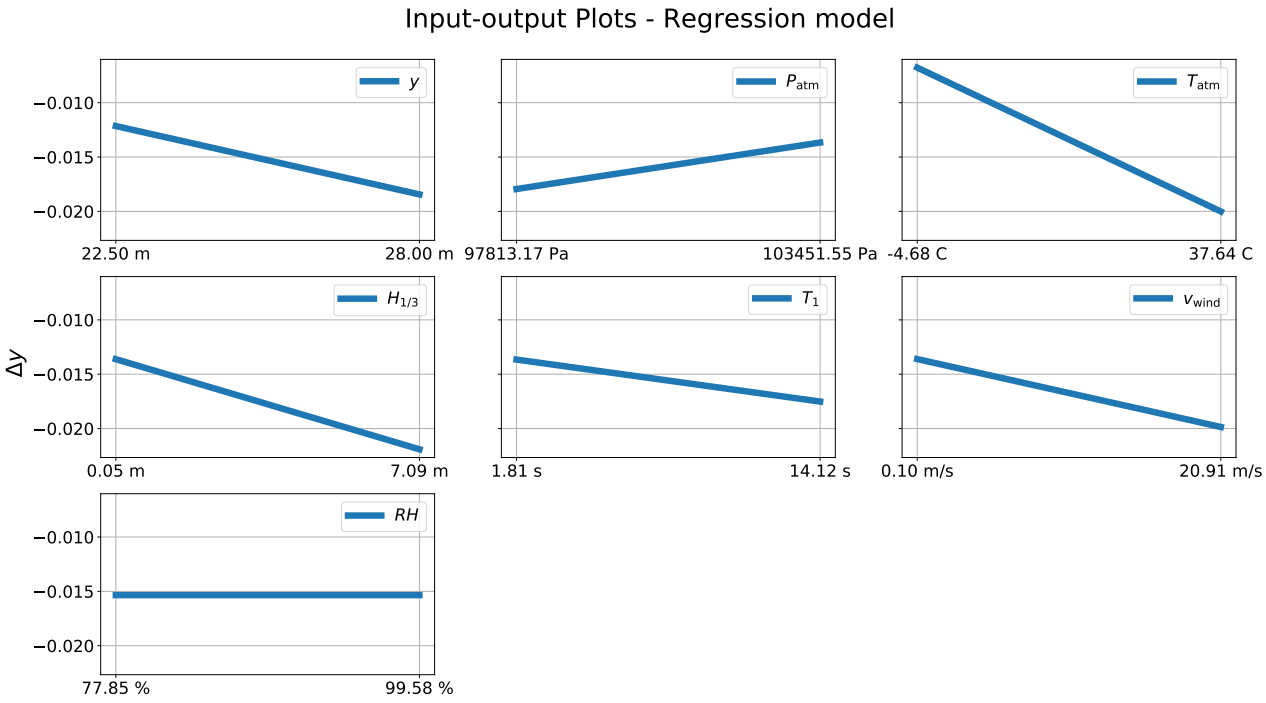

Figure 5.18: Input-output plots for the regression model while varying one variable at a time.

5.1.4.2 Nearest-Neighbors Regression

For the nearest-neighbors regression models only case 1 and case 2 with untransformed inputs were considered. The optimal number of neighbors k was chosen by 10-fold cross-

validation. The cross-validation training and test errors as a function of k are shown in Figure 5.19 and 5.20.

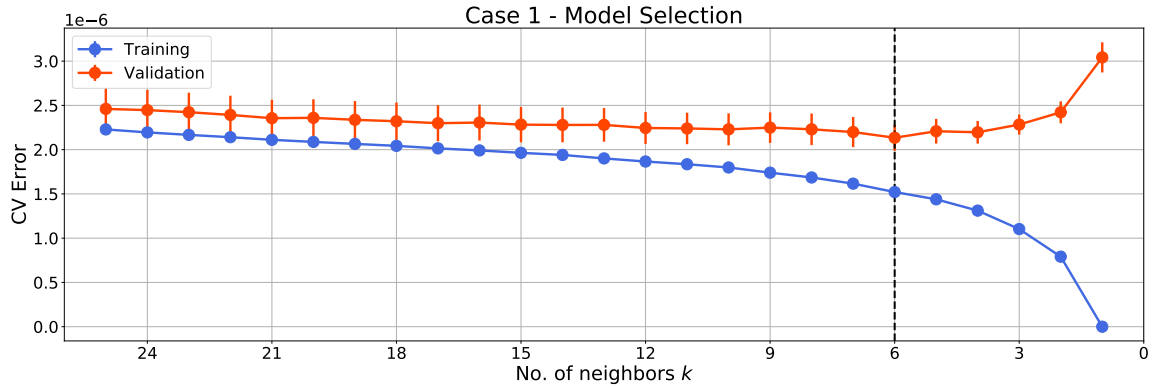

Figure 5.19: Case 1: Cross-validation errors as a function of subset size k with model selection.

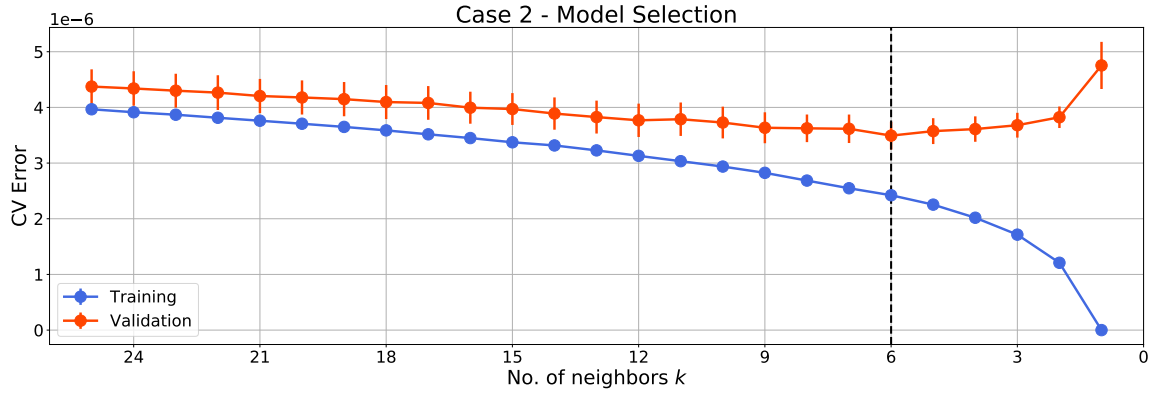

Figure 5.20: Case 2: Cross-validation errors as a function of no. of neighbors k with model selection.

Table 5.7 summarizes the results obtained with the two nearest-neighbors models. Compared to the linear regression models, the nearest-neighbors model was inferior for case 1, and slightly better for case 2, *i.e.* when ocean-wave data were missing. However, the nearest-neighbors models do not offer any mathematical model for further interpretation or analysis. Furthermore, as the true relationship is linear in the simulated data, the nearest-neighbors models were not expected to perform significantly better than the linear ones.

Case	**No. of neighbors** k	**CV error** $\widehat{\mathrm{Err}} \cdot 10^6$	**Test error** $\mathrm{Err}_{\mathcal{T}} \cdot 10^6$	**Reduction of base error rate**
1	6	2.7713 ± 0.3538	2.5416	79.57 %
2	6	3.9973 ± 0.5656	3.9447	68.29 %

Table 5.7: Summary of the nearest-neighbors regression models selected for the first two cases.

5.2 Real-World Data

Using the datasets explained in Section 4.4, the same methodology as presented in Section 5.1 was applied to investigate the relationship between the real cargo level measurements and the ambient conditions. The voyage lengths for each of the 14 voyages for each dataset is summarized in Table 5.8, together with the mean change in cargo level over all voyages. The shortest and longest voyage are colored in red and green respectively, ranging from 4 to 39 days. The cargo level decreases faster in tank 1, as it is smaller than the other three, while the differences between tank 2, 3 and 4 are less clear. One hypothesis is that the steeper change in tank 3 and 4 is due to forced boil-off used for propulsion, as these tanks are located closer to the engine room. Figure 5.21 shows the position of the vessel during the three-year period.

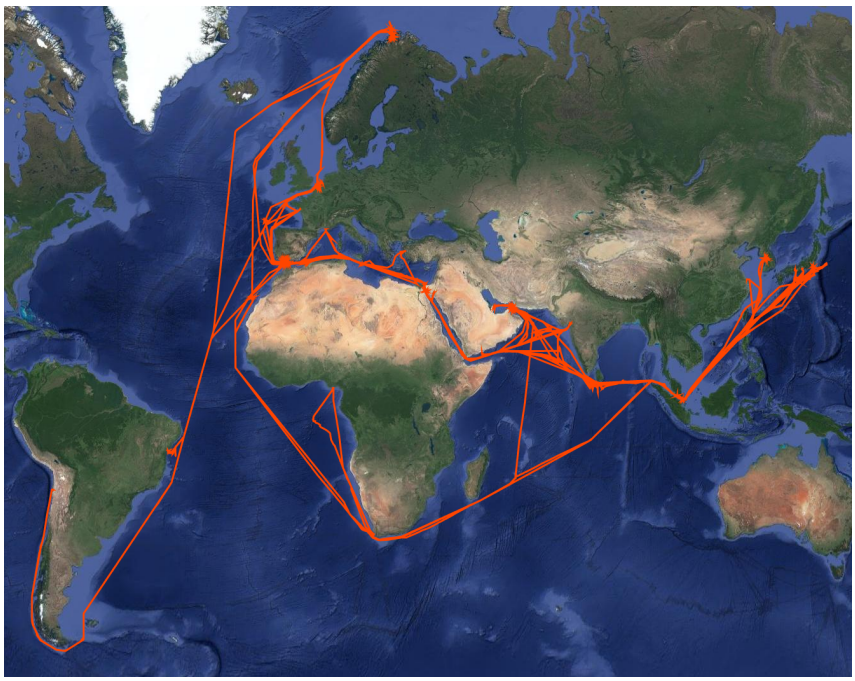

Figure 5.21: Vessel position during the three year period.

Voyage no.	**Voyage length [days]**				
	Tank 1	Tank 2	Tank 3	Tank 4	Combined
1	30.00	30.25	30.00	30.00	29.50
2	21.00	20.75	20.75	20.75	20.75
3	10.25	10.50	10.50	10.25	10.25
4	24.25	24.00	24.00	24.25	23.75
5	7.25	7.00	7.00	7.00	7.00
6	22.00	22.25	22.25	22.25	22.00
7	32.25	32.25	32.25	32.50	32.25
8	24.50	24.25	24.50	24.50	24.25
9	39.25	39.25	39.25	39.00	39.00
10	16.25	16.00	16.00	16.00	16.00
11	12.25	19.00	19.00	19.00	12.25
12	25.25	25.25	25.50	25.25	25.25
13	17.25	17.25	17.25	17.25	17.25
14	4.50	4.75	4.75	4.75	4.50
$\overline{\Delta \mathbf{y}}$ **[m/6h]**	-0.017440	-0.011232	-0.013281	-0.014620	-0.056501

Table 5.8: Summary of voyage lengths and mean change in cargo levels for each dataset. The shortest and longest voyage are colored red and green respectively.

With cargo level measurements from four different tanks, the differences between the individual tanks were investigated. Figure 5.22 shows each Δy plotted together for each voyage. An overall similarity between the Δy was seen, as expected, but also some deviations such as Δy_1 in voyage 11 and Δy_4 in voyage 14. Figure 5.23 provides pairwise plots between the Δy with scatterplots in the upper triangle, univariate distributions on the diagonal and kernel density estimation curves in the lower triangle. Based on the scatterplots and KDE curves we see that they tend to vary with each other, although with a significant spreading. Due to the bimodal nature of the distributions of both Δy_3 and Δy_4, their corresponding scatterplot and KDE curves display some separation between the samples. The cause of the bimodal tendency is unknown, but again it could be due to the forced boil-off mentioned above.

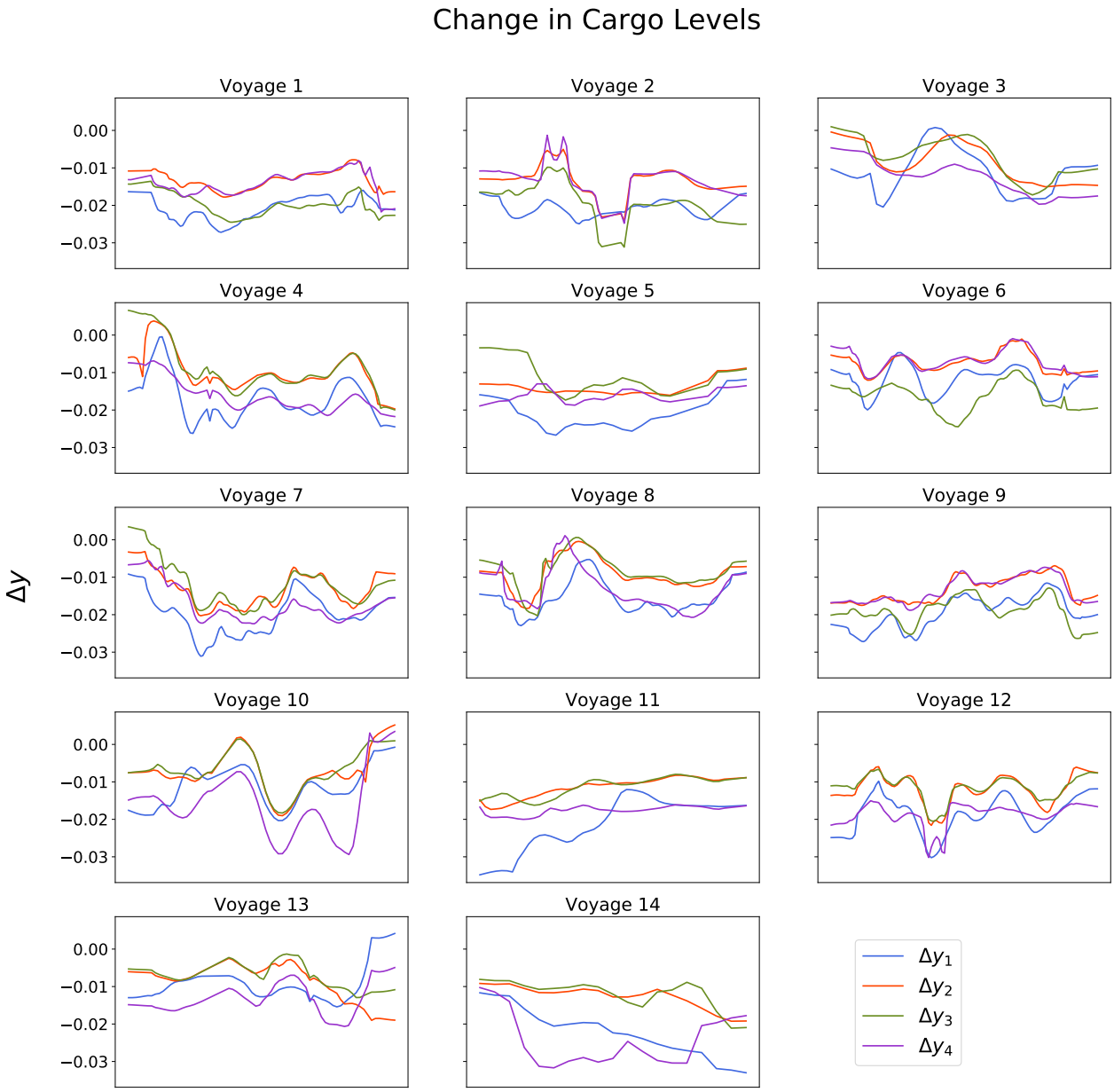

Figure 5.22: Plots of Δy for the individual tanks for each voyage

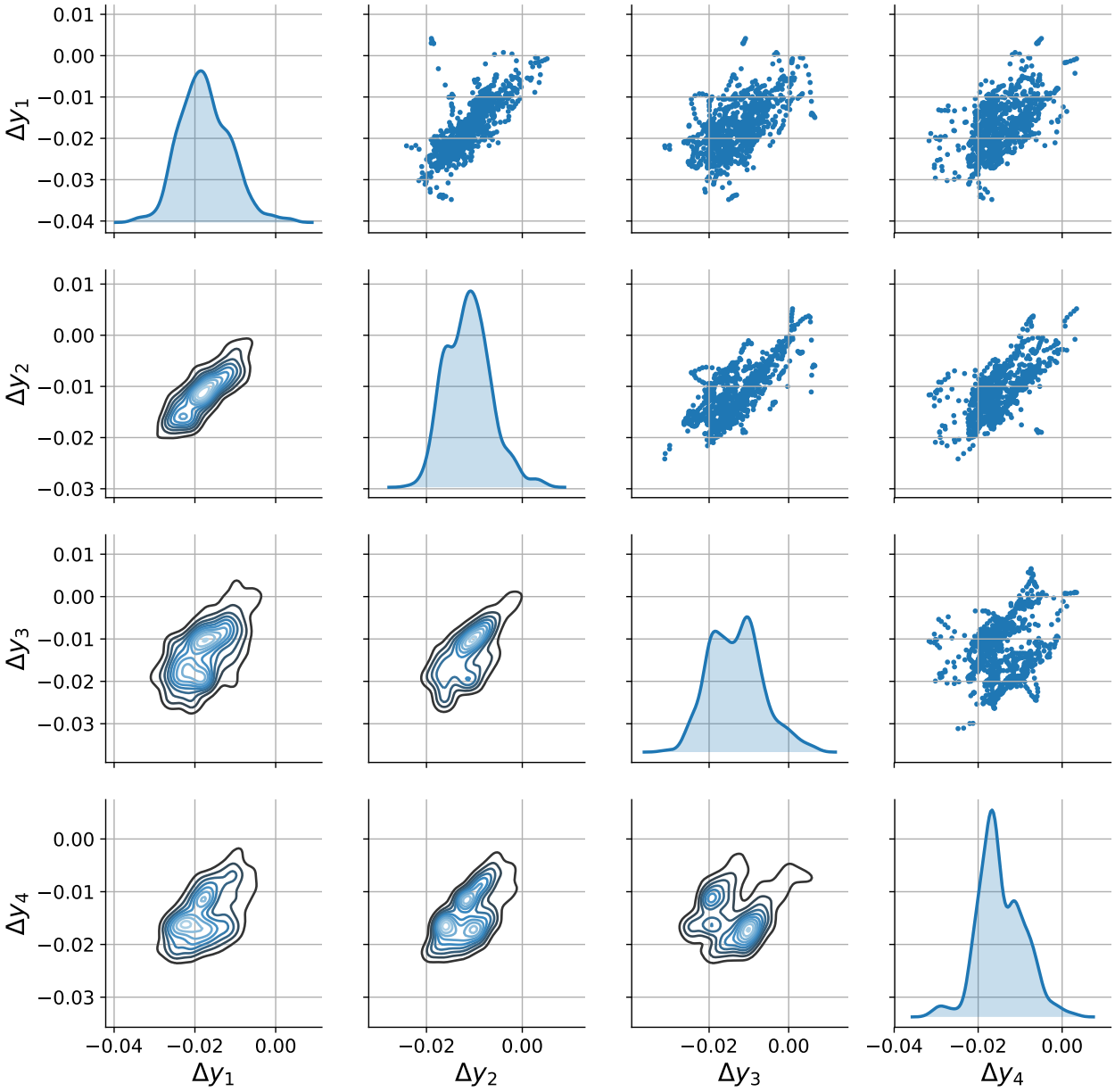

Figure 5.23: Pair plots of Δy for the individual tanks. The lower triangle shows bivariate kernel density estimations, while the upper triangle shows scatterplots. The diagonal shows univariate distributions.

5.2.1 Principal Component Analysis

For the principal component analysis, the cargo level measurements over all four tanks were combined, such that $\Delta y_{\text{tot}} = \Delta y_1 + \Delta y_2 + \Delta y_3 + \Delta y_4$ and $y_{\text{tot}} = y_1 + y_2 + y_3 + y_4$. Thus, the overall effects were investigated rather than differences between the individual tanks. A dataset $\mathbf{X}$ with Δy_{tot}, y_{tot}, and the ambient conditions was standardized and decomposed into its principal components. Figure 5.24 shows the cumulative explained variance of the model

as a function of principal components included. Note that for the six first components the explained variance was lower than for the dataset with the simulated cargo level.

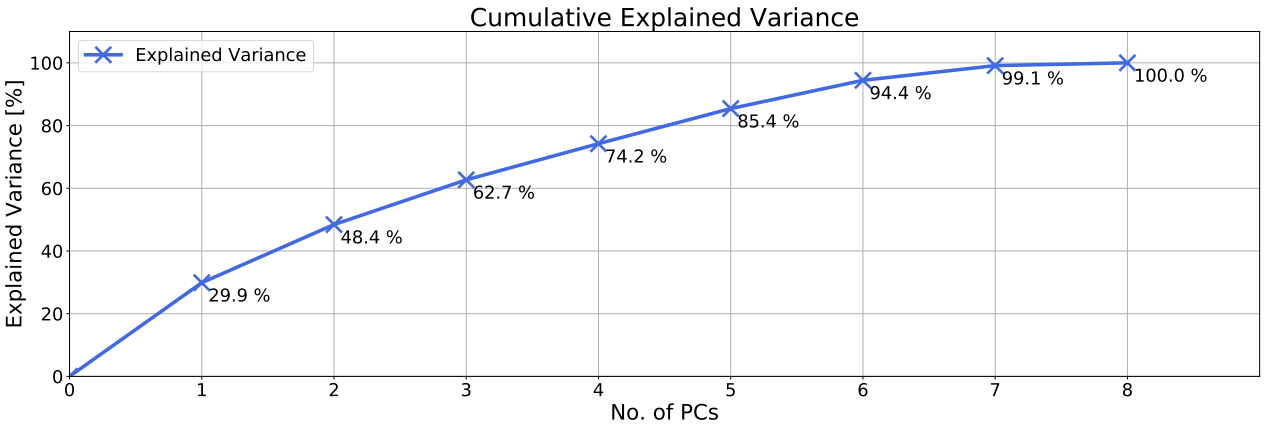

Figure 5.24: Cumulative explained variance as a function of principal components.

Figure 5.25 visualizes the loadings for all variables over all components. In PC 1 we see a correlation between the wave and wind variables, but almost no influence from Δy_{tot} in contrast to the simulated data. In PC 2 we see the same tendencies as in the simulated data, with Δy_{tot} being correlated with P_{atm} and negatively correlated with y_{tot} and T_{atm}. The loadings in the first three components are further visualized in Figure 5.26. Together these components explained 66.1 % of the variance in the data. The results were similar for that of the simulated data, except that PC 2 and 3 are reversed.

The samples were projected onto the three first components, as shown in Figure 5.27 and 5.28. Again we see that samples where Δy is most negative, *i.e.* where the BOR is highest, have a large y and low pressure with either high temperature or rough sea conditions.

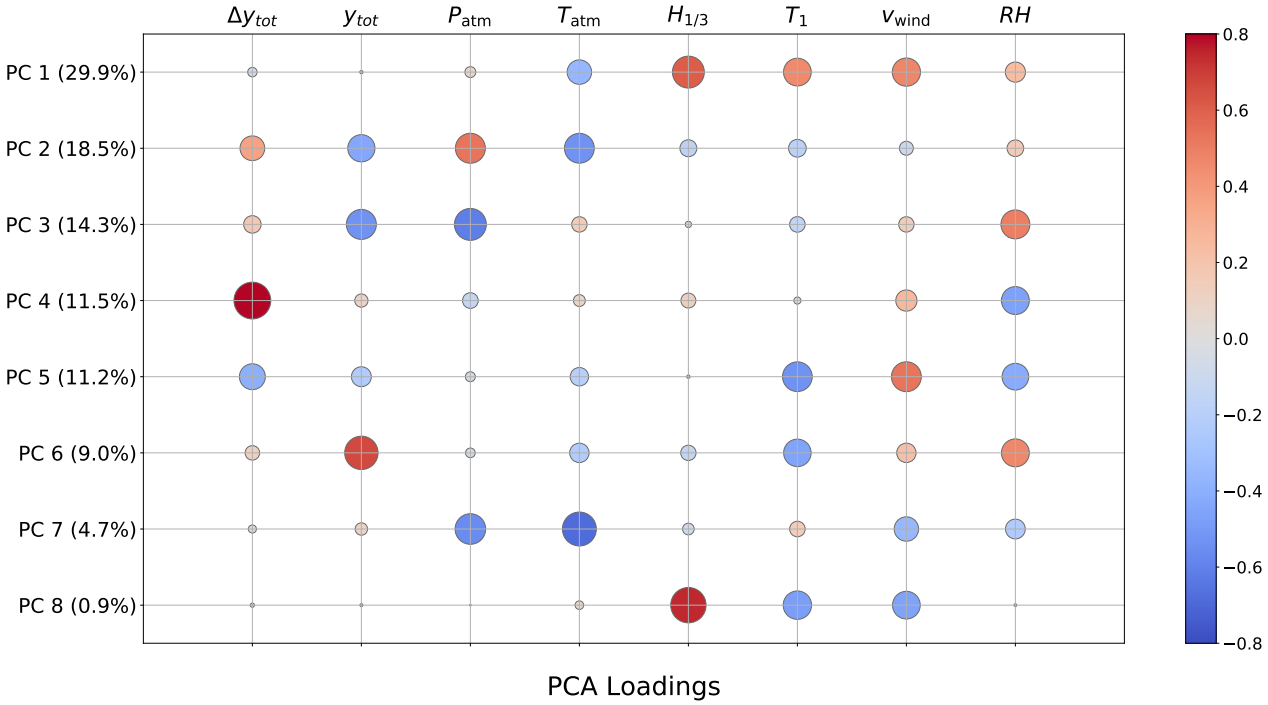

Figure 5.25: PC loadings for all components.

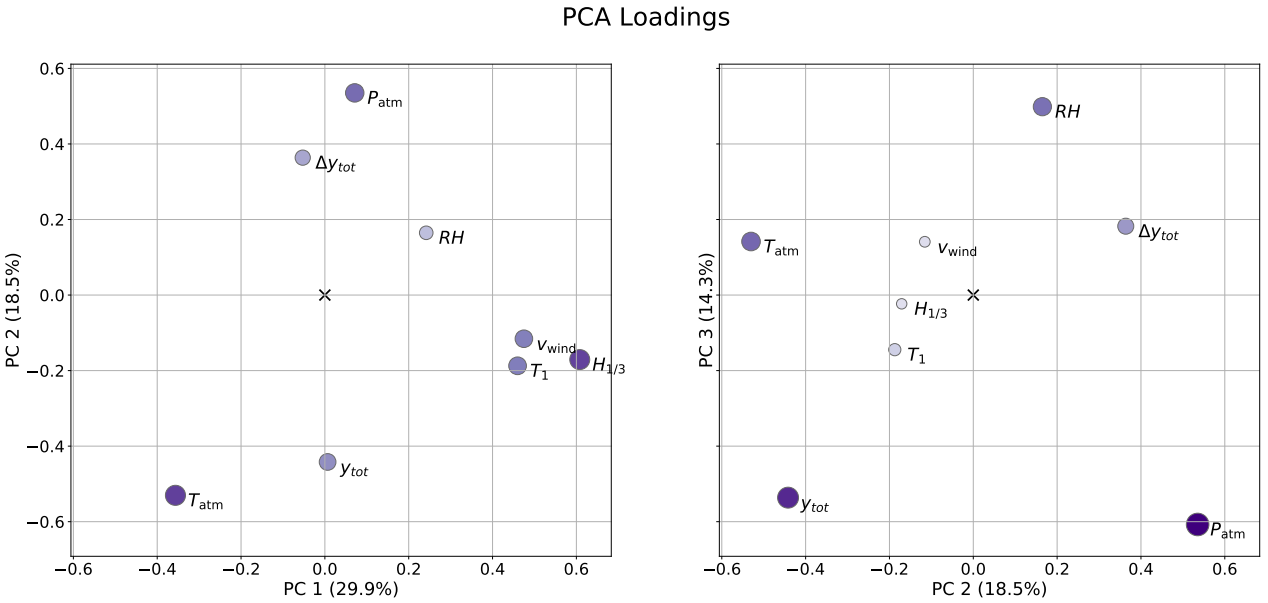

Figure 5.26: PC loadings plot for PC 1, PC 2 and PC 3.

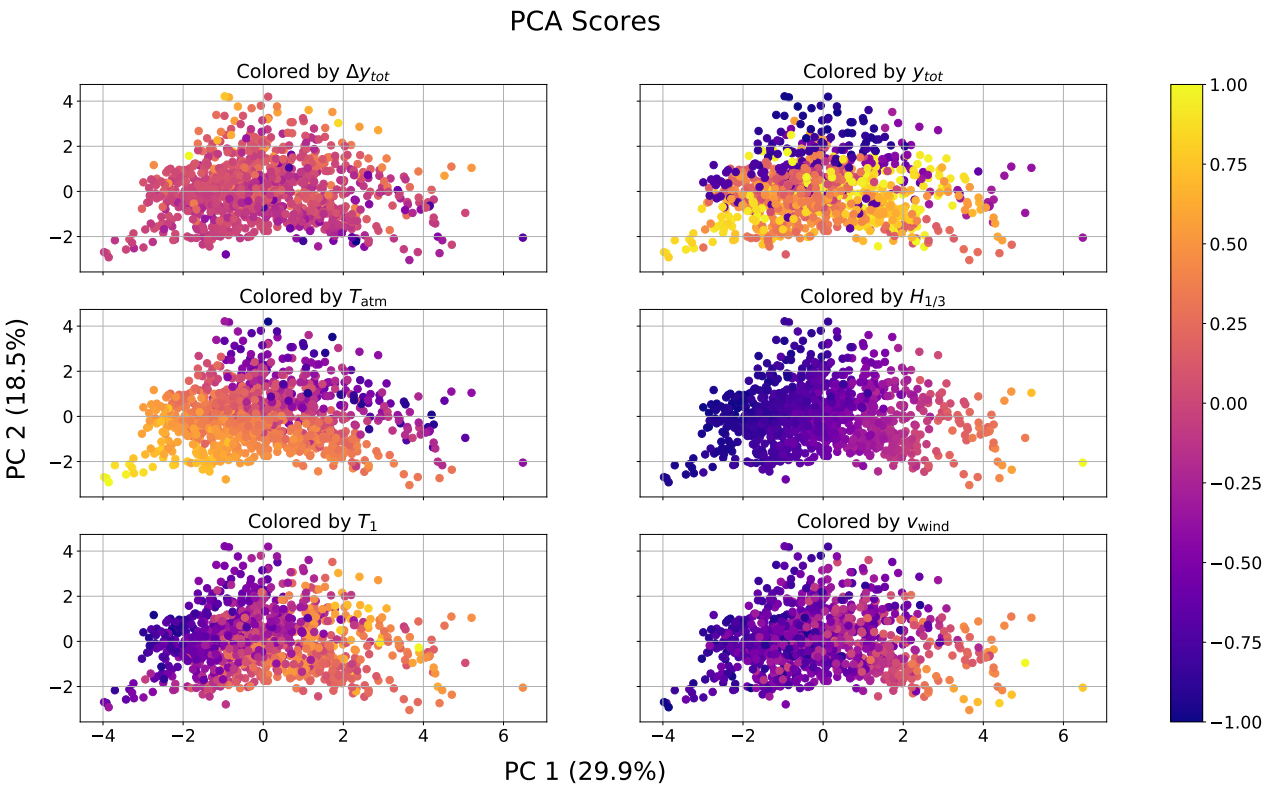

Figure 5.27: PC 1 and PC 2 scores colored by different variables.

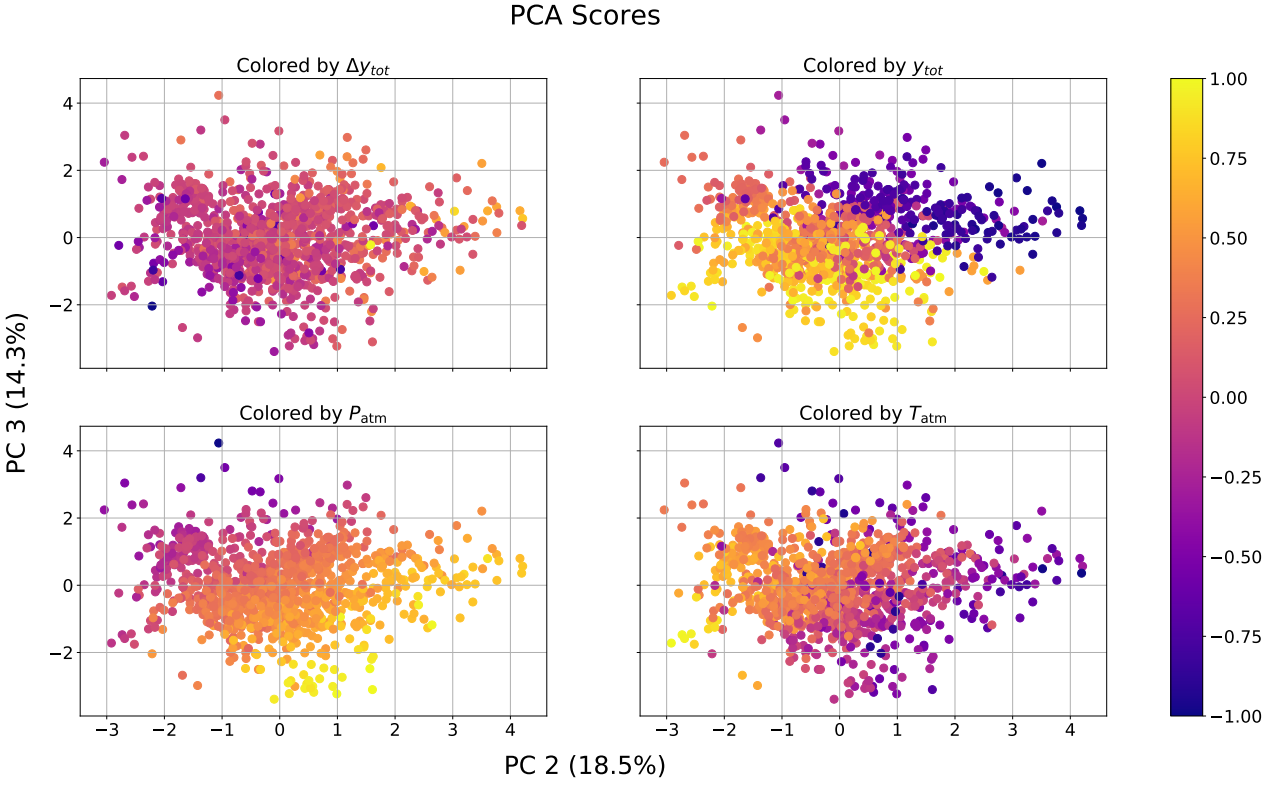

Figure 5.28: PC 2 and PC 3 scores colored by different variables.

5.2.2 Regression Analysis

Regression analyses were performed on five separate datasets: for each individual tank and all tanks combined, *i.e.* with Δy_{tot} and y_{tot}. For each dataset we considered three cases: (1) Linear regression with untransformed input variables without variable selection, termed the *simple model*. (2) Linear regression with transformed input variables and variable selection, termed the *parsimonious model*. (3) Nearest-neighbors regression with untransformed input variables without variable selection. The transformed input variables consisted of (in addition to the untransformed variables) 2nd order, 3rd order, inverse and logarithmic transformations of all variables and cross-term interactions between all ambient conditions. Thus the input set consisted of 7 variables in (1) and (3) and 50 variables in (2).

To limit our scope in this chapter the linear regression models were summarized by the number of regressors and the reduction of the base error rate. Appendix B presents more detailed results for each model. The obtained linear models were analyzed and compared using sensitivity analysis to assess the relative importance of each variable.

All input variables were standardized. The data were shuffled and split into 70 % training data and 30 % independent test data. For (2) and (3) the training data were used for model selection with 10-fold cross-validation. The final models were trained on the full training set. The test data were used at the end to evaluate the each model. All models were compared based on the reduction of the base error rate.

5.2.2.1 Linear Regression

The linear regression models are summarized in Table 5.9, where the simple and parsimonious models are compared for each dataset. The best and worst model of each type is colored green and red respectively. Large differences between the tanks were seen, with specifically poor results for tank 4, but also considerably lower for tank 2. This may indicate that a linear model was not able to capture the relationships between the change in cargo level and the ambient conditions, or that important variables were missing, such as forced boil-off or LNG composition.

Tank no.	Model type	No. of variables	Reduction of base error rate	Note
1	Simple	7	30.69 %	See Appendix B.1.1
	Parsimonious	15	44.66 %	
2	Simple	7	11.99 %	See Appendix B.1.2
	Parsimonious	12	19.75 %	
3	Simple	7	17.81 %	See Appendix B.1.3
	Parsimonious	19	45.20 %	
4	Simple	7	3.10 %	See Appendix B.1.4
	Parsimonious	11	5.89 %	
Combined	Simple	7	20.70 %	See Appendix B.1.5
	Parsimonious	14	30.81 %	

Table 5.9: Regression summary for each individual tank and combined dataset.

As an initial assessment of variable importance, we looked at the variable selections for the parsimonious models, as shown in Figure 5.29. Of the cross-terms, $H_{1/3} \cdot T_1$ was selected in all models, while $T_{atm} \cdot H_{1/3}$ was selected in three models. Of the transformed variables y^{-1}, T_1^3 and $\log(T_1)$ were selected in four models each, while y^3, T_{atm}^3, $H_{1/3}^3$, and $H_{1/3}^{-1}$ were selected in three models each. v_{wind} and RH were least frequently selected.

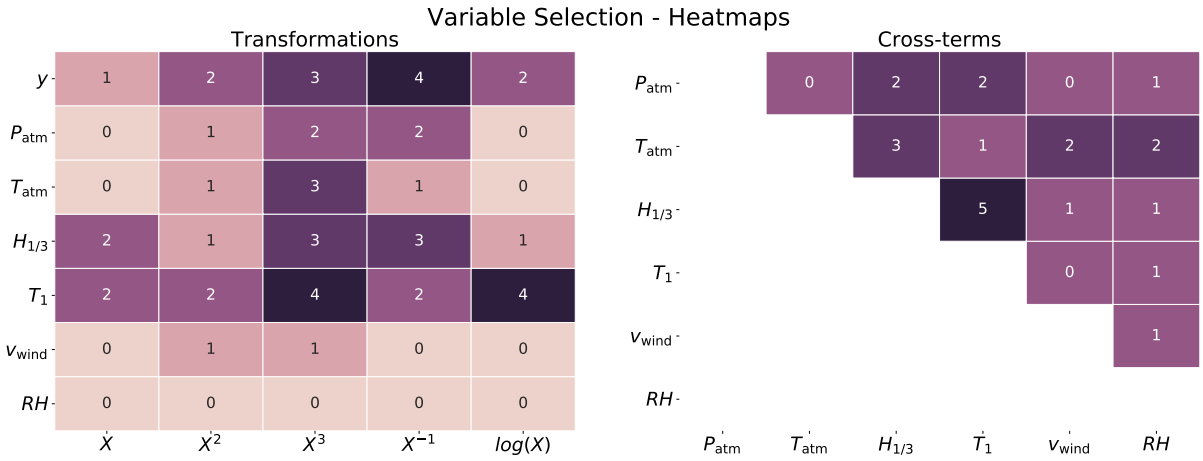

Figure 5.29: Heatmaps of variable selection for all parsimonious models.

Figures 5.30 to 5.34 present the results of the sensitivity analyses for both model type for each dataset. All models except tank 4 and the parsimonious model for tank 2 showed a significant sensitivity in y, indicating that a lot of the variation in Δy can be contributed to y. This was expected due to the geometry of the tanks. Of the ambient conditions, T_{atm}, $H_{1/3}$, and T_1 seemed to dominate, while P_{atm}, v_{wind}, and RH were often ranked low. The results are in

agreement with the heatmaps above.

The three best performing models, the parsimonious models for tank 1, 3 and all tanks combined, were sensitive to y to some extent, while $H_{1/3}$ and T_1 were the dominant ambient conditions. With the exception of all tanks combined, the models were least sensitive to RH. The remaining ambient conditions displayed a similar influence. This suggests that tank sloshing through the action of waves is most important for the overall BOG production on the vessel.

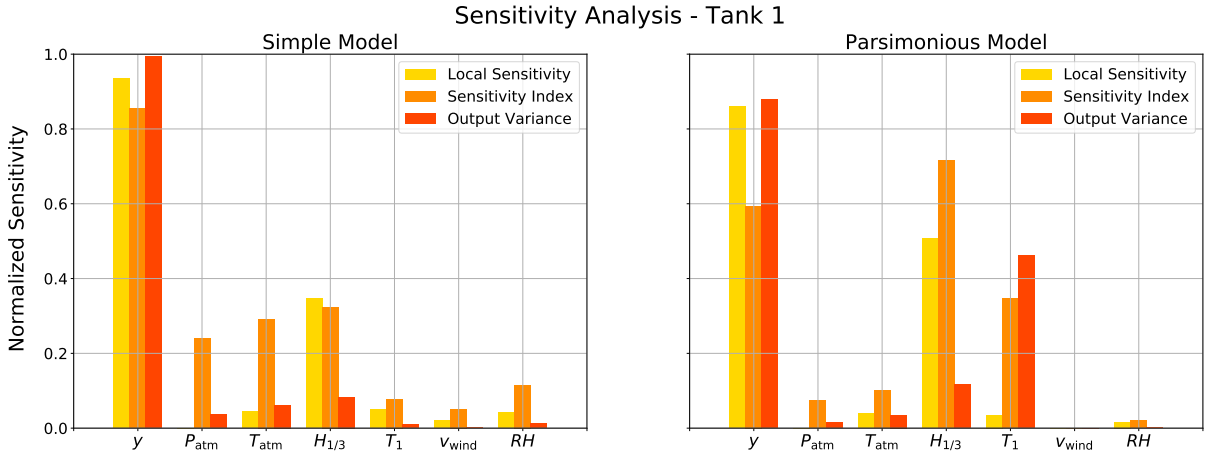

Figure 5.30: Sensitivity analysis of simple and parsimonious model for tank 1.

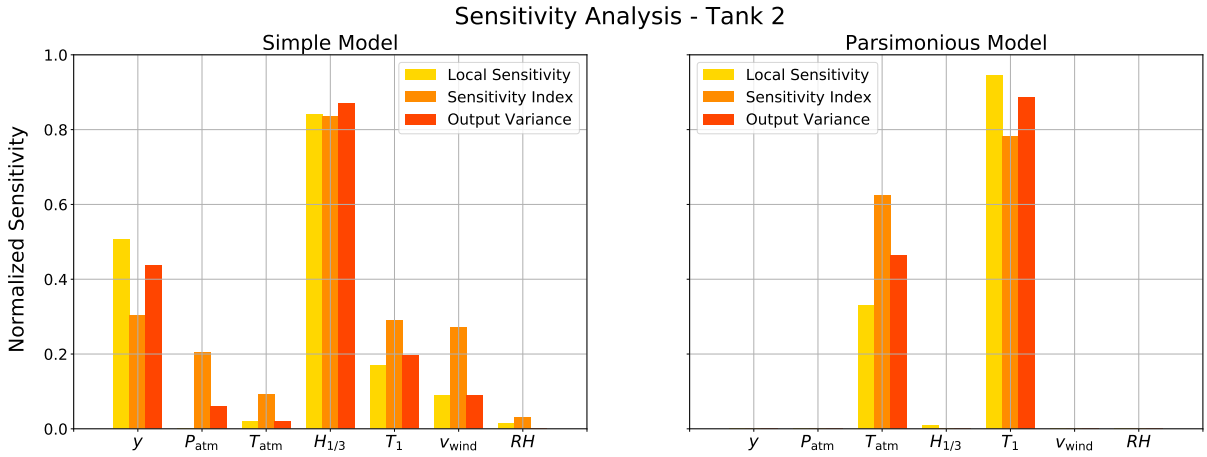

Figure 5.31: Sensitivity analysis of simple and parsimonious model for tank 2.

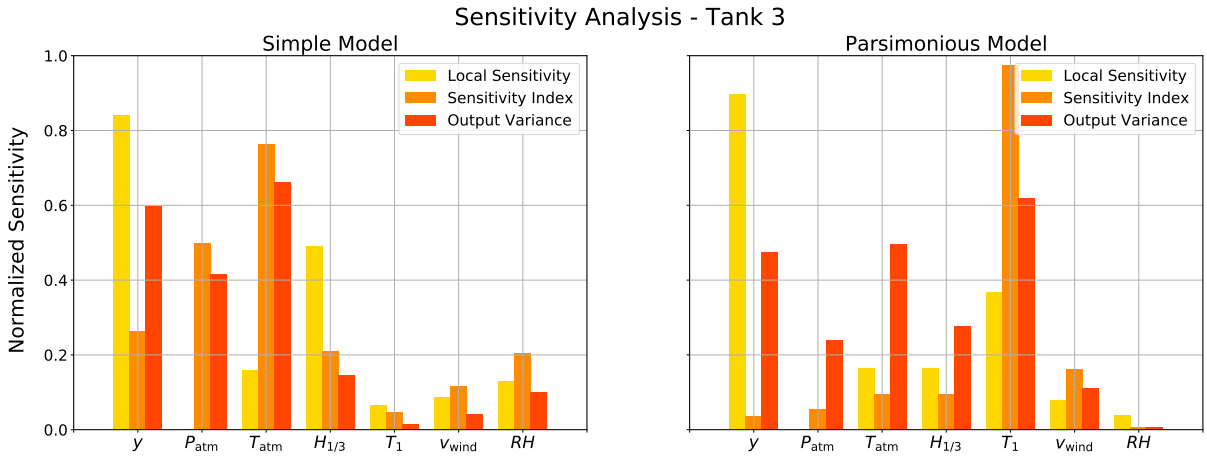

Figure 5.32: Sensitivity analysis of simple and parsimonious model for tank 3.

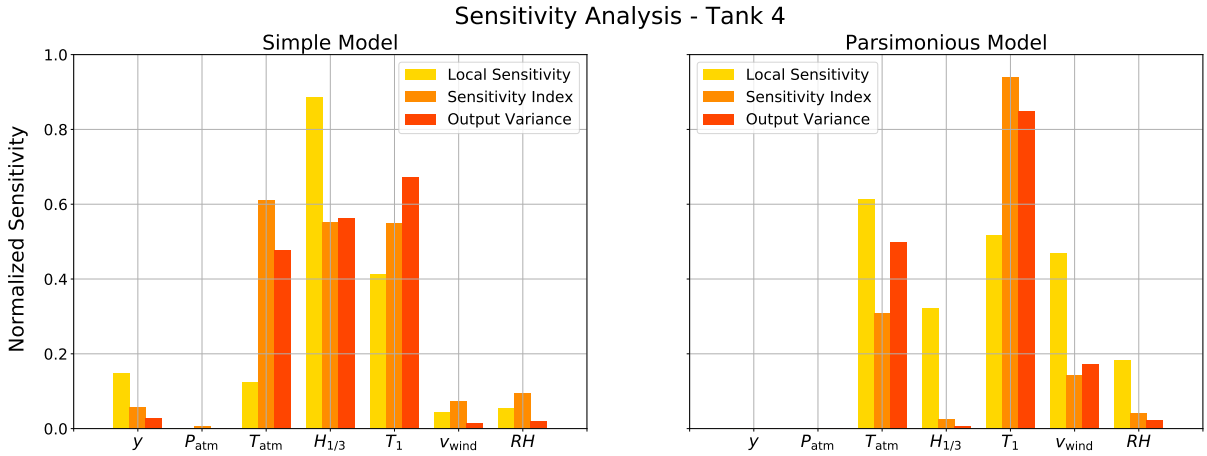

Figure 5.33: Sensitivity analysis of simple and parsimonious model for tank 4.

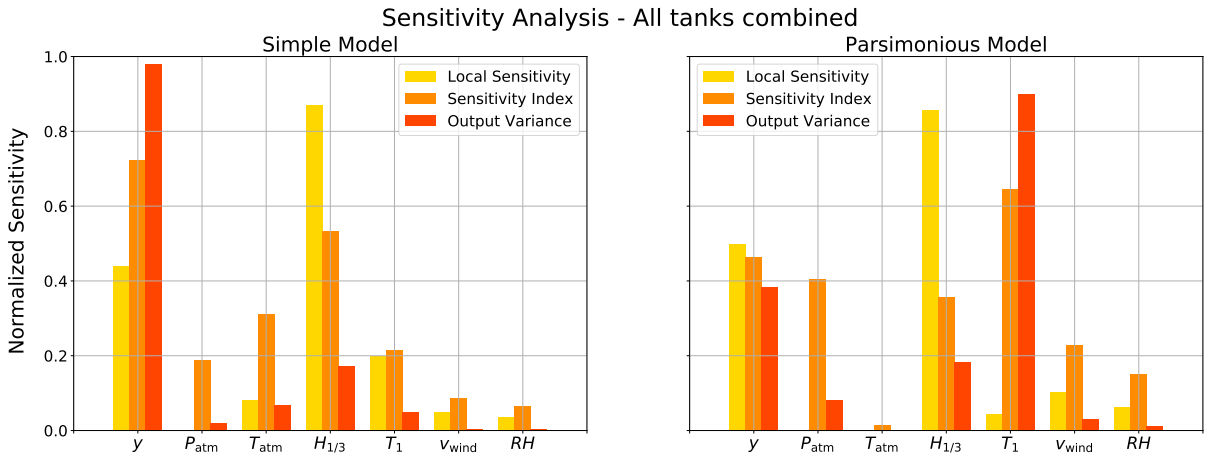

Figure 5.34: Sensitivity analysis of simple and parsimonious model for all tanks combined.

Figure 5.35 shows the input-output plots for the parsimonious model for tank 3 using randomly sampled input data and varying one variable at a time. The boil-off was found to vary

linearly with T_{atm}, as shown by Hasan et al.[3]. It also showed a linear relationship with y, P_{atm}, and RH. Both $H_{1/3}$ and T_1 showed strong nonlinearities. If we look at the joint frequency between $H_{1/3}$ and T_1, shown in Table 5.10, the largest values of $H_{1/3}$ occur for some intermediate value of T_1, before decreasing again for large values of T_1. This explains the shape of T_1 in Figure 5.35. v_{wind} also showed a nonlinear relationship with Δy, but it is not clear why high wind speeds corresponded to a lower BOR, as v_{wind} is strongly correlated with $H_{1/3}$.

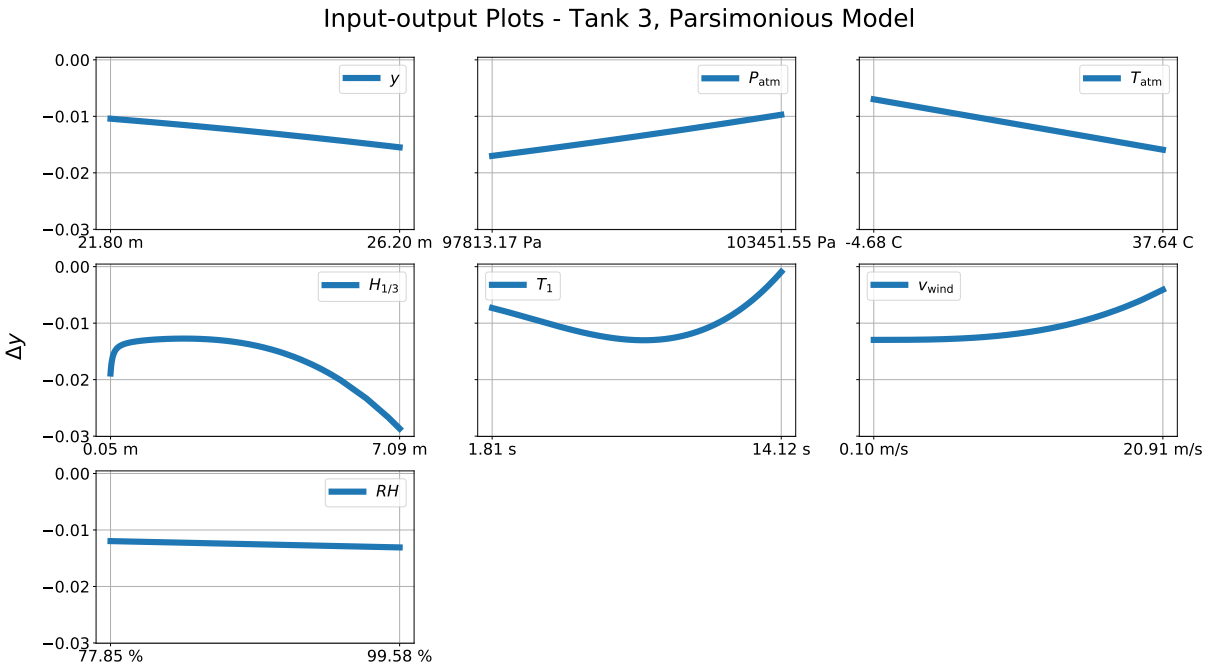

Figure 5.35: Input-output plots for the parsimonious model for tank 3 while varying one variable at a time.

Significant wave height [m]	**Mean wave period [s]**														
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	**Sum**
1	6	134	363	462	260	128	69	68	22	12	4	1	1		1530
2				97	211	248	277	281	208	138	63	16	19	4	1562
3					3	75	140	154	122	134	89	27	5		749
4							24	100	78	42	38	26	10	1	319
5								11	29	24	15	4	1		84
6								1	7	14	6	1		1	31
7										3	5				8
8											3				3
9											1				1
Sum	6	134	363	559	474	451	510	615	466	367	224	76	36	6	4287

Table 5.10: Joint frequency table of $H_{1/3}$ and T_1. The row and columns headers represent the upper limits of the intervals.

5.2.2.2 Nearest-Neighbors Regression

The nearest-neighbors regression models were trained on the untransformed inputs. The optimal number of neighbors k was chosen by 10-fold cross-validation. Table 5.11 summarizes the obtained models for each dataset, with the best and worst model colored green and red respectively. Compared to the linear models the nearest-neighbors models consistently performed better. As before models for tank 2 and 4 showed the poorest results, but with a significant improvement from the linear models (over 40 % higher reduction of the base error rate for tank 4). This may confirm the assumption that the real model is nonlinear in nature.

Tank no.	**No. of neighbors** k	**CV error** $\widehat{\mathrm{Err}} \cdot 10^4$	**Test error** $\mathrm{Err}_{\mathcal{T}} \cdot 10^4$	**Reduction of base error rate**
1	5	0.1322 ± 0.0107	0.1322	68.45 %
2	4	0.1191 ± 0.0076	0.1087	52.6 %
3	2	0.1672 ± 0.0170	0.1710	61.64 %
4	4	0.1657 ± 0.0111	0.1718	48.27 %
Combined	4	1.6486 ± 0.1129	1.4966	62.95 %

Table 5.11: Summary of the nearest-neighbors regression models selected for each dataset.

Chapter 6

Summary and Recommendations for Further Work

This chapter summarizes and discusses the results before providing recommendations for further work. Section 6.1 summarizes the work done, how it was done and clearly states my own contributions. Section 6.2 presents the main results and conclusions obtained throughout the thesis, while Section 6.3 discusses the results and possible applications of the work. Section 6.4 completes the thesis with recommendations for further work.

6.1 Contributions

All the necessary work for this thesis was carried out with scripts written in Python. The necessary extraction, conversion, and handling of data were scripted using mainly Pandas and NumPy. I converted over 3000 CSV-files with more than 1 TB sensory data from the vessel to 10.5 GB of h5 files containing the relevant variables. A separate package called netCDF4 was used for reading the reanalysis data on NetCDF-4 format. The preprocessing of the data was done with Pandas and Scikit-learn, and I implemented nearest-neighbors outlier detection using a nearest-neighbors algorithm from Scikit-learn. LOWESS smoothing was performed with the StatsModels package. All principal component and regression analyses were carried out using Scikit-learn, but my own implementation of linear regression and one by StatsModels were used for comparison. The sensitivity analyses were implemented by me, drawing on SciPy for statistical distributions and random sampling. Furthermore, an application for visualizing the location of the vessel on Google Maps was implemented using the Bokeh package. All plots were made using either Matplotlib or Seaborn. All results were

acquired by writing scripts and analyzed thereafter. I selected all methods, algorithms and their parameters and used example data to gain a full understanding of them.

6.2 Summary and Conclusions

The relationship between ambient conditions and the boil-off of LNG during marine transportation has been investigated using a data-driven approach. The data consisted of sensory data collected from a vessel over a three year period coupled with ambient conditions from an atmospheric reanalysis. As the process of boil-off causes the cargo levels to decrease, the relationship was investigated indirectly through the computed change in cargo level by means of statistical learning methods.

A review of the relevant literature on the boil-off phenomenon was presented in Chapter 2 together with descriptions of the vessel and the different datasets. A review of relevant statistical learning methods and data preprocessing with examples was presented in Chapter 3.

The necessary preprocessing was covered in Chapter 4. Six vessel specific variables measuring cargo levels and atmospheric conditions were successfully coupled with atmospheric reanalysis data using the available AIS data. Consistency was checked for within the reanalysis data and between the datasets. The laden conditions were extracted and cleaned using nearest-neighbors outlier removal, central moving average filtering and LOWESS smoothing to obtain equally spaced and smoothed measurements. The preprocessing reduced the amount of AIS data and cargo level measurements by about 95 %. The changes in cargo level over six hours were computed and combined the ambient conditions. Five datasets were constructed: one for each of the four tanks, and one for all tanks combined. The datasets spanned 14 distinct voyages, containing 9 variables and about 1100 data points.

In Section 5.1 a crude polynomial model was used to simulate a virtual cargo level using the real ambient conditions. The simulated data were inspected in the unsupervised framework using PCA to verify the relationships in the model. Both linear and nearest-neighbors regression models were trained on a portion of the data and assessed on an independent test set. Model selection was performed with 10-fold cross-validation. The best linear model reduced the base error rate with 87.29 % and successfully uncovered the important relationships in

the model, but without accounting for higher order terms. Sensitivity analysis was used to compare the best regression model against the true model, and effectively showed strong similarities between the models. The nearest-neighbors models did not outperform the linear ones, as the true model was linear.

The same methodology was employed on the real-world data in Section 5.2. Compared to the simulated data, inspections with PCA revealed lower explained variance in the real-world data and smaller contributions from Δy in the first components. However, the results were overall similar. Linear and nearest-neighbors regression were performed on all five datasets. Simple linear models using untransformed input variables without variable selection were constructed and compared to parsimonious models using transformed input variables and variable selection. Large differences between the tanks were uncovered, with 45.20 % reduction of the base error rate for tank 3 and only 5.89 % for tank 4 using the parsimonious models. The nearest-neighbors models outperformed the linear models overall, with a reduction of the base error rate varying from 48.27 % for tank 4 to 68.45 % for tank 1. Overall the ambient conditions displayed predictive capabilities on boil-off.

The variable selection during model training showed that cargo levels, waves, and ambient temperature were most frequently selected. To further assess the relative importance sensitivity analyses were performed on the obtained linear models. The cargo level was found to be overall important in tank 1 and 3 while contributing less in explaining the variation in tank 2 and 4. Waves and temperature were consistently the highest ranking ambient conditions. Of these two, the results suggested that sloshing, either through $H_{1/3}$ or T_1 was more important than the temperature differential.

6.3 Discussion

The data-driven methodology employed proved effective for the simulated data when the true model was linear and the signal-to-noise ratio reasonably low. However, the models trained on the real-world data displayed poorer performance and the nearest-neighbors models outperformed the linear ones. This could indicate that a linear model of the true relationship is an oversimplification. Large differences were uncovered between the tanks but remain unexplained.

Variable selection was performed using a greedy sub-optimal algorithm, not taking into account all possible subsets. By using an algorithm for optimal subset selection, more certain results can be obtained. Sensitivity analyses were performed while varying one variable at a time to assess the importance of the variables. This method does not take into account variable correlations or interactions but only look at one variable at a time. This is a limitation, as we know that the ambient conditions are correlated while interaction terms could be important. Moreover, only a portion of the full input space is sampled using this method. However, the results gave good indications about variable sensitivity and importance.

The results presented were found to be in agreement with the literature. Waves and ambient temperature were found to be the most important factors, in agreement with Dobrota et al. [5] and Hasan et al. [3]. A linear relationship between T_{atm} and boil-off was uncovered, in agreement with Hasan et al. [3].

Referring to the limitations listed in Section 1.3, only one vessel was considered in this thesis. To account for bias, data from several vessels should be analyzed. Moreover, with only 14 voyages during the three-year period and a limited temporal resolution of the reanalysis data, the size of the training data was constrained. More voyages would provide a richer dataset and possibly reduce voyage specific bias. As boil-off was investigated indirectly through the change of cargo levels, and tank sloshing indirectly through wave height and wave period, the results are prone to uncertainty. By using direct measurements of the boil-off amount and vessel motions, the same relationships could be uncovered to strengthen the results.

By investigating the ambient conditions it is clear that rough sea conditions are correlated with lower temperatures and the other way around. As such, there seems to be a tradeoff between ambient temperature and waves in terms of boil-off. One possible application of the type of models presented here would be to incorporate forecast weather data in path planning to optimize boil-off losses. This would require a significant amount of training data beforehand, and would only be viable at most one week ahead due to uncertainties in the forecasts. Nevertheless, as more than 1.2 billion USD are lost due to boil-off each year, it could prove useful, especially for a future fleet of autonomous LNG tankers.

6.4 Recommendations for Further Work

For further work, I will differentiate between (1) recommendations with regards to the data, and (2) recommendations with regards to the methodology.

For (1) I would recommend to

(a) Use data gathered from more than one vessel to varify the relationships across vessels.

(b) Use data of operating pressure and nitrogen content in the tanks, as these are known to have a significant effect on boil-off.

(c) Use data on vessel motions to assess directly the importance of tank sloshing on boil-off.

(d) Use data that directly measures the amount of BOG. This would remove the dependency on cargo level as one could use the BOR directly.

For (2) I would recommend to

(a) Implement the bound and branch algorithm for variable selection to ensure that the optimal subset of variables is selected without the need for exhaustive computation.

(b) Perform global sensitivity analyses of the obtained models using Monte Carlo simulation, as this will sample the full input space and account for higher order interactions.

(c) Assess the performance of other nonlinear regression models such as support vector machines or neural nets.

For long-term recommendations, it would be interesting to look into the possibilities of using weather forecast data in path planning for optimizing of boil-off losses.

Bibliography

[1] IEA. *CO2 Emissions From Fuel Combustion Highlights 2016*, 2016. https://www.iea.org/publications/freepublications/publication/co2-emissions-from-fuel-combustion-highlights-2016.html [Accessed: 2017-03-17].

[2] IGU. *2016 World LNG Report*, 2016. http://www.igu.org/publications/2016-world-lng-report [Accessed: 2017-03-17].

[3] MM Faruque Hasan, Alfred Minghan Zheng, and I. A. Karimi. Minimizing boil-off losses in liquefied natural gas transportation. *Industrial & engineering chemistry research*, 48 (21):9571–9580, 2009.

[4] Bluegold Research. *Global LNG Prices*, 2017. https://bluegoldresearch.com/global-lng-prices [Accessed: 2017-06-01].

[5] Đorđe Dobrota, Branko Lalić, and Ivan Komar. Problem of boil-off in LNG supply chain. *Transactions on Maritime Science*, 2(02):91–100, 2013.

[6] Angel Benito and S. A. Enagás. Accurate determination of LNG quality unloaded in Receiving Terminals: An Innovative Approach. In *24th World Gas Conference*, 2009.

[7] IMO. *Amendments to the International Code for the Construction and Equipment of Ships Carrying Liquefied Gases in Bulk (IGC Code)*, 2014. http://www.imo.org/en/KnowledgeCentre/IndexofIMOResolutions/Maritime-Safety-Committee-(MSC)/Documents/MSC.370(93).pdf [Accessed: 2017-03-19].

[8] Francis Brown. Auto-refrigeration: When bad things happen to good pressure vessels., 2002. *The National Board of Boiler and Pressure Vessel Inspectors Bulletin*, 57(3):4-5, 2002.

[9] The Engineering Toolbox. *Boiling Points for common Liquids and Gases*, 2017. http://www.engineeringtoolbox.com/boiling-points-fluids-gases-d_155.html [Accessed: 2017-06-09].

[10] The Engineering Toolbox. *Fuel Gases Heating Values*, 2017. http://www.engineeringtoolbox.com/heating-values-fuel-gases-d_823.html [Accessed: 2017-06-09].

[11] The Engineering Toolbox. *Molecular Weight - Gases and Vapors*, 2017. http://www.engineeringtoolbox.com/molecular-weight-gas-vapor-d_1156.html [Accessed: 2017-06-09].

[12] GTT. *NO96 System*, 2016. http://www.gtt.fr/en/technologies-services/our-technologies/no96 [Accessed: 2017-03-17].

[13] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597, April 2011. ISSN 1477-870X. doi: 10.1002/qj.828.

[14] D.P. Dee, J. Fasullo, D. Shea, and J. Walsh. *The Climate Data Guide: Atmospheric Reanalysis: Overview & Comparison Tables*, 2016. https://climatedataguide.ucar.edu/climate-data/atmospheric-reanalysis-overview-comparison-tables [Accessed: 2017-02-07].

[15] ECMWF. *Data Assimilation*, 2017. http://www.ecmwf.int/en/research/data-assimilation [Accessed: 2017-02-07].

[16] P. Berrisford, D.P. Dee, P. Poli, R. Brugge, K. Fielding, M. Fuentes, P.W. Kållberg, S. Kobayashi, S. Uppala, and A. Simmons. The era-interim archive version 2.0. Shinfield Park, Reading, November 2011.

[17] Python Foundation. *About Python*, 2017. https://www.python.org/about/ [Accessed: 2017-04-28].

[18] Pandas. *Pandas - Python Data Analysis Library*, 2017. http://pandas.pydata.org/ [Accessed: 2017-04-28].

[19] NumPy. *NumPy- Numerical Python*, 2017. http://www.numpy.org/ [Accessed: 2017-04-28].

[20] SymPy. *SymPy - Symbolic Python*, 2017. http://www.sympy.org/en/index.html [Accessed: 2017-05-26].

[21] SciPy. *SciPy - Scientific Computing Tools for Python*, 2017. https://www.scipy.org/about.html [Accessed: 2017-04-28].

[22] SciKit-Learn. *SciKit-Learn: Machine Learning in Python*, 2017. http://scikit-learn.org/stable/ [Accessed: 2017-04-28].

[23] Matplotlib. *Matplotlib 1.5.3 Documentation*, 2017. http://matplotlib.org/ [Accessed: 2017-04-28].

[24] Spyder. *Spyder IDE Wiki*, 2017. https://github.com/spyder-ide/spyder/wiki [Accessed: 2017-04-28].

[25] Ajay Malik. *How Python Is Different From Other Languages*, 2016. http://www.c-sharpcorner.com/article/how-python-is-different-from-other-languages/ [Accessed: 2017-04-28].

[26] Karlijn Willems. *Choosing R or Python for data analysis? An infographic*, 2015. https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis#gs.tnzXtZc [Accessed: 2017-04-28].

[27] Graham Williams. *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer Science & Business Media, 2011.

[28] Erling Paulsen, Christian De Jonge, and Anders Gravdal. Multivariate Analysis of Ship Data. Project Thesis, NTNU, December 2016.

[29] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. Springer series in statistics. Springer, New York, 2001. ISBN 978-0-387-95284-0.

[30] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[31] Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, and Jörg Stork. Comparison of different Methods for Univariate Time Series Imputation in R. *arXiv preprint arXiv:1510.03924*, 2015.

[32] Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.

[33] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 427–438, New York, NY, USA, 2000. ACM. ISBN 978-1-58113-217-5. doi: 10.1145/342009.335437.

[34] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[35] Jerome H. Friedman. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, March 1997. ISSN 1384-5810, 1573-756X. doi: 10.1023/A:1009778005914.

[36] Scott Fortmann-Roe. *Understanding the Bias-Variance Tradeoff*, 2012. http://scott.fortmann-roe.com/docs/BiasVariance.html [Accessed: 2017-05-02].

[37] Leo Breiman. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[38] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.

[39] Leo Breiman and Philip Spector. Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique*, 60(3):291–319, 1992. ISSN 0306-7734. doi: 10.2307/1403680.

[40] William Briggs. *Example of how easy it is to mislead yourself: stepwise regression*, 2008. http://wmbriggs.com/post/92/ [Accessed: 2017-05-11].

[41] Peter L. Flom and David L. Cassell. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. *NorthEast SAS Users Group (NESUG): Statistics and Data Analysis*, 2007.

[42] Patrenahalli M. Narendra and Keinosuke Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 26(9):917–922, 1977.

[43] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, November 1901. ISSN 1941-5982, 1941-5990. doi: 10.1080/14786440109462720.

[44] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[45] Gene H. Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420, 1970.

[46] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

[47] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

[48] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[49] Vaisala. *Humidity Conversion Formulas: Calculation Formulas for Humidity*, 2013. http://www.vaisala.com/Vaisala%20Documents/Application%20notes/Humidity_Conversion_Formulas_B210973EN-F.pdf [Accessed: 2017-03-17].

[50] Willard J. Pierson and Lionel Moskowitz. A proposed spectral form for fully developed wind seas based on the similarity theory of S. A. Kitaigorodskii. *Journal of Geophysical Research*, 69(24):5181–5190, December 1964. ISSN 2156-2202. doi: 10.1029/JZ069i024p05181.

[51] D. M. Hamby. A review of techniques for parameter sensitivity analysis of environmental models. *Environmental monitoring and assessment*, 32(2):135–154, 1994.

[52] Ilya M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.

[53] Jérôme Morio. Global and local sensitivity analysis methods for a physical system. *European Journal of Physics*, 32(6):1577, 2011.

Appendix A

Sensitivity Analysis

This appendix provides some simple methods to assess the parameter sensitivity of a model. Among the reasons to conduct a sensitivity analysis, we highlight the need to determine[51]:

1. which parameters are insignificant and can be eliminated from the final model.
2. which inputs contribute most to output variability.
3. which parameters are most highly correlated with the output.
4. what consequence results from changing a given input parameter.

In general, sensitivity analyses are conducted by[51]:

(a) defining the model and its input and output variables.
(b) assigning probability density functions to each input parameter.
(c) generating input data through random sampling and computing output data.
(d) assessing the influences and relative importance of each input/output relationship.

After a sensitivity analysis has been carried out, the input variables can be ranked according to their measure of sensitivity, *i.e.* how much influence they have on the model output. Several different methods exist to perform such a ranking, and their results may differ. Different results among the lower ranking variables are not of practical concerns, as it is the variables that consistently achieve a high ranking that has the highest influence on the model output.

A.1 One-at-a-time Sensitivity Measures

We can distinguish between methods that explore the full parameter space and methods that explore a subset of the parameter space. A popular variance-based method that explores the

full input space is given in Sobol[52], where the Monte Carlo methods are used to estimate the sensitivity indices. Here we will focus on methods that explores a subset of the parameter space by varying one variable at a time (OAT) while keeping the others at their mean value. Assuming that we have assigned probability density functions to our input variables, generated input data and computed the corresponding output values, we will look at three OAT sensitivity measures: the sensitivity index (SI), the local sensitivity (LS) and the output variance σ_y^2.

A.1.1 Sensitivity Index

The sensitivity index for a variable is calculated as the fractional difference in the output when the input variable is varied from its minimum value to its maximum value. Thus, the sensitivity index for a variable $\mathbf{x}_j$ is given as

$$\mathrm{SI}_j = \frac{D_{\max}}{D_{\max} - D_{\min}}, \tag{A.1}$$

where $D_{\min}$ and $D_{\max}$ represent the minimum and maximum output values, resulting from varying $\mathbf{x}_j$ over its entire range[51].

A.1.2 Local Sensitivity

The local sensitivity determines how small perturbations near a fixed point in input space $x^0 = [x_1^0, \ldots, x_p^0]^T$ influence the output value[53]. To obtain the local sensitivities, we compute the partial derivatives

$$A_j = \left.\frac{\partial y}{\partial \mathbf{x}_j}\right|_{x=x^0}, \tag{A.2}$$

and evaluate them at x^0. For our purpose the fixed point will be equal to mean input values, *i.e.* $x^0 = [\overline{x_1}, \ldots, \overline{x_p}]$. To compare the magnitudes of the local sensitivity, the absolute value of A_j is used.

A.1.3 Output Variance

The output variance σ_{yj}^2 is a measure of the variance of the output when one input at a time is varied over its entire range while the other variables are kept fixed at their mean values. Thus,

the output variance for a variable $\mathbf{x}_j$ provides a measure of how varying $\mathbf{x}_j$ causes variability in y.

Appendix B

Regression Results

This appendix supplements Section 5.2 and presents all the results from both the linear and nearest-neighbors regression models for all five datasets.

B.1 Linear Regression Results

For the simple models, the coefficients, their confidence intervals and p-values and the regression diagnostic plots are presented. For the parsimonious models, additional plots with variable selection through cross-validation and input-output plots are presented.

B.1.1 Tank 1

Simple Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.017469	$[-0.017809, -0.017130]$	0
y	$\hat{\beta}_1$	-0.003084	$[-0.003441, -0.002727]$	$6.77 \cdot 10^{-55}$
P_{atm}	$\hat{\beta}_2$	0.000573	$[0.000178, 0.000968]$	0.00454
T_{atm}	$\hat{\beta}_3$	-0.000682	$[-0.001116, -0.000247]$	0.00214
$H_{1/3}$	$\hat{\beta}_4$	-0.000969	$[-0.001944, 0.000007]$	0.0518
T_1	$\hat{\beta}_5$	-0.000311	$[-0.001017, 0.000395]$	0.388
v_{wind}	$\hat{\beta}_6$	-0.000170	$[-0.000848, 0.000507]$	0.622
RH	$\hat{\beta}_7$	0.000318	$[-0.000046, 0.000683]$	0.0871
Reduction of base error rate: 30.69 %				

Table B.1: Tank 1 simple model: Estimated coefficients with 95 % confidence intervals. Significant variables are colored green.

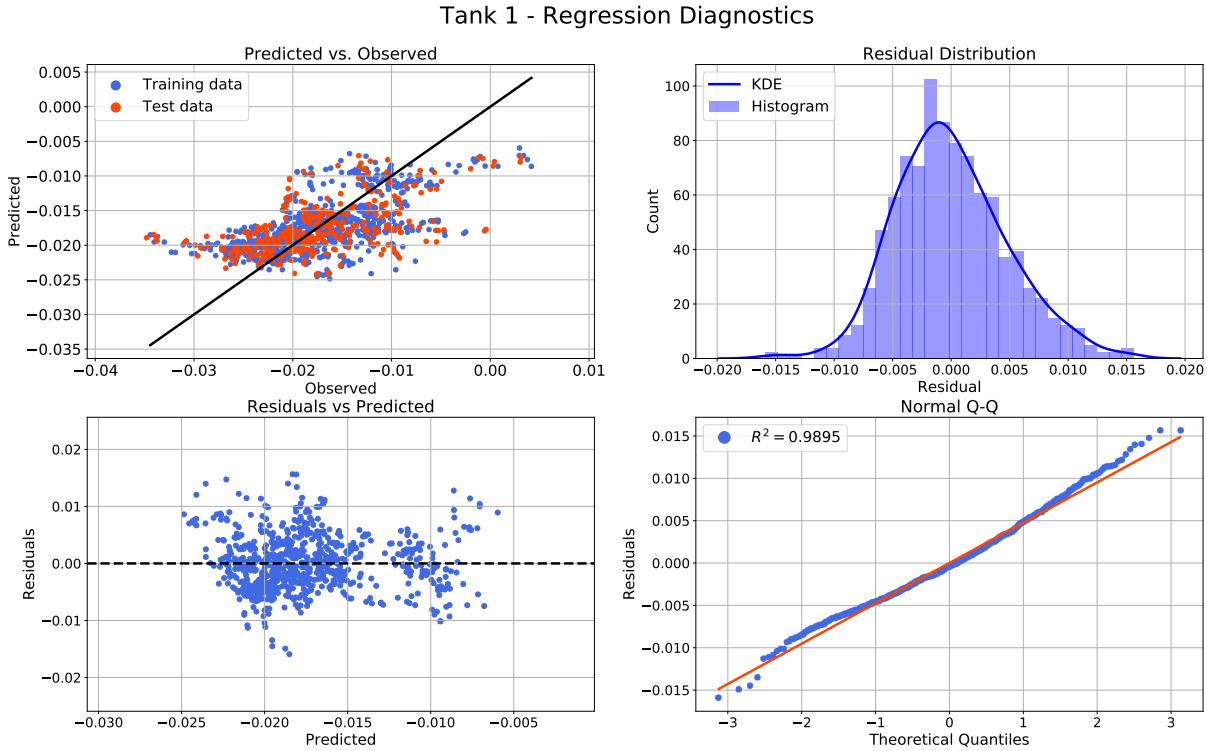

Figure B.1: Tank 1 simple model: Regression diagnostic plots.

Parsimonious Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.017469	$[-0.017778, -0.017161]$	0
y^{-1}	$\hat{\beta}_1$	26.6276	$[20.9373, 32.3179]$	$3.88 \cdot 10^{-19}$
$T_{\text{atm}} \cdot H_{1/3}$	$\hat{\beta}_2$	0.001224	$[0.000027, 0.002421]$	0.0450
P_{atm}^{-1}	$\hat{\beta}_3$	0.000661	$[-0.000034, 0.001356]$	0.0624
T_1^{-1}	$\hat{\beta}_4$	-0.018141	$[-0.028446, -0.007835]$	$5.80 \cdot 10^{-4}$
$T_1 \cdot RH$	$\hat{\beta}_5$	0.000012	$[-0.003827, 0.003851]$	0.995
$H_{1/3}^{-1}$	$\hat{\beta}_6$	-0.000957	$[-0.001559, -0.000355]$	0.00186
$H_{1/3} \cdot T_1$	$\hat{\beta}_7$	-0.005352	$[-0.007753, -0.002951]$	$1.38 \cdot 10^{-5}$
$T_{\text{atm}} \cdot RH$	$\hat{\beta}_8$	-0.000993	$[-0.001705, -0.000281]$	0.00635
$H_{1/3}^3$	$\hat{\beta}_9$	0.001546	$[0.000589, 0.002504]$	0.00158
T_1^3	$\hat{\beta_{10}}$	-0.008899	$[-0.014980, -0.002818]$	0.00418
$\log(T_1)$	$\hat{\beta_{11}}$	-0.059067	$[-0.087667, -0.030468]$	$5.55 \cdot 10^{-5}$
$P_{\text{atm}} \cdot T_1$	$\hat{\beta_{12}}$	0.050853	$[0.026244, 0.075461]$	$5.50 \cdot 10^{-5}$
$\log(y)$	$\hat{\beta_{13}}$	53.1939	$[41.8896, 64.4983]$	$2.53 \cdot 10^{-19}$
y^3	$\hat{\beta_{14}}$	26.7827	$[21.1793, 32.3862]$	$7.40 \cdot 10^{-20}$
y^2	$\hat{\beta_{15}}$	-53.3512	$[-64.5686, -42.1337]$	$1.10 \cdot 10^{-19}$
Reduction of base error rate: 44.66 %				

Table B.2: Tank 1 parsimonious model: Estimated coefficients with 95 % confidence intervals. Significant variables at $\alpha = 0.05$ are colored green.

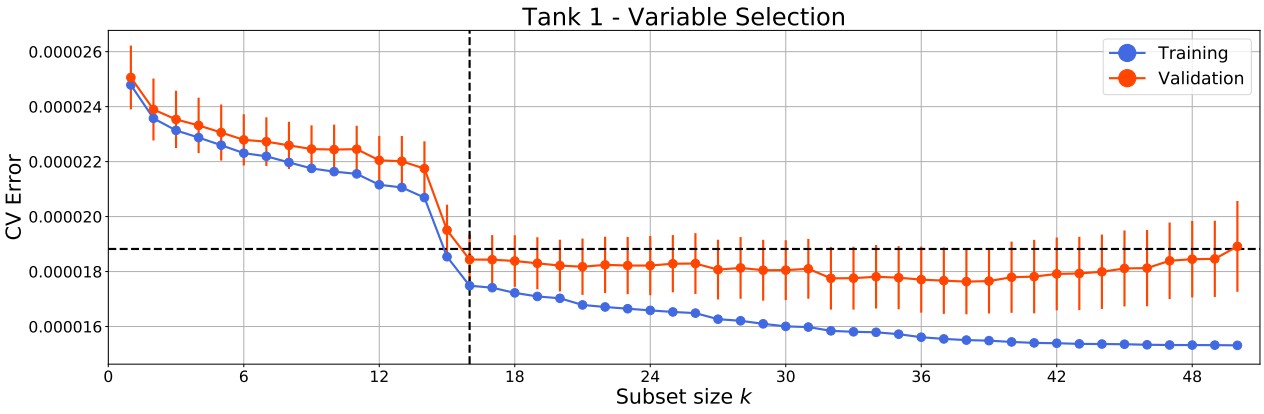

Figure B.2: Tank 1 parsimonious model: Variable selection by cross-validation.

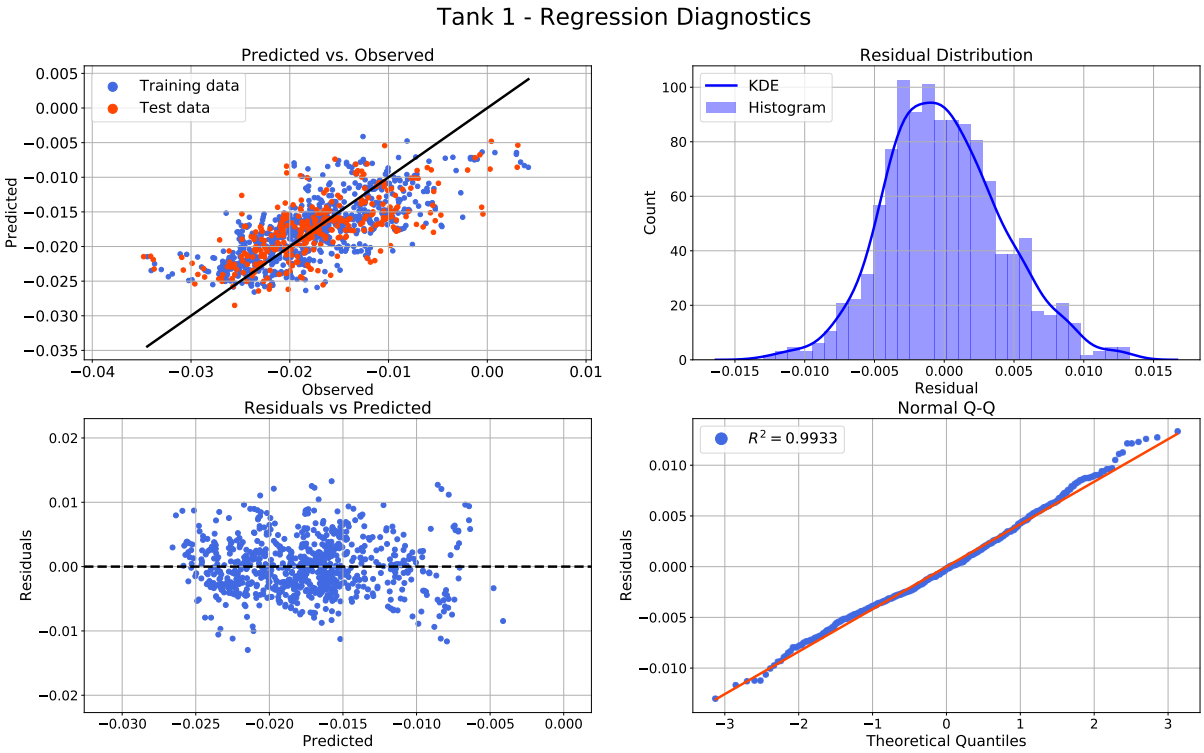

Figure B.3: Tank 1 parsimonious model: Regression diagnostic plots.

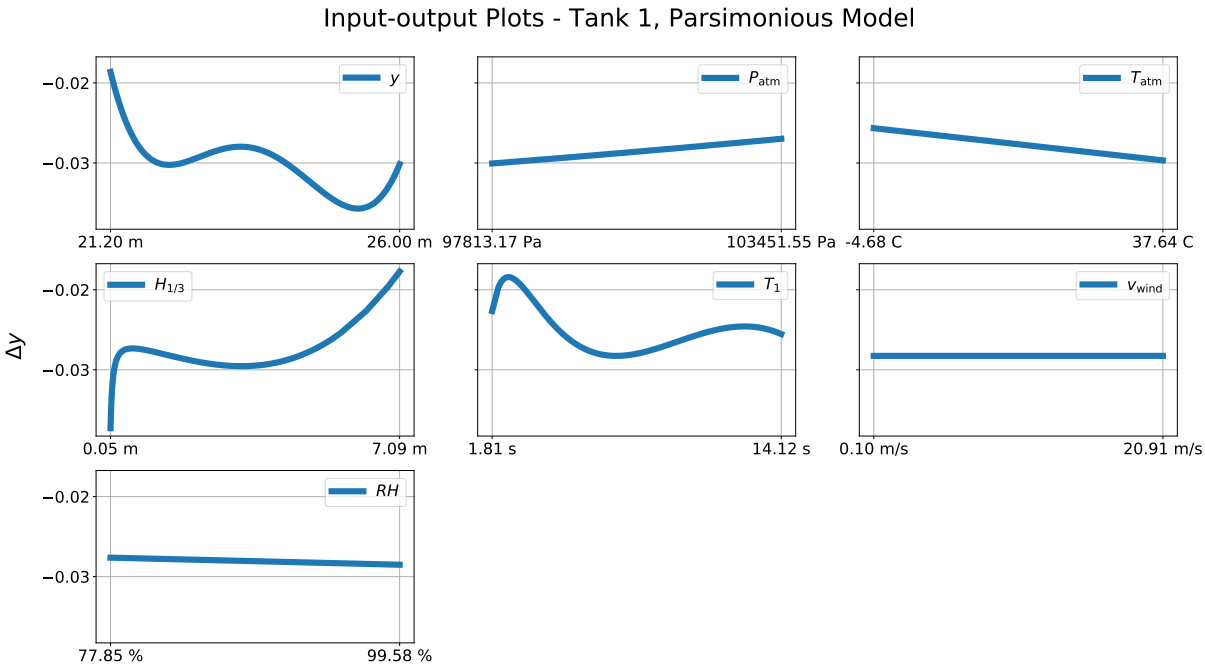

Figure B.4: Tank 1 parsimonious model: Input-output plots.

B.1.2 Tank 2

Simple Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.011242	$[-0.011538, -0.010946]$	0
y	$\hat{\beta}_1$	-0.001410	$[-0.001714, -0.001106]$	$7.76 \cdot 10^{-19}$
P_{atm}	$\hat{\beta}_2$	0.000484	$[0.000136, 0.000832]$	0.00646
T_{atm}	$\hat{\beta}_3$	-0.000250	$[-0.000629, 0.000129]$	0.196
$H_{1/3}$	$\hat{\beta}_4$	-0.002122	$[-0.002987, -0.001257]$	$1.77 \cdot 10^{-6}$
T_1	$\hat{\beta}_5$	0.000932	$[0.000317, 0.001546]$	0.00302
v_{wind}	$\hat{\beta}_6$	0.000674	$[0.000077, 0.001272]$	0.0270
RH	$\hat{\beta}_7$	0.000091	$[-0.000225, 0.000406]$	0.573
Reduction of base error rate: 11.99 %				

Table B.3: Tank 2 simple model: Estimated coefficients with 95 % confidence intervals. Significant variables are colored green.

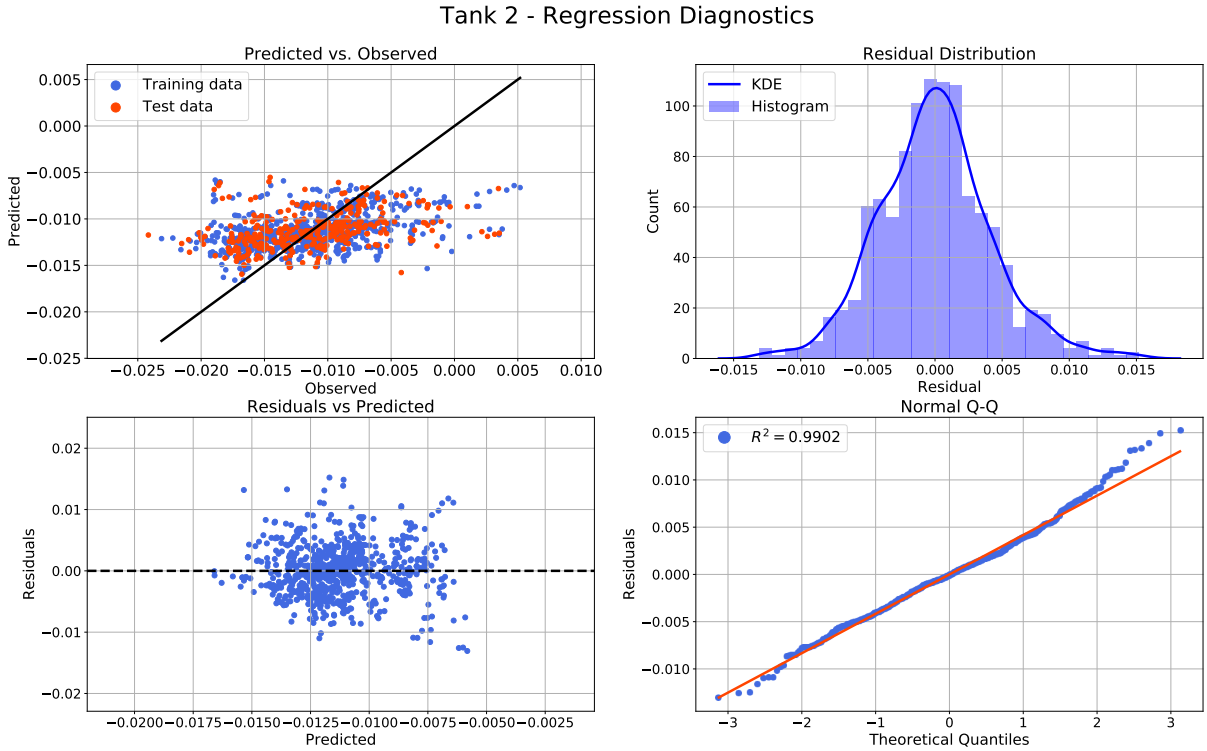

Figure B.5: Tank 2 simple model: Regression diagnostic plots.

Parsimonious Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.011242	$[-0.011526, -0.010958]$	0
y^{-1}	$\hat{\beta}_1$	0.001492	$[0.001198, 0.001786]$	$4.84 \cdot 10^{-22}$
$H_{1/3}^2$	$\hat{\beta}_2$	0.000299	$[-0.002200, 0.002798]$	0.815
T_{atm}^3	$\hat{\beta}_3$	-0.000921	$[-0.001402, -0.000441]$	$1.79 \cdot 10^{-4}$
T_1^3	$\hat{\beta}_4$	-0.035154	$[-0.055154, -0.015153]$	$5.91 \cdot 10^{-4}$
T_1	$\hat{\beta}_5$	-0.090304	$[-0.13525, -0.045354]$	$8.76 \cdot 10^{-5}$
$H_{1/3}^{-1}$	$\hat{\beta}_6$	0.000407	$[-0.000563, 0.001378]$	0.410
$H_{1/3} \cdot v_{\text{wind}}$	$\hat{\beta}_7$	0.001652	$[0.000308, 0.002997]$	0.0161
$H_{1/3} \cdot T_1$	$\hat{\beta}_8$	-0.006476	$[-0.009975, -0.002978]$	$2.98 \cdot 10^{-4}$
T_{atm}^{-1}	$\hat{\beta}_9$	-0.000585	$[-0.001033, -0.000137]$	0.0106
T_1^2	$\hat{\beta}_{10}$	0.10298	$[0.051998, 0.15396]$	$8.02 \cdot 10^{-5}$
$\log(T_1)$	$\hat{\beta}_{11}$	0.025160	$[0.010996, 0.039323]$	$5.16 \cdot 10^{-4}$
$\log(H_{1/3})$	$\hat{\beta}_{12}$	0.002173	$[0.000433, 0.003912]$	0.0145
Reduction of base error rate: 19.75 %				

Table B.4: Tank 2 parsimonious model: Estimated coefficients with 95 % confidence intervals. Significant variables at $\alpha = 0.05$ are colored green.

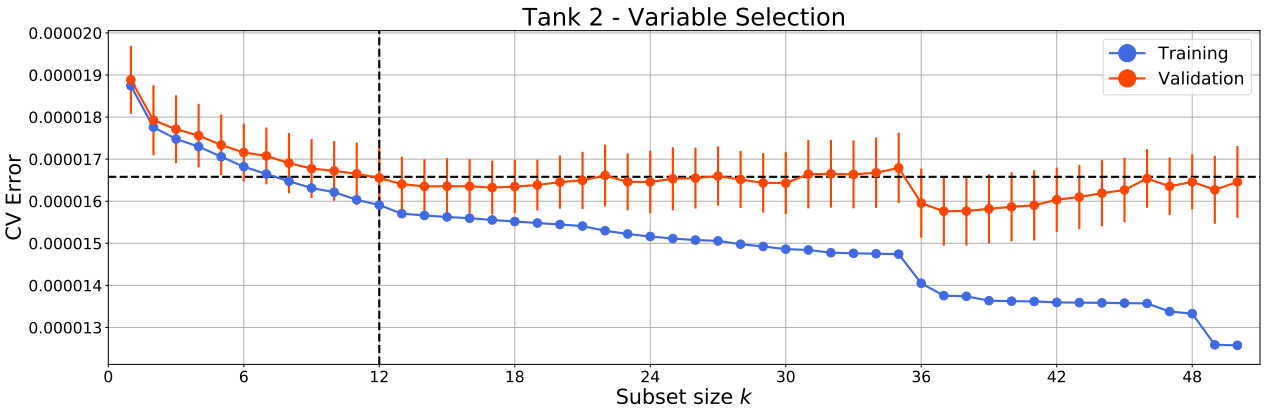

Figure B.6: Tank 2 parsimonious model: Variable selection by cross-validation.

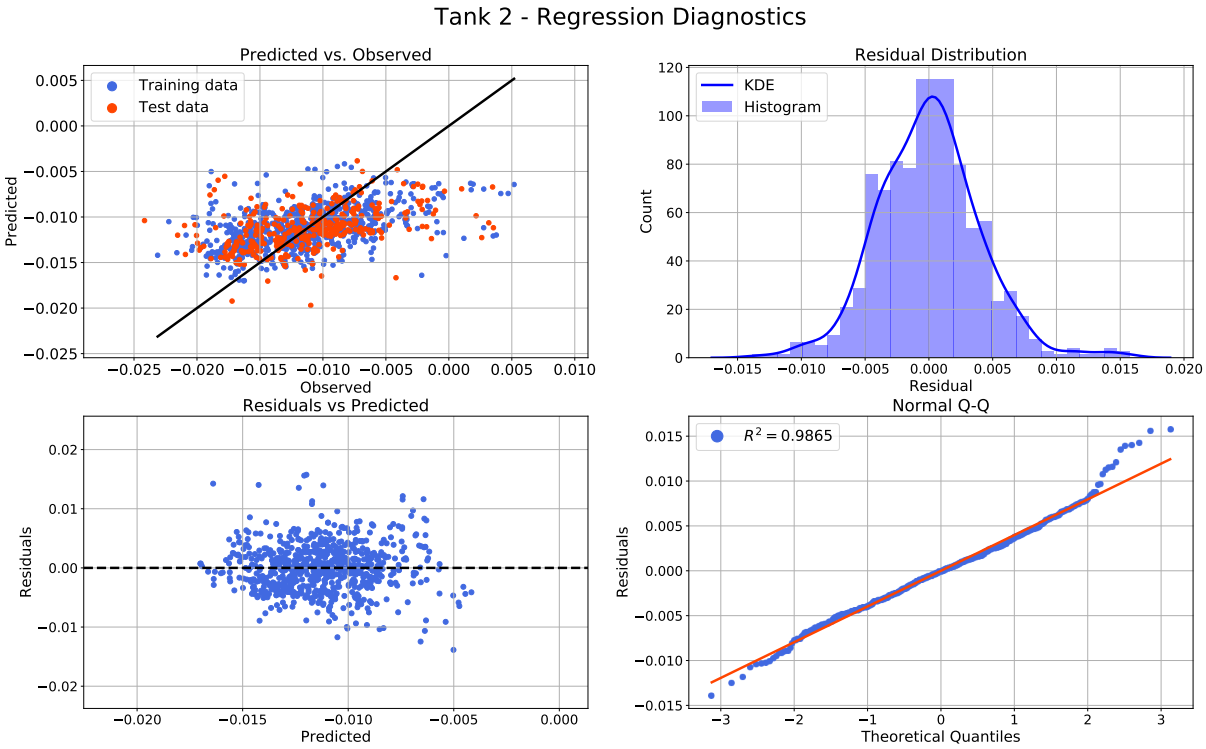

Figure B.7: Tank 2 parsimonious model: Regression diagnostic plots.

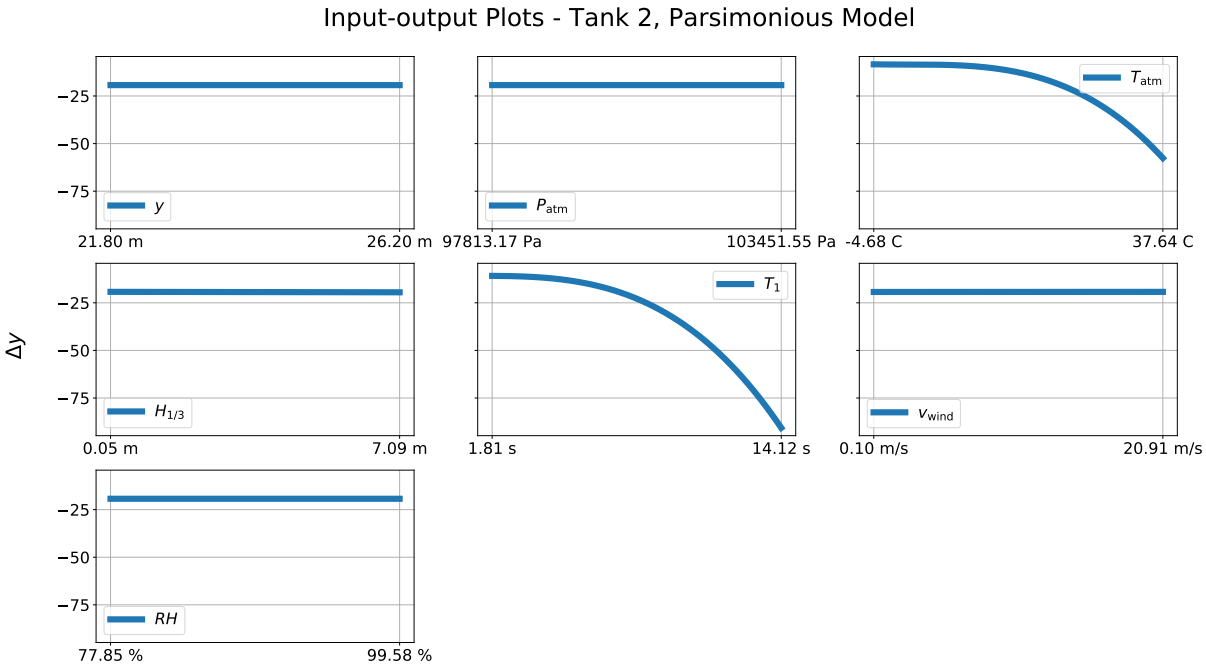

Figure B.8: Tank 2 parsimonious model: Input-output plots.

B.1.3 Tank 3

Simple Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.013256	$[-0.013677, -0.012834]$	0
y	$\hat{\beta}_1$	-0.001577	$[-0.002015, -0.001138]$	$3.71 \cdot 10^{-12}$
P_{atm}	$\hat{\beta}_2$	0.001172	$[0.000671, 0.001674]$	$5.29 \cdot 10^{-6}$
T_{atm}	$\hat{\beta}_3$	-0.001347	$[-0.001894, -0.000800]$	$1.63 \cdot 10^{-6}$
$H_{1/3}$	$\hat{\beta}_4$	-0.000803	$[-0.002019, 0.000413]$	0.195
T_1	$\hat{\beta}_5$	0.000230	$[-0.000625, 0.001086]$	0.597
v_{wind}	$\hat{\beta}_6$	0.000426	$[-0.000420, 0.001272]$	0.323
RH	$\hat{\beta}_7$	-0.000531	$[-0.000977, -0.000085]$	0.0198
Reduction of base error rate: 17.81 %				

Table B.5: Tank 3 simple model: Estimated coefficients with 95 % confidence intervals. Significant variables are colored green.

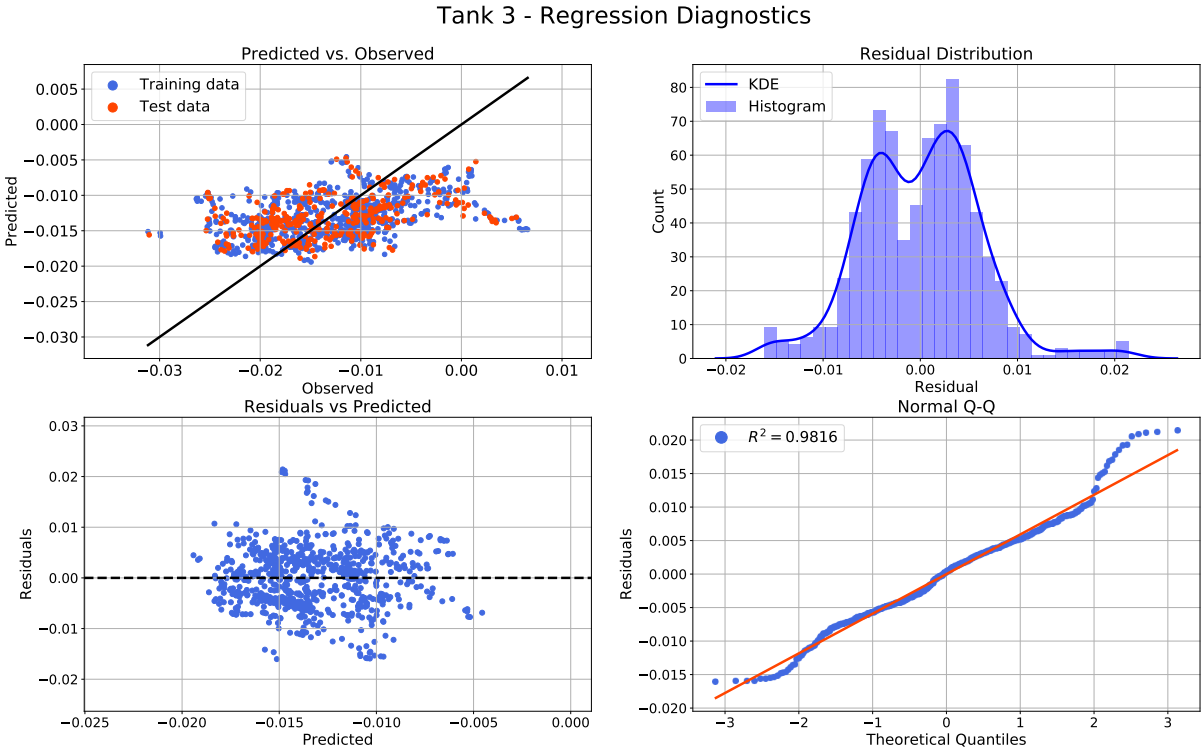

Figure B.9: Tank 3 simple model: Regression diagnostic plots.

Parsimonious Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.013256	$[-0.013611, -0.012901]$	0
$T_{\text{atm}} \cdot RH$	$\hat{\beta}_1$	0.000192	$[-0.000344, 0.000729]$	0.481
y^{-1}	$\hat{\beta}_2$	-0.667838	$[-1.038393, -0.297282]$	$4.28 \cdot 10^{-4}$
y^3	$\hat{\beta}_3$	0.263433	$[0.140886, 0.385981]$	$2.74 \cdot 10^{-5}$
P_{atm}^3	$\hat{\beta}_4$	-0.970372	$[-3.919349, 1.978605]$	0.518
$\log(y)$	$\hat{\beta}_5$	-0.932355	$[-1.425017, -0.439692]$	$2.18 \cdot 10^{-4}$
$H_{1/3}^3$	$\hat{\beta}_6$	-0.000122	$[-0.001330, 0.001087]$	0.843
P_{atm}^{-1}	$\hat{\beta}_7$	0.334692	$[-0.642450, 1.311834]$	0.502
P_{atm}^2	$\hat{\beta}_8$	1.304714	$[-2.620813, 5.230240]$	0.514
v_{wind}^3	$\hat{\beta}_9$	0.000051	$[-0.000861, 0.000963]$	0.912
T_1^3	$\hat{\beta_{10}}$	-0.138753	$[-0.198992, -0.078514]$	$7.12 \cdot 10^{-6}$
$P_{\text{atm}} \cdot T_1$	$\hat{\beta_{11}}$	0.117198	$[0.034925, 0.199471]$	0.00530
$H_{1/3}^{-1}$	$\hat{\beta_{12}}$	-0.000320	$[-0.001129, 0.000488]$	0.437
T_1^2	$\hat{\beta_{13}}$	0.430188	$[0.242080, 0.618296]$	$8.25 \cdot 10^{-6}$
T_1^{-1}	$\hat{\beta_{14}}$	0.050272	$[0.019444, 0.081101]$	0.00143
T_1	$\hat{\beta_{15}}$	-0.601217	$[-0.847707, -0.354726]$	$2.02 \cdot 10^{-6}$
$\log(T_1)$	$\hat{\beta_{16}}$	0.245103	$[0.114590, 0.375616]$	$2.43 \cdot 10^{-4}$
$H_{1/3} \cdot T_1$	$\hat{\beta_{17}}$	-0.009696	$[-0.015939, -0.003453]$	0.00238
$H_{1/3}$	$\hat{\beta_{18}}$	0.128378	$[0.058820, 0.197936]$	$3.11 \cdot 10^{-4}$
$P_{\text{atm}} \cdot H_{1/3}$	$\hat{\beta_{19}}$	-0.121141	$[-0.190095, -0.052187]$	$5.94 \cdot 10^{-4}$
Reduction of base error rate: 45.20 %				

Table B.6: Tank 3 parsimonious model: Estimated coefficients with 95 % confidence intervals. Significant variables at $\alpha = 0.05$ are colored green.

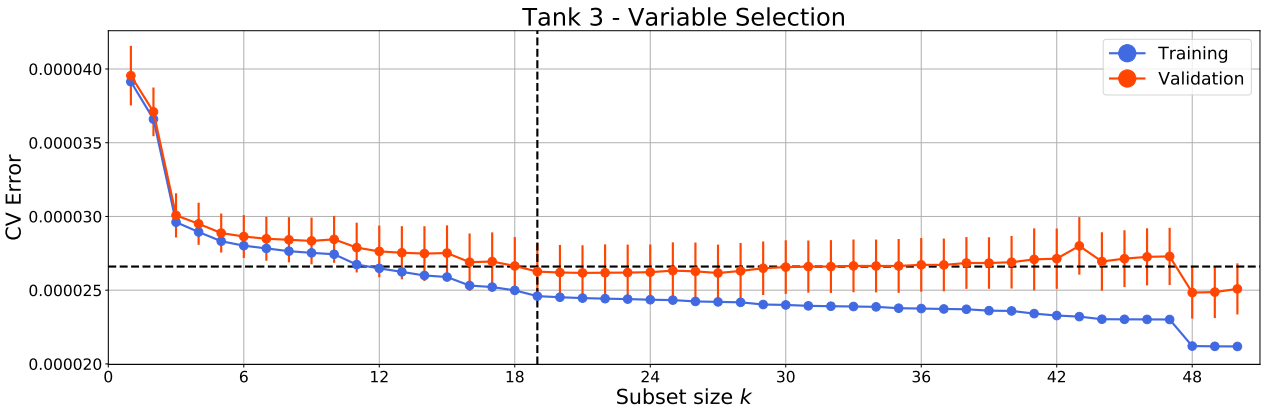

Figure B.10: Tank 3 parsimonious model: Variable selection by cross-validation.

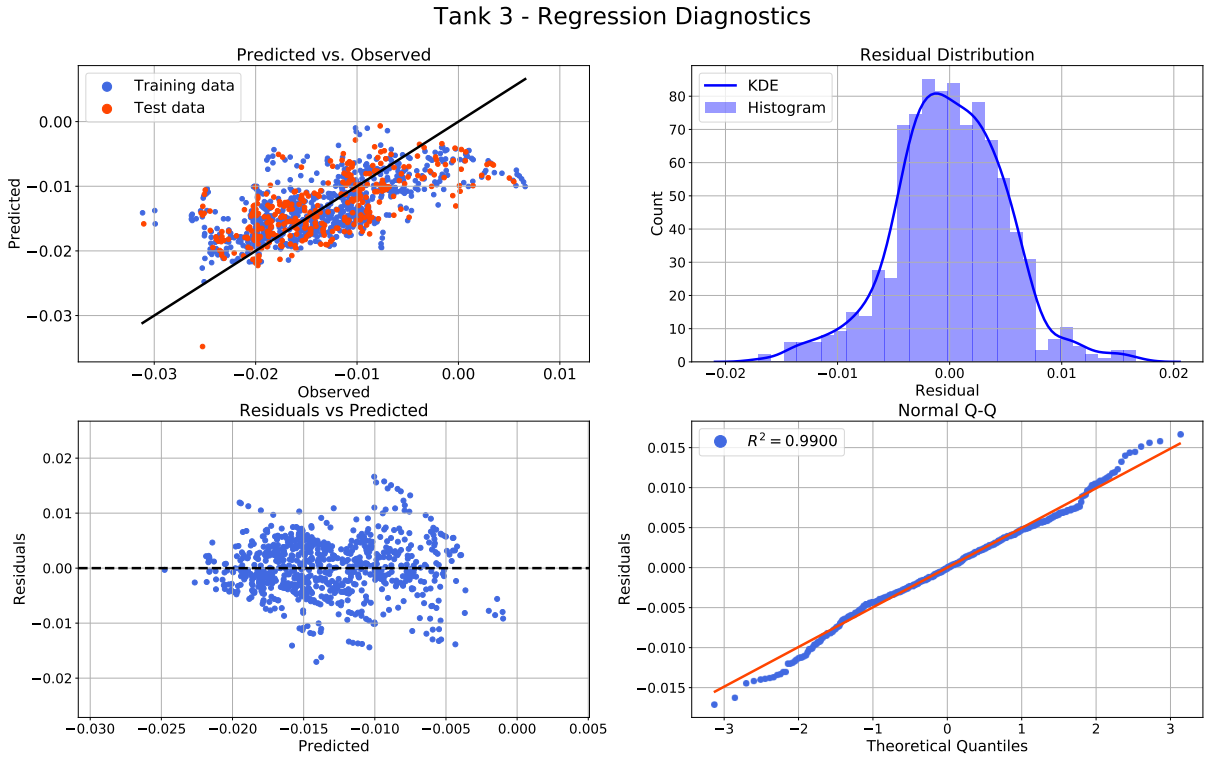

Figure B.11: Tank 3 parsimonious model: Regression diagnostic plots.

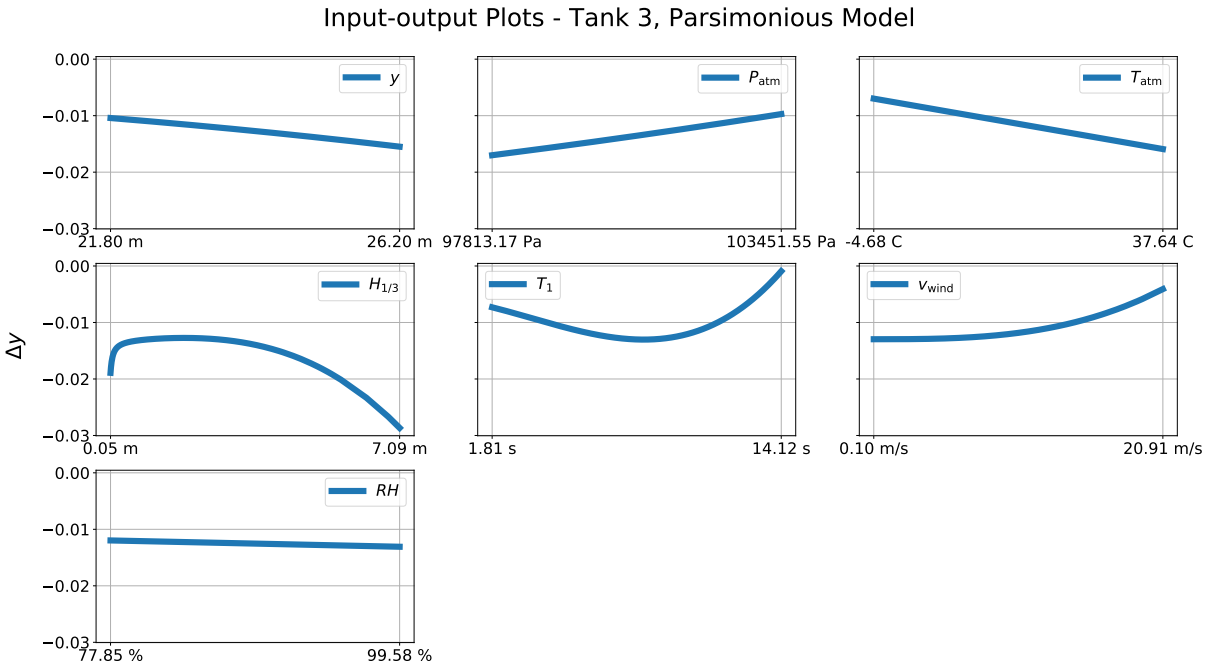

Figure B.12: Tank 3 parsimonious model: Input-output plots.

B.1.4 Tank 4

Simple Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.014598	$[-0.014959, -0.014238]$	0
y	$\hat{\beta}_1$	-0.000282	$[-0.000653, 0.000090]$	0.137
P_{atm}	$\hat{\beta}_2$	0.000013	$[-0.000414, 0.000441]$	0.952
T_{atm}	$\hat{\beta}_3$	-0.001038	$[-0.001501, -0.000574]$	$1.26 \cdot 10^{-5}$
$H_{1/3}$	$\hat{\beta}_4$	-0.001431	$[-0.002468, -0.000394]$	0.00690
T_1	$\hat{\beta}_5$	0.001453	$[0.000718, 0.002187]$	$1.12 \cdot 10^{-4}$
v_{wind}	$\hat{\beta}_6$	-0.000219	$[-0.000936, 0.000497]$	0.548
RH	$\hat{\beta}_7$	0.000223	$[-0.000158, 0.000604]$	0.251
Reduction of base error rate: 3.10 %				

Table B.7: Tank 4 simple model: Estimated coefficients with 95 % confidence intervals. Significant variables are colored green.

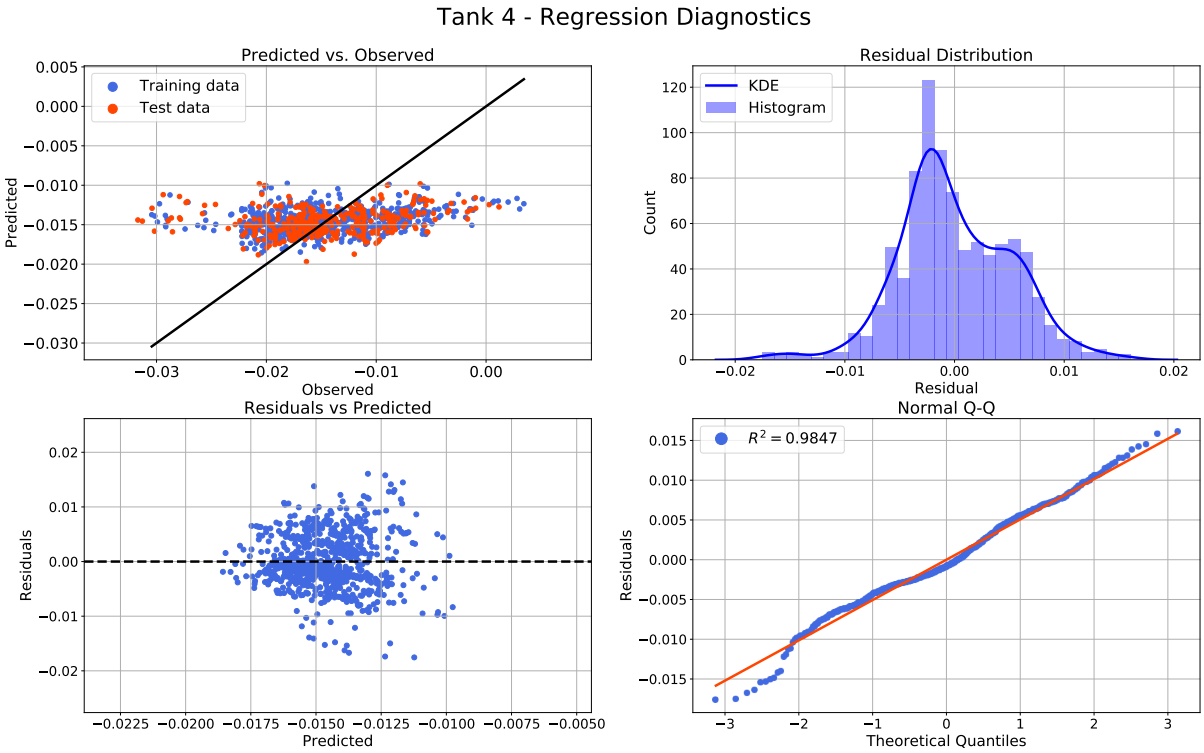

Figure B.13: Tank 4 simple model: Regression diagnostic plots.

Parsimonious Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.014598	$[-0.014950, -0.014247]$	0
$T_{\text{atm}} \cdot v_{\text{wind}}$	$\hat{\beta}_1$	0.006965	$[0.004460, 0.009470]$	$6.48 \cdot 10^{-8}$
$P_{\text{atm}} \cdot RH$	$\hat{\beta}_2$	0.000253	$[-0.002710, 0.003217]$	0.867
$T_{\text{atm}} \cdot H_{1/3}$	$\hat{\beta}_3$	-0.010022	$[-0.013446, -0.006598]$	$1.32 \cdot 10^{-8}$
T_{atm}^3	$\hat{\beta}_4$	0.009051	$[0.005203, 0.012899]$	$4.57 \cdot 10^{-6}$
v_{wind}^2	$\hat{\beta}_5$	-0.000416	$[-0.001898, 0.001065]$	0.518
T_{atm}^2	$\hat{\beta}_6$	-0.015766	$[-0.021032, -0.010500]$	$6.22 \cdot 10^{-9}$
$T_{\text{atm}} \cdot T_1$	$\hat{\beta}_7$	0.010591	$[0.007051, 0.014130]$	$6.34 \cdot 10^{-9}$
$\log(T_1)$	$\hat{\beta}_8$	-0.007000	$[-0.009423, -0.004578]$	$1.98 \cdot 10^{-8}$
$H_{1/3} \cdot RH$	$\hat{\beta}_9$	0.014341	$[0.008987, 0.019695]$	$1.89 \cdot 10^{-7}$
$v_{\text{wind}} \cdot RH$	$\hat{\beta}_{10}$	-0.007206	$[-0.010340, -0.004072]$	$7.39 \cdot 10^{-6}$
$H_{1/3} \cdot T_1$	$\hat{\beta}_{11}$	-0.005912	$[-0.008893, -0.002932]$	$1.07 \cdot 10^{-4}$
Reduction of base error rate: 5.89 %				

Continued on next page

Variable	Coefficient	Value	95 % confidence interval	p-value

Table B.8: Tank 4 parsimonious model: Estimated coefficients with 95 % confidence intervals. Significant variables at $\alpha = 0.05$ are colored green.

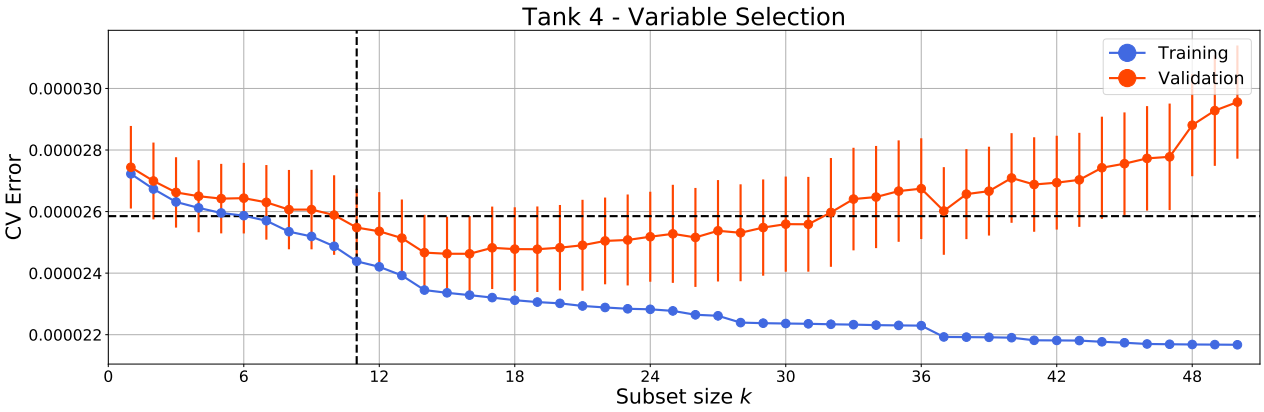

Figure B.14: Tank 4 parsimonious model: Variable selection by cross-validation.

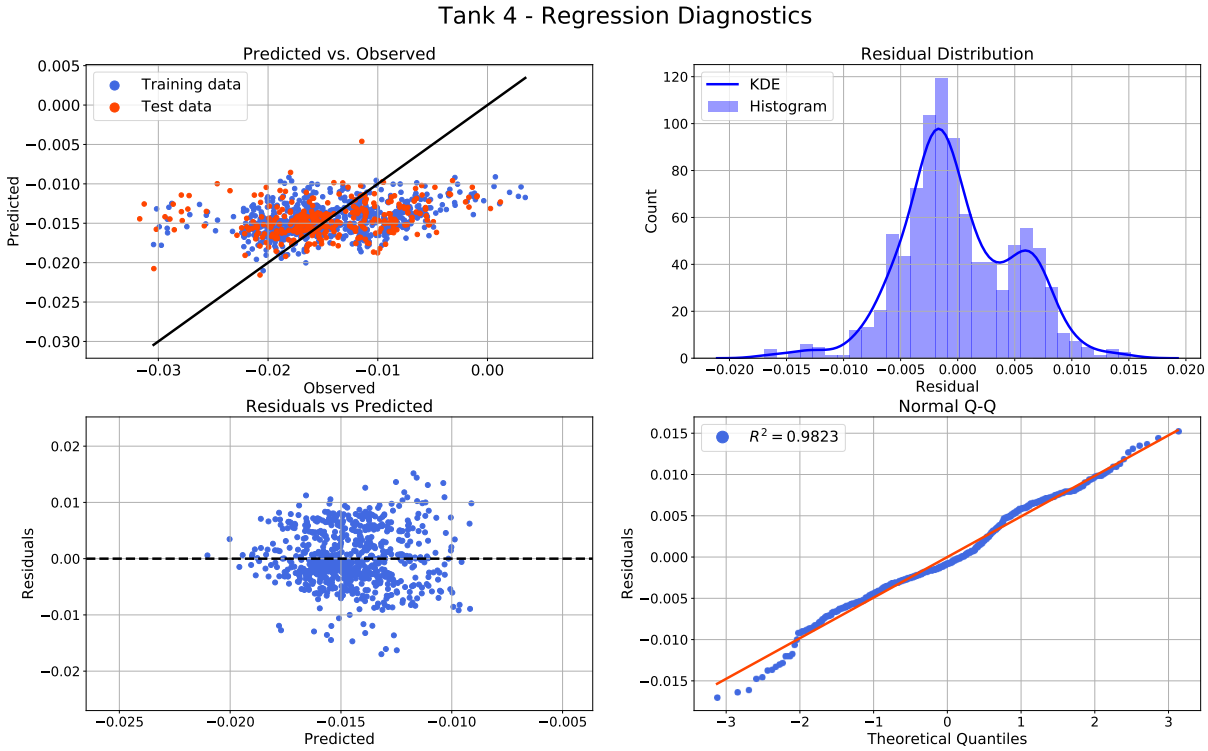

Figure B.15: Tank 4 parsimonious model: Regression diagnostic plots.

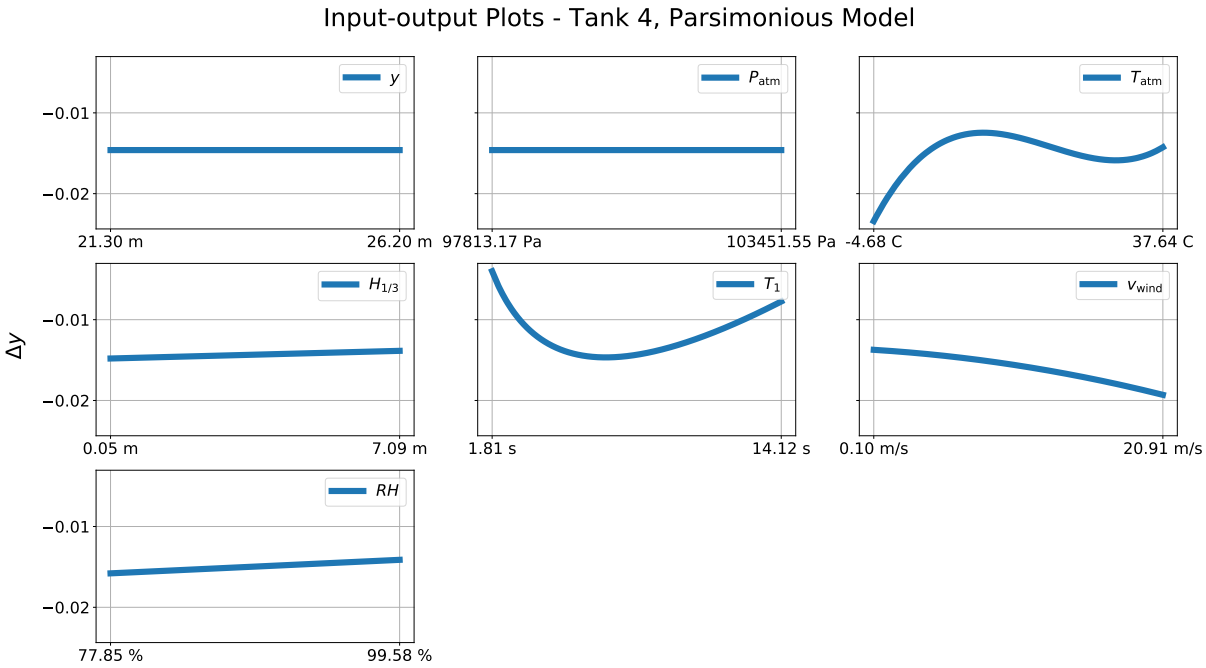

Figure B.16: Tank 4 parsimonious model: Input-output plots.

B.1.5 All Tanks Combined

Simple Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.056929	$[-0.058080, -0.055778]$	0
y	$\hat{\beta}_1$	-0.006665	$[-0.007855, -0.005475]$	$3.43 \cdot 10^{-26}$
P_{atm}	$\hat{\beta}_2$	0.001872	$[0.000522, 0.003221]$	0.00661
T_{atm}	$\hat{\beta}_3$	-0.003196	$[-0.004658, -0.001733]$	$2.03 \cdot 10^{-5}$
$H_{1/3}$	$\hat{\beta}_4$	-0.006528	$[-0.009779, -0.003277]$	$8.82 \cdot 10^{-5}$
T_1	$\hat{\beta}_5$	0.003178	$[0.000890, 0.005466]$	0.00654
v_{wind}	$\hat{\beta}_6$	0.001107	$[-0.001158, 0.003371]$	0.338
RH	$\hat{\beta}_7$	-0.000666	$[-0.001887, 0.000556]$	0.285
Reduction of base error rate: 20.70 %				

Table B.9: All tanks combined simple model: Estimated coefficients with 95 % confidence intervals. Significant variables are colored green.

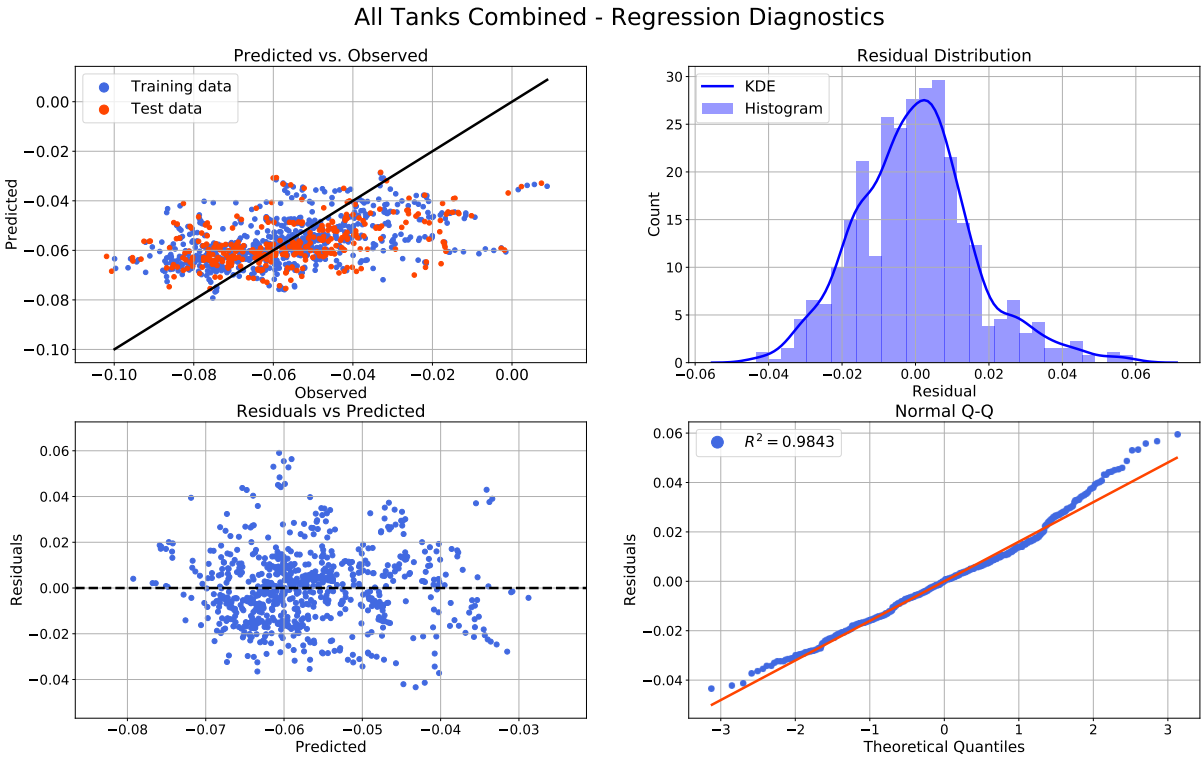

Figure B.17: All tanks combined simple model: Regression diagnostic plots.

Parsimonious Model

Variable	Coefficient	Value	95 % confidence interval	p-value
-	$\hat{\beta}_0$	-0.056929	$[-0.058000, -0.055858]$	0
y^{-1}	$\hat{\beta}_1$	30.499	$[19.075, 41.922]$	$1.67 \cdot 10^{-7}$
P_{atm}^3	$\hat{\beta}_2$	0.004429	$[0.0015954, 0.007262]$	0.00219
$T_{\text{atm}} \cdot H_{1/3}$	$\hat{\beta}_3$	0.001710	$[-0.006144, 0.002725]$	0.450
T_1^3	$\hat{\beta}_4$	0.022959	$[0.015474, 0.030444]$	$1.84 \cdot 10^{-9}$
$T_{\text{atm}} \cdot v_{\text{wind}}$	$\hat{\beta}_5$	0.001863	$[-0.000548, 0.00427]$	0.130
$H_{1/3}^3$	$\hat{\beta}_6$	0.001683	$[-0.001678, 0.00504]$	0.326
T_{atm}^3	$\hat{\beta}_7$	-0.000232	$[-0.002550, 0.00209]$	0.845
$T_1 \cdot RH$	$\hat{\beta}_8$	-0.010179	$[-0.014850, -0.00551]$	$1.95 \cdot 10^{-5}$
$H_{1/3} \cdot T_1$	$\hat{\beta}_9$	-0.029125	$[-0.044830, -0.01342]$	$2.78 \cdot 10^{-4}$
$H_{1/3}$	$\hat{\beta_{10}}$	0.20996	$[0.061173, 0.35875]$	0.00568
$P_{\text{atm}} \cdot H_{1/3}$	$\hat{\beta_{11}}$	-0.19436	$[-0.342953, -0.04578]$	0.0104
y^3	$\hat{\beta_{12}}$	97.721	$[62.923, 132.519]$	$3.71 \cdot 10^{-8}$

Continued on next page

Variable	Coefficient	Value	95 % confidence interval	p-value
y^2	$\hat{\beta_{13}}$	-255.96	$[-348.228, -163.698]$	$5.40 \cdot 10^{-8}$
y	$\hat{\beta_{14}}$	188.74	$[119.847, 257.630]$	$7.89 \cdot 10^{-8}$
Reduction of base error rate: 30.81 %				

Table B.10: All tanks combined parsimonious model: Estimated coefficients with 95 % confidence intervals. Significant variables at $\alpha = 0.05$ are colored green.

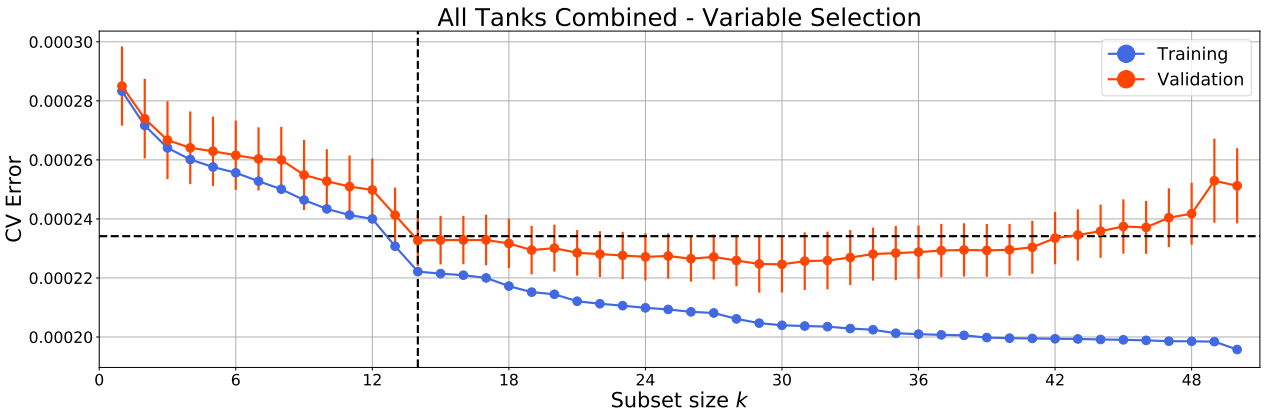

Figure B.18: All tanks combined parsimonious model: Variable selection by cross-validation.

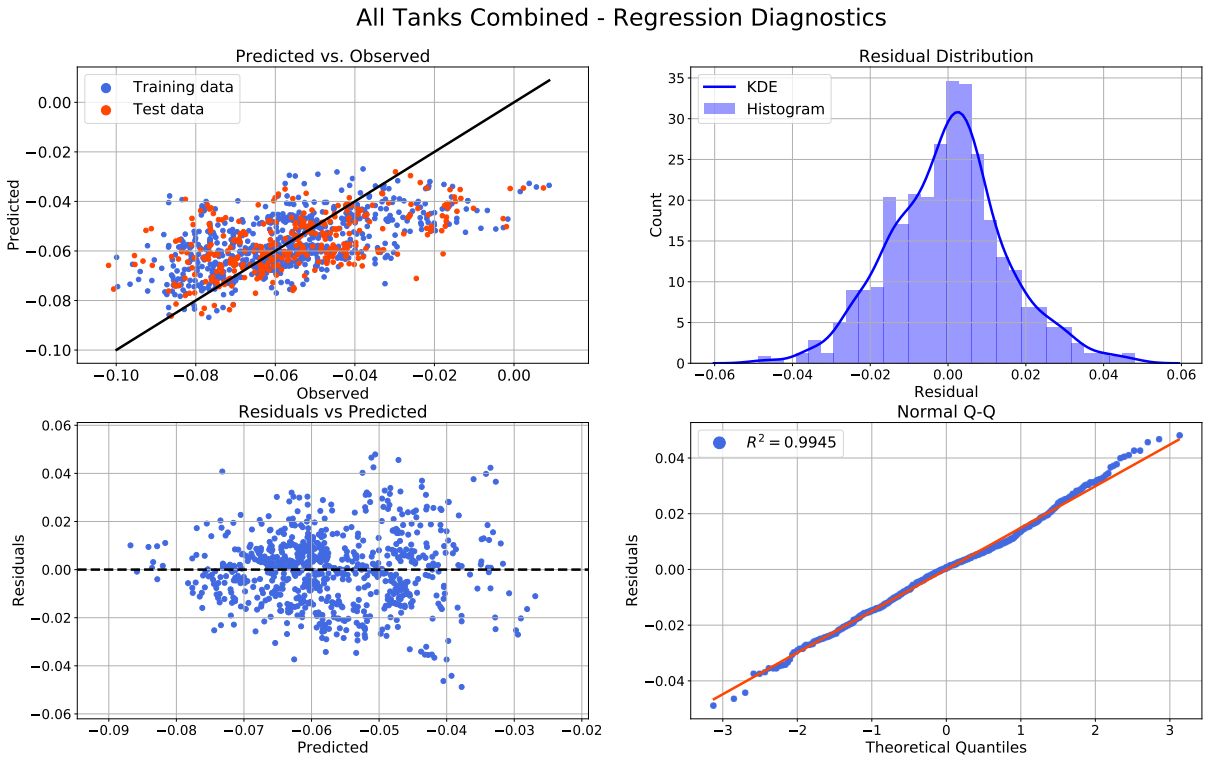

Figure B.19: All tanks combined parsimonious model: Regression diagnostic plots.

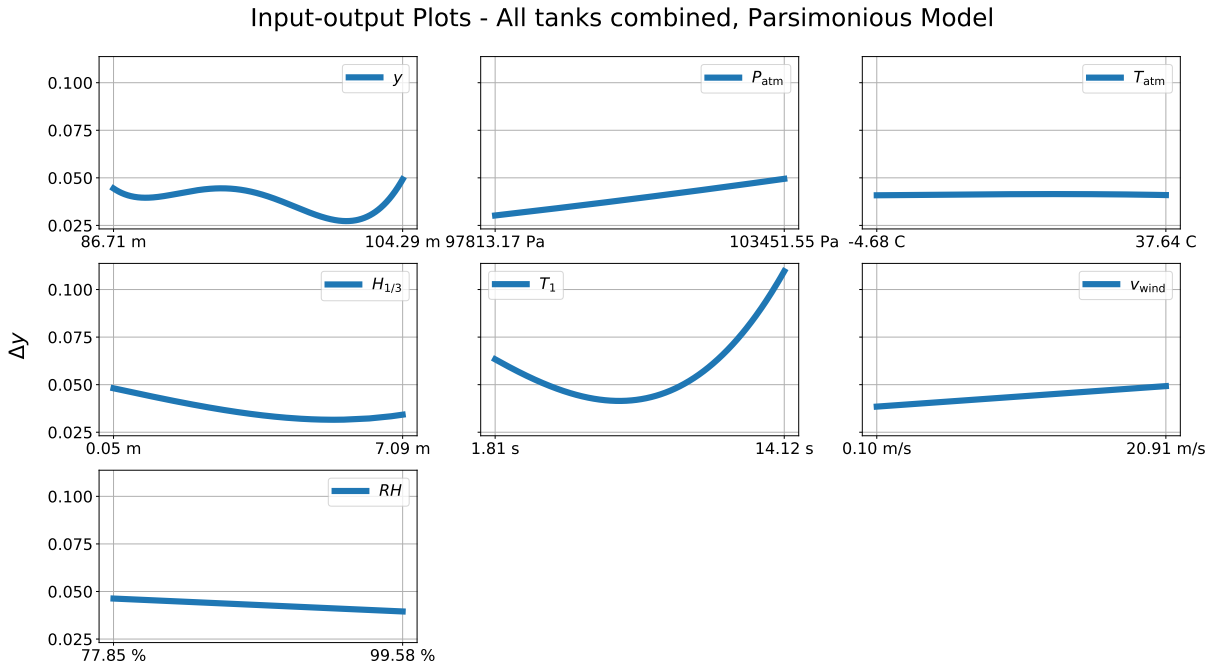

Figure B.20: All tanks combined parsimonious model: Input-output plots.

B.2 Nearest-Neighbors Regression Results

Figures B.21 to B.25 shows the cross-validation training and test error as a function of neighbors k for all five datasets.

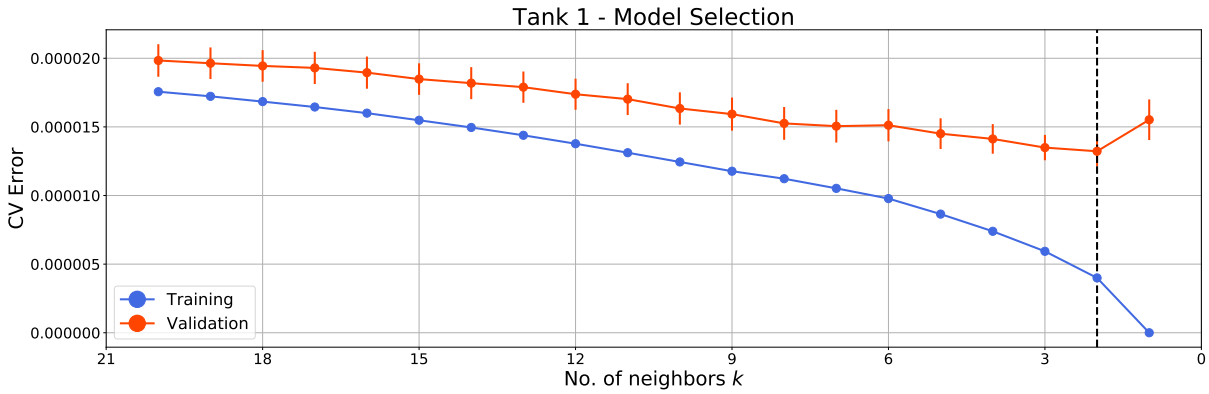

Figure B.21: Tank 1: Cross-validation errors as a function of subset size k with model selection.

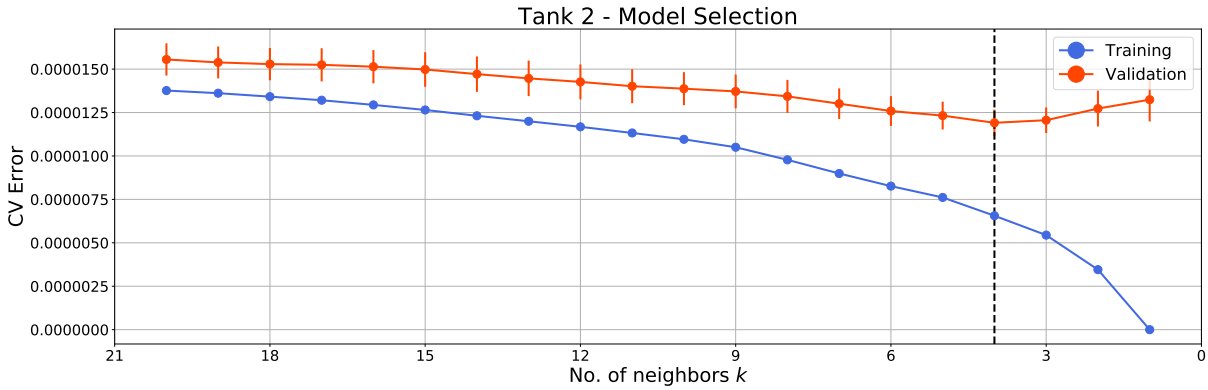

Figure B.22: Tank 2: Cross-validation errors as a function of subset size k with model selection.

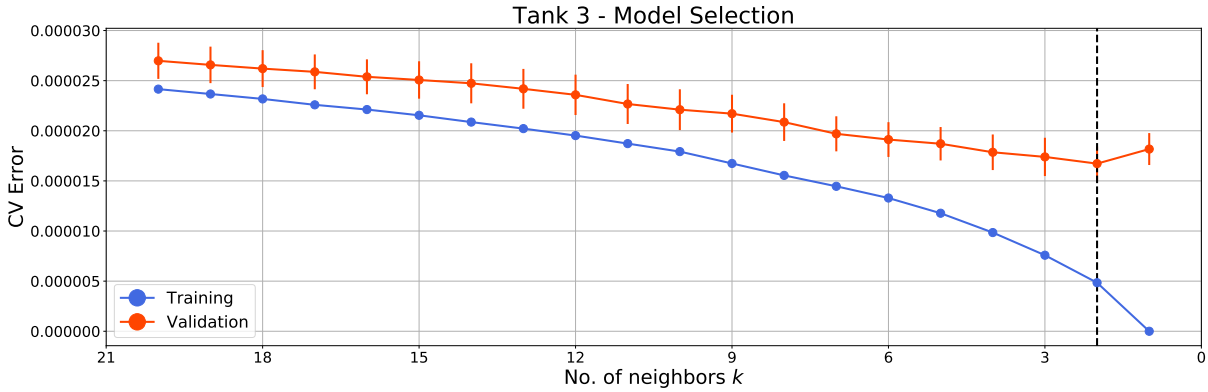

Figure B.23: Tank 3: Cross-validation errors as a function of subset size k with model selection.

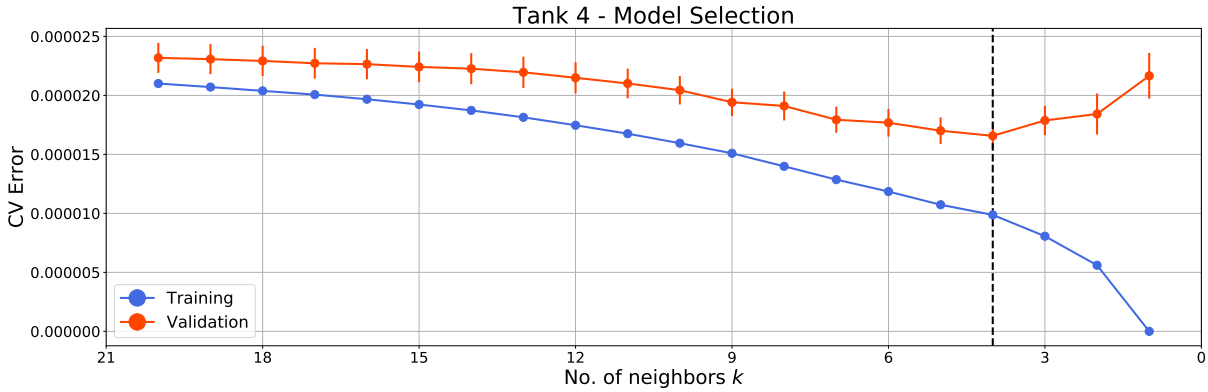

Figure B.24: Tank 4: Cross-validation errors as a function of subset size k with model selection.

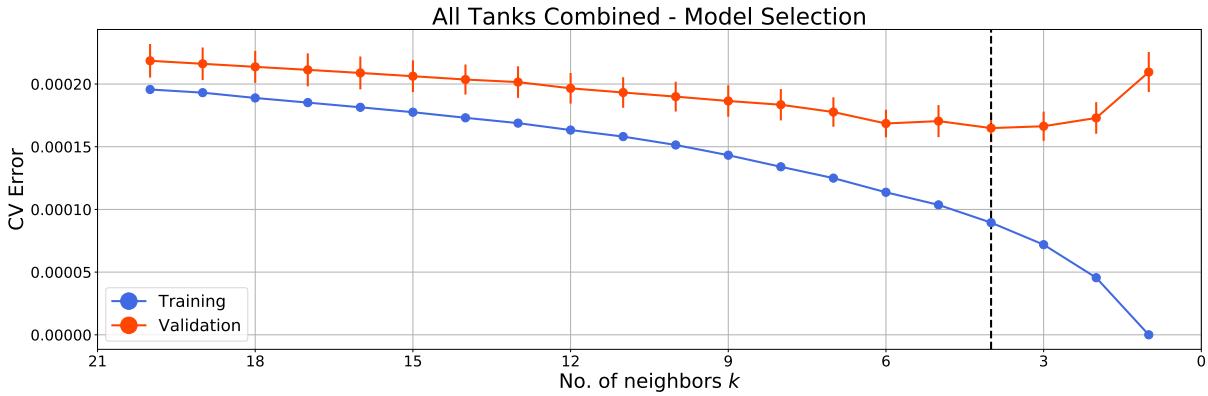

Figure B.25: All tanks combined: Cross-validation errors as a function of subset size k with model selection.