Statistical Monitoring for Failure Detection of Royal Netherlands Navy Vessels

Stijns, Esmée

Award date:
2021

Eindhoven University of Technology

Royal Netherlands Navy

Master Thesis

Department of Mathematics & Computer Science

Statistical Monitoring for Failure Detection of Royal Netherlands Navy Vessels

Author:
Esmée Stijns

Supervisors:
Alessandro Di Bucchianico
Wieger Tiddens
Tiedo Tinga

August 10, 2021

Koninklijke Marine

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Abstract

In this thesis, statistical monitoring techniques for failure detection are applied to data of Royal Netherlands Navy vessels. This is in line with the goal of the Royal Netherlands Navy to achieve smart data-driven maintenance in 2030. This project is still in an early stage. Therefore, in this thesis we aim at designing statistical process control methods to detect failures for the main bearings of the diesel engines of the ocean-going patrol vessels. Since the two most important failure modes (abrasive wear and cavitation) cause an increase in the bearing temperatures, these temperatures are monitored.

A regression model is used to correct for other factors that influence the bearing temperatures. To select the variables that are included in the regression model, LASSO variable selection is applied. The regression model is used to compute predictive and recursive residuals. Methods based on solely predictive or solely recursive residuals proved to be inappropriate to use in practice. A hybrid method based on a combination of recursive and predictive residuals was able to predict one of the two failures in the historical data in advance. This hybrid method is designed to detect weak but persistent upward trends via EWMA control charts.

Another method that proved to be useful is based on regression adjusted variables. This method is designed to detect when a bearing shows deviant behaviour from what is expected given the other bearings. This method is able to detect the other failure in the historical data set in advance. Since both methods can detect different patterns leading to failures, it is recommended to implement both of them at the Royal Netherlands Navy to achieve online monitoring of the bearing temperatures.

Keywords: Predictive maintenance, statistical process control, multiple linear regression, predictive and recursive residuals, LASSO variable selection, regression adjusted variables

Preface

With this master's thesis I complete six years of studying Mathematics at the Eindhoven University of Technology. This was an amazing time for me where I learned a lot, not only about mathematics but also about myself. For this master's thesis I researched data of the Royal Netherlands Navy. Due to Covid-19 this all went a little different than I expected, but in the end it all worked out well and I am happy with the final result.

I would like to thank Alessandro Di Bucchianico for helping me to get in touch with the Royal Netherlands Navy in the first place. I would also like to thank him for the ideas, help and feedback he gave me throughout the project.

Next, I would like to thank Wieger Tiddens and Tiedo Tinga. They helped me a lot with the practical side of the project. Their feedback was really useful and I also appreciate that they pushed me to think about how these methods can be implemented in practice. I would also like to thank Jacco for discussing the results of the various methods with me and Bart and Dennis for their practical support.

I also want to thank Kathrin Möllenhoff and Jaron Sanders for joining the assessment committee.

Finally, I would like to thank Tom, my parents and my friends for their great support not only during this thesis but also during the remainder of my six years in Eindhoven.

Contents

List of Acronyms

ACF Autocorrelation Function.

ARA Adaptive Regression Adjusted.

ARL Average Run Length.

BLUE Best Linear Unbiased Estimator.

CL Centre Line.

CUSUM Cumulative Sum.

EWMA Exponentially Weighted Moving Average.

IC In-Control.

IPMS Integrated Platform Management System.

LASSO Least Absolute Shrinkage and Selection Operator.

LCL Lower Control Limit.

LOOVC Leave One Out Cross Validation.

MAE Mean Absolute Error.

MEWMA Multivariate Exponentially Weighted Moving Average.

MLR Multiple Linear Regression.

MSE Mean Squared Error.

MYT Mason–Young–Tracy.

OC Out-of-Control.

OPVs Ocean-going Patrol Vessels.

PCA Principal Component Analysis.

RAV Regression Adjusted Variables.

RMSE Root Mean Squared Error.

RPM Revolutions Per Minute.

RUL Remaining Useful Life.

SPC Statistical Process Control.

UCL Upper Control Limit.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to the Royal Netherlands Navy

The fleet, navy personnel and marines of the Royal Netherlands navy are deployed worldwide in the interest of ensuring security at sea and from the sea. Both at home and abroad, the Royal Netherlands Navy takes part in crisis management operations and provides humanitarian assistance and disaster relief. The Royal Netherlands Navy is the oldest of the four services of the Netherlands armed forces.

The protection of shipping routes and hubs at sea is essential for unhampered trade and safe energy transport. The navy also intercepts drug transports and combats piracy, human trafficking, gun running and terrorism. To that end, Royal Netherlands Navy ships carry out patrols and boarding and blockade operations. Submarines are used to gather intelligence in sea areas all around the world. Closer to home the Navy is also regularly deployed to clear explosives at sea and in seaports, to carry out search-and-rescue operations and to patrol coastal waters. [1]

1.2 Background

The research leading to this thesis is performed at the department 'Data voor Onderhoud' (English: Data for Maintenance). This is a relatively new department of the Royal Netherlands Navy whose goal is to achieve smart data-driven maintenance by 2030. 'Data voor Onderhoud' only has a few employees and works together with students from different programs and universities to achieve this goal.

At the moment, data from sensors of the Ocean-going Patrol Vessels (OPVs) is investigated. These vessels are designed for the surveillance of coastal waters and operate in a lower violence spectrum mainly at the North Sea and the Caribbean. However, they can also be deployed in anti-piracy and anti-terrorism operations. The sensors that provide the data were already present on the vessels for operational purposes. The goal is to see if those existing sensors can also be used for predictive maintenance.

An important difference between the Royal Netherlands Navy and other organisations is that maintenance cannot be done at all times. The vessels sail all around the world. Small adjustments can be done at sea, but when larger problems occur, the ship usually must sail back to Den Helder before any maintenance can be done. This makes predictive maintenance a very important topic for the Royal Netherlands Navy.

Predictive maintenance also fits well within the defence vision of 2035. In this vision it is stated that the Ministry of Defence wants to be a strongly innovative and information-driven organization. Also, in the Sailplan 2030 (the vision document of the Navy) the importance of big data is mentioned. Using predictive maintenance instead of pre-scheduled maintenance contributes to becoming more information-driven. As a consequence, failures can be avoided and the operational availability of the vessels will be increased.

[1] Source: Royal Netherlands Navy, `https://english.defensie.nl/organisation/navy/tasks`

1.3 Problem description

In the past few years some research has already been done in the area of predictive maintenance at the Royal Netherlands Navy. The work of Heek (2021) showed that a combination of regression models and Statistical Process Control (SPC) can be used to predict failures in the main bearings of the diesel engines from the OPVs. The main bearings of the diesel engine are a critical part of the OPVs, since a failure in these bearings will cause operational unavailability. Heek was able to obtain well developed signals approximately 200 operating hours before bearing failures occurred. However, he also concluded that the model is very sensitive to maintenance. In his recommendations he writes that areas of improvement are the data sampling, the regression model and the use of multivariate SPC. Therefore, our research will focus on these topics. Since Heek already showed some promising results, we will continue with this research and use some more advanced statistical methods.

The main research question that will be studied in this thesis can be found in the block below. In order to properly answer the research question, five sub-questions are formulated.

Research questions

Main question: How can statistical process control be applied to predict failures in the main bearings of the diesel engines from the OPVs?

Sub-questions:

- Which (output) variables should be monitored to timely detect failures in the main bearings of the diesel engines?
- How can the relation between output variables and input variables of the main bearings of the diesel engines be modelled?
- What is an appropriate (multivariate) control chart to monitor the main bearings of the diesel engines making use of the relation between output and input variables?
- Can a root cause analysis be developed to determine from the control chart the cause of the change when a change in the process is detected?
- How can the Remaining Useful Life (RUL) be predicted when a change in the process is detected?

1.4 Thesis outline

In Chapter 2 the data set that is used in this research is described. The pre-processing steps are discussed and an exploratory data analysis is performed. In Chapter 3 a literature review on the topics that are discussed in this thesis is presented. We finish this chapter with the solution strategy that will be followed to find answers to the research questions. Chapter 4 describes the statistical methods that we will use in great detail. In this chapter, the multiple linear regression model and predictive and recursive estimates that can be generated using this model are described. Next, various univariate and multivariate statistical process control methods are discussed. In Chapter 5 variables that are included in the multiple linear regression model are selected. Next, the statistical process control methods are applied to the historical data set provided by the Royal Netherlands Navy. In Chapter 6 the most important conclusions are summarized in order to answer the research questions. Finally in Chapter 7 suggestions for future research are given and recommendations regarding the implementation of the models at the Royal Netherlands Navy are listed.

On top of this, the thesis contains several appendices. In Appendix A additional figures are presented. In Appendix B additional tables are provided. Some prerequisites regarding multiple linear regression are listed in Appendix C. Appendix D gives and proofs some additional lemmas and theorems. Finally, in Appendix E the most important R functions that were written for this research are listed.

Chapter 2

Data

In this chapter, various aspects of the data set that is used in this research are discussed. First, we describe the data set and the pre-processing steps that were taken to make this data set usable. Next, an exploratory data analysis is performed. In this exploratory data analysis, the focus lies on the correlations between the variables that are included in the data set. Furthermore, the influence of maintenance actions is discussed.

The data that is available for this research is the Integrated Platform Management System (IPMS) data set. This data set contains data that is used by the vessels for operational purposes, but the research of Heek (2021) already showed that this data can also be used for maintenance purposes. The IPMS data set consists of the data from approximately 700 sensors. In this thesis, a subset of the IPMS data is used. The sensors that are included in this subset are selected by the engineers of the Royal Netherlands Navy, since they know which variables are associated with the quality of the diesel engines. The subset that was selected contains data from two types of sensors: a total of 59 value sensors and 4 switch sensors. The value sensors measure quantities such as temperature and pressure from different elements of the diesel engines. This data is highly frequent, with frequencies varying between 1/3 Hz and 1 Hz. The switch sensors are used to determine the propulsion mode of the vessel. This data is saved differently than the data from the value sensors. The states of the switches are saved every hour and upon change.

Eventually, the goal is to monitor the quality of the main bearings of the diesel engine. According to Vencl and Rac (2014) the four main causes of bearing failures are abrasive wear, adhesive wear, surface fatigue wear and cavitation. In his research, Heek concluded, after gaining expert knowledge of the engineers at the Royal Netherlands Navy, that abrasive wear and cavitation are the most important ones for the diesel engines of the OPVs. Both wear and cavitation lead to an increased bearing temperature. For this reason, the bearing temperatures give an indication of the quality of the bearings. Each bearing can fail independently from the other bearings. Therefore, to get a good idea of the quality of each of the bearings, the temperature sensors at the bottom of the bearings are good sensors to monitor. In the data set, these sensors are named TEMPERATUUR.HOOFDLAGER.X.HVD.BB, where X represents the number of the bearing.

2.1 Pre-processing of the data

Before the data can be used, some pre-processing should take place. For example, the data from the switch sensors and the data from the value sensors is stored very differently. To overcome these issues, a data pipeline was designed in Python. An elaborate discussion (in Dutch) on how this pipeline was designed can be found in Teunisse (2021). This pipeline results in two ready-to-use data sets. For completeness, we give a concise overview of the steps that are taken in the pipeline.

1. Retrieve raw data sets per day for the period of interest. Select only the relevant sensors and store the files per day for both the switch and value sensors.

2. Transform the data from the switch values to the same format as the data from the value sensors, so that these can be coupled.

3. Remove periods where the diesel engine was turned off or running stationary, since in these periods no abnormal temperatures are expected. Doing this, the only remaining propulsion modes are the transit and manoeuvre mode. These are the only modes in which the ship sails on the diesel engine.

4. Add operating hours to the data set, since these contain the information on the "age" of the diesel engine.

5. Remove the data directly after a change in RPM, since the engine does not run in a stable manner during these transitions.

6. Remove data from periods when there was no load on the diesel engine, since these are periods that contain unstable behaviour.

7. Make separate data sets for each propulsion mode. This is important because the engine is used in different ways for the transit and manoeuvre mode.

8. The highly frequent daily data sets are too large to merge into one large data set. It is however crucial to have one complete data set for each propulsion mode. Since the degradation process takes place at a large time scale, the data can be resampled to one observation per minute. Doing this, the data sets can be merged to two complete (multiple year) data sets for each of the two propulsion modes.

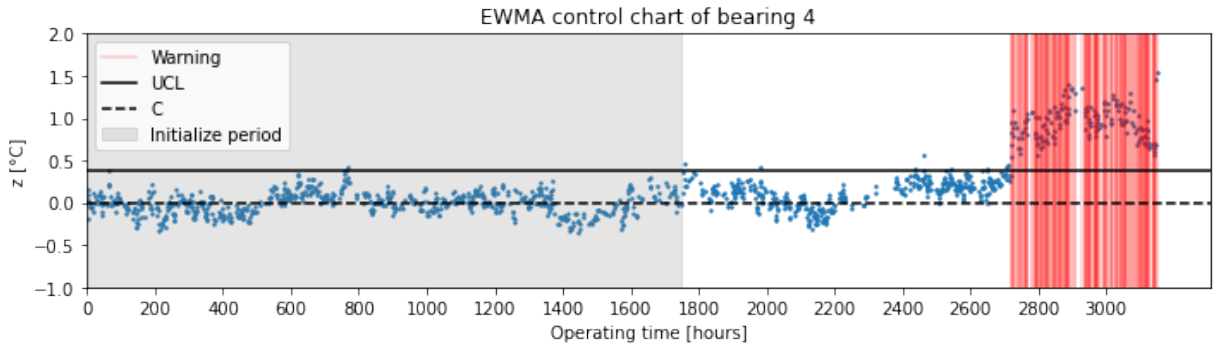
The last step of the pre-processing pipeline differs from the pipeline as described in Teunisse (2021). Therefore, it is explained in some more detail in this paragraph. To resample the data the `pandas.DataFrame.resample()` function was used. The sample frequency can be given as an argument to this function. A sample frequency of one observation per minute is used in this research. The resample function creates a new data frame with the desired sample frequency. If no data is available at some point, a row of NAs will be added to the data frame. Methods are available to impute these NAs, but we chose to delete these rows. This is because, the NAs do not represent missing data. In previous steps it was decided that this data should be removed, because it is associated to unstable behaviour. Therefore, it should be removed again after the resampling process.

2.1.1 Validation of pre-processing

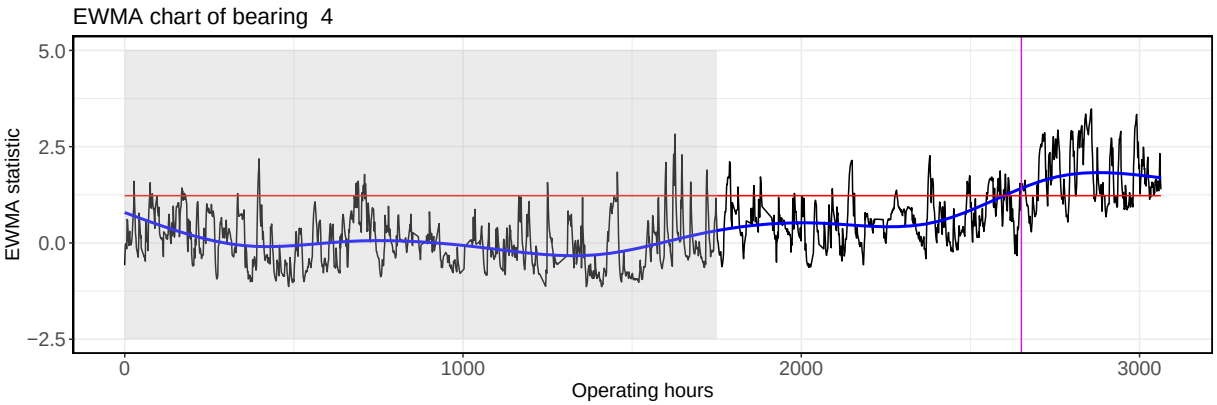
To validate the pre-processing of the data sets, we make a comparison with the results of Heek (2021). The research of Heek showed that the data contains upward trends followed by bearing failures. It is very important that this behaviour is also present in our data set, since these trends help to predict failures. By using the same regression model and the same SPC methods on the same time period, we can compare the two ways in which the data was pre-processed. Figure 2.1a shows the results of Heek in the period that he defined as case II. Figure 2.1b shows the results in the data set that was pre-processed as described above. When we compare the two figures, it can be seen that the trend in the Exponentially Weighted Moving Average (EWMA) statistic is similar. However, Figure 2.1b contains more noise than Figure 2.1a. Also, the absolute EWMA values are larger in Figure 2.1b than in Figure 2.1a. This is because Heek aggregated his data. He took averages over the hour and then used a regression model on these averages. Since this makes the exact mathematics difficult, we want to perform in this research more complicated, we did not aggregate the data, but only resampled it. This difference is clearly visible in the two figures.

From Figure 2.1a and Figure 2.1b, we can conclude that the way in which the data is pre-processed influences the results. However, the pre-processed data set that is used in this research, shows a similar trend as the data set pre-processed by Heek. This validation shows that the trend is still present in our pre-processed data set. Knowing this trend is present, we can design models to detect the trend. Therefore, we conclude that our pre-processed data sets can use in the continuation of this research.

As a side note we want to add that the results from Heek look less noisy than the results obtained by our pre-processing steps. Anyhow, Heek optimized his methods on his data set. It is to be expected that his methods are not optimal in our data set. In this research we will explore new statistical methods and optimize them on our data set. In the end, we can compare these results to the results in Heek (2021). In this way we can decide if our extended research improves the model designed by Heek.



(a) EWMA chart of case II as described in Heek (2021).



(b) EWMA chart on the same period as case II using the pre-processed data set as described in Section 2.1.

Figure 2.1: Validation of pre-processing steps.

2.2 Exploratory data analysis

As discussed, the variables of interest are the temperatures of the main bearings. In Figure 2.2, a plot is provided to give an idea of the general behaviour of these temperatures. From this plot, it can already be seen very clearly that the bearing temperatures are strongly correlated.

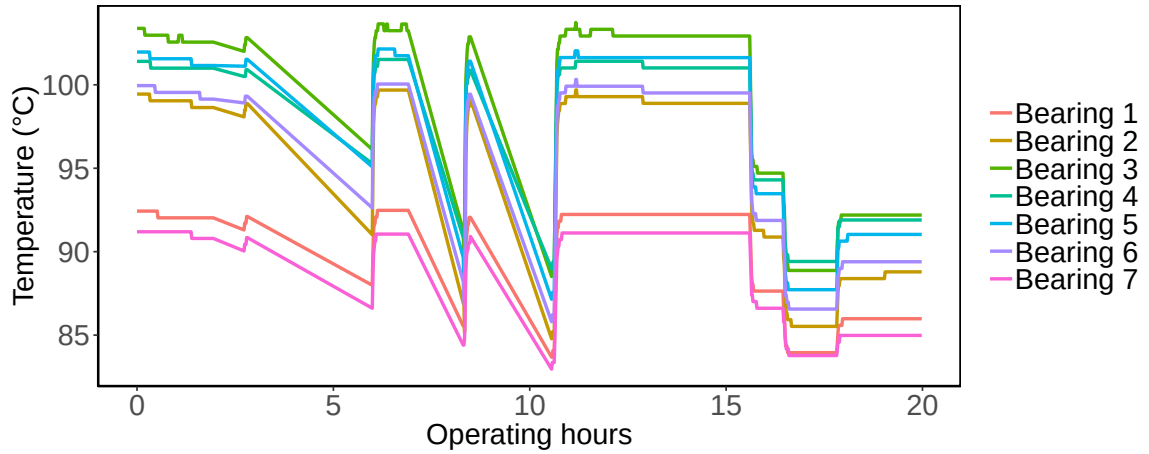


Figure 2.2: Bearing temperature over time in the transit mode.

In the previous section it is mentioned that we include two propulsion modes: the transit and the manoeuvre mode. This is because these two modes are the only modes in which the diesel engine is used to propel the ship. The two modes are used in different scenarios. Think of the manoeuvre mode as the sport mode of your car and the transit mode as the "regular" settings. Since, the engine is used in different ways by these modes,

the mode might be a factor that should be included in the model. To get a first impression on the difference between the propulsion modes a boxplot of the bearing temperatures for both modes is displayed in Figure 2.3. It can be seen that the temperatures in the transit mode are structurally higher than the temperatures in the manoeuvre mode. This shows that there is indeed a difference between the two modes. For now, it is decided to treat the data from these two modes separately. Later in this report, it will be examined if the modes should indeed be separated or that the difference between the modes can be fully explained by other factors that are included in the model.

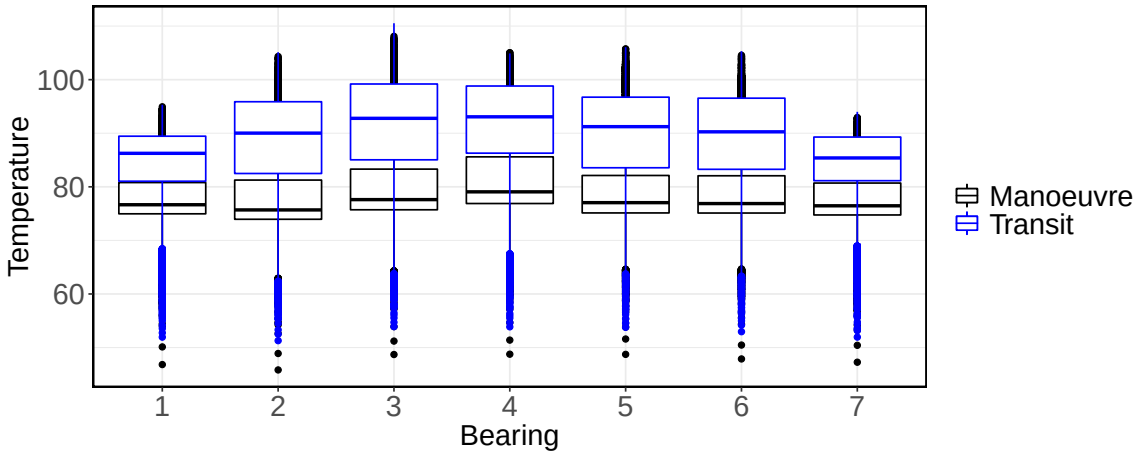


Figure 2.3: Boxplot of the bearing temperatures for the seven main bearings for both propulsion modes.

2.2.1 Correlations

The temperatures of the bearings do not only depend on the quality of the bearings. For example, a higher RPM leads to more friction which leads to a higher temperature. In Table B.3 (in the appendix) it can be seen that the correlation between the bearing temperature of the first bearing and the RPM is 0.98. This means that the bearing temperature strongly depends on the RPM. As can be seen in the table, the RPM is not the only variable that has a high correlation with the bearing temperature. When the bearing temperatures are used as a measure for the quality of the bearing, there should be corrected for other variables that influence the bearing temperature. This will be discussed in more detail later in the report.

So, the temperatures of the bearings do not only depend on the quality of the bearings, but also on other variables in the data set. These other variables, also called the predictors, depend on each other. In Figure A.1 the correlation matrix of the predictors is visualized. Here, it can be seen that many predictors have a high positive correlation. In the top right corner, a bright yellow block is visible. This block indicates very strong correlations between the all the exhaust gas temperatures. This is what you would expect, since these are measurements of the same matter, but only at different locations. The same reasoning holds for the large correlations between the oil temperatures. It is important to keep in mind that many of the predictors are highly correlated, since this could lead to problems in the variable selection process. This will be discussed later in the report.

As mentioned before, there are seven main bearings. It is interesting to investigate how similar these bearings are. If the bearings all behave the same, they could be modelled using the same model. If the bearings behave very similar, we expect high correlations between the bearing temperatures of the seven bearings. The correlations between the seven main bearings can be found in Table 2.1 and they are indeed very high. The similarity between the bearings can also be seen in Figure 2.2. However, this figure also shows that the absolute values of the temperatures vary between the bearings. From Table 2.1 and Figure 2.2 we can conclude that the variables that should be included in the models for the seven bearings could be the same, but the coefficients might be different. Also, since an important failure mechanism is cavitation, which causes each bearing to fail individually, it is really important to monitor the quality of each bearing individually.

Table 2.1: Correlation between the temperatures of the seven main bearings.

Bearing	1	2	3	4	5	6	7
1	1.00	0.99	0.99	0.99	0.99	0.99	1.00
2	0.99	1.00	1.00	0.99	0.99	0.99	0.99
3	0.99	1.00	1.00	1.00	0.99	0.99	0.99
4	0.99	0.99	1.00	1.00	0.99	0.99	0.99
5	0.99	0.99	0.99	0.99	1.00	0.99	0.99
6	0.99	0.99	0.99	0.99	0.99	1.00	0.99
7	1.00	0.99	0.99	0.99	0.99	0.99	1.00

2.2.2 The influence of maintenance

In his thesis, Heek recommended to look into the influence of maintenance actions on the model. In Figure 2.1a a control chart is given for the fourth bearing. At the end of timeline this bearing fails. It can be seen that from 2700 operating hours onward, many signals are generated by the control chart. Lay people might interpret these signals as a sign that something is wrong with this bearing. This would mean that the control chart gives signals approximately 400 operating hours before the failure occurs. However, this is not quite true. In Figure 2.1b a vertical, purple line is drawn around 2700 operating hours. This purple line represents a maintenance action that is performed on the diesel engine. While Figure 2.1a might suggest that there is a clear sign of an upcoming failure approximately 400 hours before it occurs, these signals are very likely caused by the maintenance action.

Heek suggest refitting the regression model after maintenance actions are performed. However, this would mean that a new initialization period of several hundreds of hours is needed to refit and validate the new regression model. During this initialization period, the process cannot be monitored. It appears that both failures that are present in the data set take place approximately 400 hours after a maintenance action is performed. It therefore is very important to be able to monitor the process as soon as possible after maintenance is executed. The focus of our research should lie on finding methods that do not require a long initialization period after maintenance, so that the models can actually be used in practice by the Royal Netherlands Navy.

Summary and conclusions

In this section, we discussed how the data can be pre-processed. We validated the pre-processing by using the model designed by Heek. While the results on our data set contained more noise than the results of Heek, the overall pattern was similar. From this we can conclude that the data set we created can be used during the continuation of our research.

The exploratory data analysis showed that both the temperature sensors of the main bearings and the sensors included as possible predictors are highly correlated. Furthermore, we showed that there is a significant difference between the temperature profiles used by the transit and manoeuvre modes. These are important aspects to keep in mind when we are going to model the data. Finally, we showed that maintenance affects the predictive qualities of the regression model. It is important to further investigate how to deal with maintenance actions.

Chapter 3

Literature and solution strategy

In this chapter a literature review on the topics that are related to the research questions is given. We pay attention to literature on univariate and multivariate SPC, root-cause analysis, remaining useful life predictions, the combination between SPC and linear regression, the combination between SPC and predictive maintenance and some data science techniques that will be useful for our research. Next, using this literature review, a solution strategy will be proposed. This solution strategy will help us to answer the research questions.

3.1 Literature review

Statistical process control (SPC) is a collection of methodologies to monitor the quality of a process via statistical procedures. SPC is also referred to as SQC (Statistical Quality Control) or SPM (Statistical Process Monitoring). While we think the term SPM is more accurate, the term SPC will be used in this report, since this is common practice. SPC was pioneered by Walter A. Shewhart, who invented the first control chart in the 1920s at the Bell Labs company. For more information on the history of statistical process control we refer to Miranti (2005) and Juran (1997). Shewhart developed a tool, the Shewhart chart, to distinguish the special cause variation from the common cause variation. Common cause variation is inevitable. The common cause variation can only be changed if the process itself is changed. Special cause variation is the variation that needs to be removed from the process. It can be caused by, for example, defective materials. If only common cause variation is present, the process is considered to be statistically in-control. If special cause variation is present, the process is considered to be out-of-control. After Shewhart published his famous original work (Shewhart (1931)), much progress has been made in the field of statistical process control (see Frisén (2007)). Some very well-known variations on the Shewhart chart are the Cumulative Sum (CUSUM) chart (Page (1955)) and the Exponentially Weighted Moving Average (EWMA) chart (Roberts (1959), Shiryaev (1963)). The main difference between these two charts and the Shewhart control chart is that the CUSUM and EWMA chart take the history of the process into account, while the Shewhart chart only uses the current observation. We refer to Qiu (2013), for an elaborate introduction in the field of statistical process control. This book discusses the basic concepts of SPC and describes some well-known SPC procedures.

3.1.1 SPC in predictive maintenance

While statistical process control was originally used to monitor production processes, it also proved to be useful in the area of predictive maintenance. In Linderman et al. (2005) it is claimed that jointly optimizing statistical process control and maintenance policies can minimize the total costs of quality inspection and maintenance. In this research the traditional Shewhart chart is used, but it is suggested to use EWMA charts to expand the applicability of the maintenance model to diverse environments. The claim that including statistical process control in the maintenance strategy is useful is supported by Panagiotidou and Tagaras (2010) and Mehrafrooz and Noorossana (2011).

3.1.2 SPC using regression models

An important assumption in statistical process control is that the observations that need to be monitored are independent and identically normally distributed. However, in many processes this might not be the case. If other variables influence the quality characteristic that is monitored, a correction should be made for these variables. One way to do this is using a regression model where the quality characteristic is the dependent variable and the other variables are included as predictors. The residuals that are obtained from the regression model can be used for SPC. In Brown et al. (1975) and Dufour (1982) a way to obtain these residuals is described and the distribution is given. In Van Dalen (2018) this approach is applied to a case study of wind turbines. Here, the residuals of the regression model are monitored using CUSUM charts. A distinction is made between the predictive residuals and recursive residuals. For the predictive residuals, the model needs to be fitted on a stable period, then it can be used to make predictions outside the stable period. On the other hand, one could also use a self-starting approach as described in McClurg (2016). The residuals that are obtained via this self-starting approach are referred to as recursive residuals in Brown et al. (1975). Brown proves that the recursive residuals are independent over time, while the predictive residuals are correlated.

Another well-known method in the field of control theory is the Kalman filter. In Kalman filtering mathematical or physical models are used in combination with noisy (sensor) measurements to estimate a systems state when it cannot be measured directly. This is exactly what we want to do in this research. We want to determine the quality of the bearings, which cannot be measured directly. However, regression models and sensor data can be used to estimate the quality of the bearings. The method of recursive residuals is actually a special case of a Kalman filter as is noted in Dufour (1982). For more information on the link between the recursive residuals method and Kalman filters we refer to Pollock (2003).

Instead of monitoring the residuals of a regression model, one could also monitor the coefficients of the regression model. This idea was introduced in Dufour (1982). Dufour describes a method to estimate the regression coefficients recursively. He also gives the distribution of the difference between two successive coefficient estimates. He claims that these differences are independent and identically normally distributed, which makes them appropriate to use for statistical process control. However, Dufour does not make the link to SPC in his article. He tests the stability of the regression coefficients using hypothesis testing.

3.1.3 Data science methods for SPC

As described in the previous section, regression models can be a useful tool to use in combination with SPC. However, before the regression models can be used, methods to select the variables that are included in the model must be applied. Standard variable selection methods such as forward selection, backward selection and stepwise selection are described in James et al. (2014). However, these methods are not very useful when the data set is very large. In Tibshirani (1996) the Least Absolute Shrinkage and Selection Operator (LASSO) variable selection method is introduced. This method is particularly useful when the number of available variables is large. LASSO variable selection and other variable selection methods for SPC are discussed in Capizzi (2015) and Capizzi and Masarotto (2015). Once the variables are selected and the regression model is fitted on the data, this model needs to be validated. James et al. (2014) describes several cross validation techniques that can be used to this end.

The data we use in this research can be considered as “big data”. The number of variables is large and also the number of measurements is very large. We refer to Megahed and Jones-Farmer (2015) for a discussion on how big data influences statistical process control.

3.1.4 Multivariate statistical process control

The literature we discussed so far focuses on univariate statistical process control. However, the temperatures of the seven main bearings are highly correlated. Univariate approaches do not take this correlation into account. Therefore, multivariate approaches should also be considered. In 1947, Hotelling proposed a multivariate extension of the Shewhart chart (Hotelling (1947)). Later, various multivariate extensions of the CUSUM and EWMA charts were published. In Woodall and Montgomery (1999), which is a discussion paper on statistical process control, it is stated that multivariate SPC is the region with the highest activity level in SPC. Many papers are published in this region, which makes it hard to keep track of all new developments. We refer to

Wierda (1994) for an overview on multivariate statistical process control methods and directions for future research.

One paper on multivariate SPC that has our special interest is the paper of Hawkins (1991). Hawkins has a rather different view on using regression analysis in multivariate statistical process control. In this procedure, called Regression Adjusted Variables (RAV), Hawkins suggest regressing each quality characteristic on all the other quality characteristics, which results in a Z-statistic. In this way, the quality characteristic is monitored with respect to the other characteristics. This method could be very interesting in our case, since we would like to detect (among other things) when one of the bearings starts behaving different than the others. In this paper, Hawkins uses CUSUM charts to monitor the individual Z-statistics and suggest two metrics to combine the Z-statistics, so that they can be monitored via a grouped chart.

When using a multivariate monitoring approach, it is important to be able to answer the question: "If the process is out of control, what is the problem?" Answering this question is also referred to as performing a root cause analysis or fault detection/diagnosis. In their overview paper on the interpretation of multivariate statistical process control procedures, Bersimis et al. (2005) state that performing a root cause analysis is the most challenging problem in multivariate SPC. However, when using the RAV procedure, fault detection can be performed in a rather simple way. When a signal appears in the grouped chart, one can look at the individual charts to see which variable is causing the signal. The quality characteristics that contain signals in the individual charts are then the cause(s) of the signal in the grouped chart.

For charts based on Hotelling's T^2 statistic, performing a root cause analysis takes a little more effort. Over the years many methodologies have been proposed. Principal Component Analysis (PCA) techniques are commonly used. However, these are often hard to interpret. In Mason and Tracy (1995) a fault detection method based on orthogonal decompositions of the T^2 statistic, called the Mason–Young–Tracy (MYT) decomposition, is proposed. This method leads to a more direct interpretation of the out-of-control signals. In the MYT decomposition, the T^2 statistic is decomposed in conditional and unconditional terms. When variables have a large unconditional term, this means that this single variable is outside the control limits. When variables have large conditional terms, this means that that there is something wrong with the relations between the variables included in this term. In this case, all variables included in this term need to be examined in more detail.

In Liu et al. (2006) the Adaptive Regression Adjusted (ARA) scheme is introduced which can be used as an alternative to the PCA based methods. The ARA scheme makes use of regression adjusted variables as described earlier in this section. In this scheme the variables that are assumed to be in-control are regressed on the variables that are assumed to be out-of-control. Regression adjusted in-control variables are then monitored using a T^2 statistic. The paper concludes that fault detection via the ARA scheme, in general, outperforms PCA methods.

We discussed the well-known MYT method and the ARA scheme, which is an improvement om the MYT method that uses regression adjusted variables. However, there are numerous methods that we did not discuss. For an overview on various fault diagnosis methods, we refer to Vidal-Puig and Ferrer (2013). In this discussion paper, different fault diagnosis techniques in multivariate quality control are described and tested in a simulation study on different scenarios. An important conclusion of this study is that most methods have problems with false positives that are often not reported in the original papers.

3.1.5 Remaining useful life estimation

Another aspect of statistical process control we need to discuss is the determination of the Remaining Useful Life (RUL). The RUL is defined as the length from the current time to the end of the useful life. There are various approaches to RUL prediction. Lei et al. (2018) categorizes these different approaches into four categories: Physics model-based approaches, statistical model-based approaches, AI approaches and hybrid approaches. Physical model-based approaches describe the degradation process using mathematical models based on the failure mechanisms. Using these kinds of models is out of the scope of this thesis and will not be discussed in more detail. AI approaches use black-box machine learning algorithms to learn the degradation pattern, these methods are hard to explain because of a lack of transparency. Therefore, they will not be used in this research. Statistical RUL prediction models are constructed by fitting random coefficient models or stochastic process models on the available data. These kinds of models are discussed in great detail in Si et al. (2017). This category of RUL prediction methodologies seems to be the most feasible within the context of this research.

3.1.6 SPC applied to the main bearings of diesel engines

Finally, we would like to discuss some works were SPC is applied to achieve predictive maintenance on the main bearings of diesel engines. In Salomé (2019) the RUL of bearings is studied at Damen, which is a Dutch shipbuilding company that also builds ships for the Royal Netherlands Navy. Salomé uses statistical process control on vibration data to predict the RUL. Salomé used self-starting Shewhart and EWMA charts in his research. The main reason for selecting a self-starting approach was that the change points presented themselves shortly after the start of the measurements. Therefore, phase I data could not be collected. The recursive residuals we discussed earlier are an example of a self-starting approach. Since, the context of our research is very similar to the research of Salomé, these recursive residuals will be explored in great detail.

In Heek (2021) a first attempt was made to apply statistical process control methods to the main bearings of diesel engines at the Royal Netherlands Navy. Heek used the IPMS data set to obtain a condition-based monitoring approach for the main bearings of the diesel engines. Using multiple linear regression and EWMA charts on the residuals, he was able to predict failures approximately 200 operating hours before they occurred. An important conclusion of his work is that the model is sensitive to maintenance actions, since these actions influence the physical relations in the diesel engine. The self-starting approach using recursive residuals might be a solution for this sensitivity, because the model can be reset and started again after maintenance. The regression adjusted variables approach might also be useful, since it compares one bearing with respect to the others. The physical relationships change for all bearings, which is why the RAV methods is likely to be less sensitive to such changes.

3.2 Proposed solution strategy

Using the literature review in the previous section, we came up with the proposed solution strategy as visualized in Figure 3.1. The starting point will be a subset of the IPMS data set that was pre-processed according to the steps described in Section 2.1. Next, a variable selection method must be performed to select the variables that will be included in the MLR model. When the MLR model has been fitted and validated on a stable period, several routes can be taken.

The first route is the route of the predictive residuals. The MLR model will be fitted to a stable period and this model will be used to make predictions outside the stable period. These predictions will be subtracted from the actual measurements which results in the predictive residuals. The predictive residuals will be monitored using univariate Shewhart and EWMA charts and the multivariate T^2 and MEWMA charts. Alternative to the predictive residuals, the residuals could also be calculated recursively. At time point i, the MLR model will be fitted on the data from time points $1, \ldots, i-1$. This model will then be used to make a prediction for time point i. The predictions are subtracted from the actual measurements and this is how the recursive residuals are obtained. The recursive residuals can be monitored in the same way as the predictive residuals. Since the recursive and predictive residuals will probably both have their strengths and weaknesses, a hybrid method will be investigated. In the hybrid method, the recursive residuals will be used to start up the model. When the model is good enough a switch to the predictive residuals is made. The residuals obtained via the hybrid method can be monitored in the same way as the predictive and recursive residuals.

The third route is the route of the recursive coefficients. At time point i the MLR model will be fitted to the data from time points $1, \ldots, i$. The coefficient vectors obtained at time points $i-1$ and i are subtracted. The difference of the recursive coefficient vectors will be monitored using multivariate SPC via T^2 and MEWMA charts.

Finally, we would like to investigate the RAV method as described in Section 3.1.4, which is also a multivariate method. Because the raw data does not meet the assumptions that are required by this method, the recursive residuals will be used as input. The recursive residuals of each bearing are regressed on the recursive residuals of the other bearings, from which we can derive the Z-statistics. The Z-statistics can be monitored using CUSUM charts. On top of that, metrics are developed to combine the individual Z-statistics to a grouped statistic. The grouped statistics will also be monitored using CUSUM charts.

When the (multivariate) control charts are retrieved a root cause analysis should be performed to determine the cause of the signal. In our case, this is important, since we need to know which bearings need to be inspected and maintained. Ideally, we would like to implement some of the RUL prediction methods as described. However,

there is a limited amount of failure data. This makes the RUL predictions via standard methodologies unreliable. More failure data (possibly from other ships) should be collected before a RUL analysis can be performed. This is infeasible within the time limits of this research. Therefore, this is left open for future research.

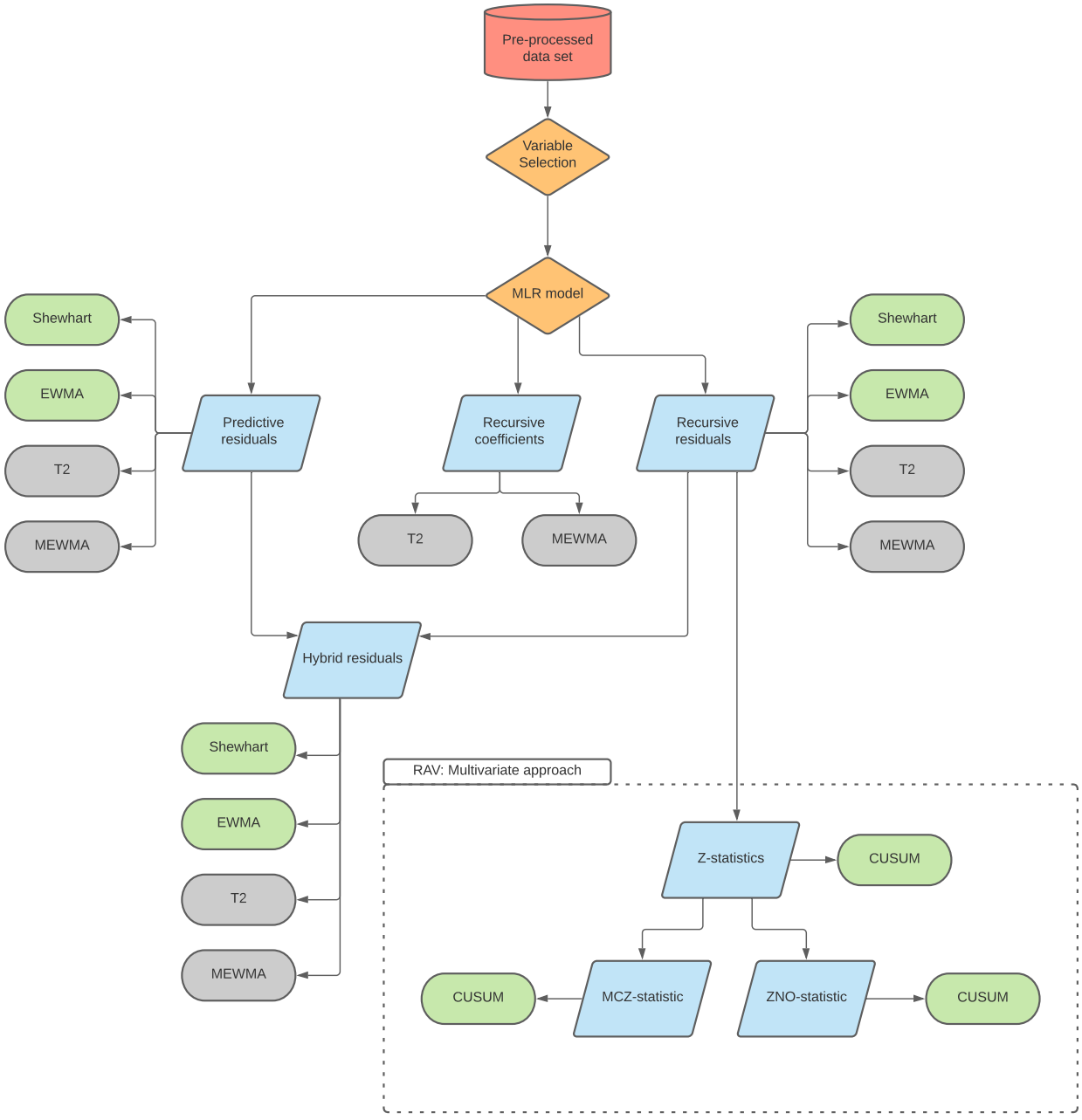


Figure 3.1: Visual representation of the proposed solution strategy. Eventually, we were unable to implement the grey parts of this proposed strategy.

Chapter 4

Statistical methods

The goal of this chapter is to describe the various statistical methods that are included in the proposed solution strategy as described in Section 3.2. First the multiple linear regression model will be described and the recursive and predictive estimates that can be derived from this model are discussed extensively. Then, the LASSO variable selection method is described. Finally, several univariate and multivariate statistical process control methods are discussed. A general description is provided and we discuss how the methods can be implemented on our data set specifically.

4.1 Multiple linear regression

In this section the Multiple Linear Regression (MLR) model is described. Next, some predictive and recursive estimates based on the MLR model will be discussed. The focus lies on the distributions of these estimates. To describe the basic concepts of the MLR model, the book of Bingham and Fry (2010) was used. The description of the predictive and recursive estimates is based on the work of Van Dalen (2018), Brown et al. (1975) and Dufour (1982).

4.1.1 The MLR model

Assume that during the in-control period n observations $(Y_1, \ldots, Y_n)$ are collected that come from a normal distribution. During the same period also n realizations of the p predictors are observed. These are also called the dependent variables. The multiple linear regression model is defined as follows.

Definition 4.1.1. *The multiple linear regression model for the real-valued observations Y_i $(i = 1, 2, \ldots)$ is defined by*

$$Y_i = X_{i\bullet}\beta + \varepsilon_i, \tag{4.1}$$

where Y_i is the ith observation in time, $X_{i\bullet} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ contains the observed values of the predictors at time i, $\beta = (\beta_0, \ldots, \beta_p)^T$ is the coefficient vector and the ε_i's are the errors for observation i, which are assumed to be i.i.d. $\mathcal{N}(0, \sigma^2)$ for $i = 1, 2, \ldots$.

Let $(Y_1, Y_2, \ldots, Y_n)$ be n observations that are collected during a stable (in-control) period, then the following model can be defined:

$$Y_{[1:n]} = X\beta + \varepsilon_{[1:n]}, \tag{4.2}$$

where $Y_{[1:n]} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ and $\varepsilon_{[1:n]} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim \mathcal{N}(0_n, \sigma^2 I_n)$.

In this model X is called the $(n \times (p+1))$-design matrix. From the normality and independence of the errors it follows that the $Y_{[1:n]}$ are also independent and $\mathcal{N}(X\beta, \sigma^2 I_n)$ distributed. In this model, β and σ^2 are unknown

and need to be estimated. The coefficient vector β is estimated by solving Gauss's normal equations. When X is assumed to be non-singular this leads to the estimate

$$\hat{\beta} = (X^T X)^{-1} X^T Y_{[1:n]}, \tag{4.3}$$

which follows a $\mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$ distribution. The variance of the errors σ^2 can be estimated as

$$\hat{\sigma}^2 = \frac{SSE}{n}, \tag{4.4}$$

where SSE is the sum of squares for errors. The estimate $\hat{\sigma}^2$ follows an $n\sigma^2 \chi^2_{n-p-1)}$ distribution. More details on the derivations and distributions of these estimates can be found in Appendix C.

Until now, we have seen what the properties of the multiple linear regression model are when it is applied to an in-control period. However, we would like to know what happens if the model is used to predict observations outside the stable (in-control) period. Various ways to use the model outside the stable period will be discussed in the remainder of this section.

4.1.2 Predictive residuals

Let us consider the observations $(Y_1, Y_2, \ldots)$. Assume that the observations $(Y_1, \ldots, Y_n)$ are collected during a stable period. After this stable period the new observations $(Y_{(n+1)}, Y_{(n+2)}, \ldots)$ are collected. The multiple linear regression model as defined in (4.1) can be used to predict the value of a new observation $Y_{(n+i)}$ (for $(i = 1, 2, \ldots)$) in the following way:

$$\hat{Y}_{(n+i)} = X_{(n+i)\bullet} \hat{\beta}, \tag{4.5}$$

where $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE) as defined in Formula 4.3. This can be rewritten as

$$\begin{aligned} \hat{Y}_{(n+i)} &= X_{(n+i)\bullet} \hat{\beta} \\ &= X_{(n+i)\bullet} (X^T X)^{-1} X^T Y_{[1:n]} \\ &= P_{(n+i)} Y_{[1:n]}, \end{aligned} \tag{4.6}$$

where $P_{(n+i)} = X_{(n+i)\bullet} (X^T X)^{-1} X^T$. So, $\hat{Y}_{(n+i)}$ are the predictions of the new observations $Y_{(n+i)}$. These predictions will be called the fitted values.

Note that the distribution of $\hat{\beta}$ is $\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X^T X)^{-1}\right)$. Making use of Lemma D.0.1 it follows that the distribution of $\hat{Y}_{(n+i)}$ is

$$\hat{Y}_{(n+i)} \sim \mathcal{N}(X_{(n+i)\bullet} \beta, \sigma^2 X_{(n+i)\bullet} (X^T X)^{-1} X^T_{(n+i)\bullet}). \tag{4.7}$$

If the fitted values, $\hat{Y}_{(n+i)}$, are compared with the actual new observations, $Y_{(n+i)}$, we obtain the residuals $e_{(n+i)}$. The residuals are defined as

$$\begin{aligned} e_{(n+i)} &= Y_{(n+i)} - \hat{Y}_{(n+i)} \\ &= Y_{(n+i)} - P_{(n+i)} Y_{[1:n]} \\ &= \begin{pmatrix} -P_{(n+i)} & 1 \end{pmatrix} \begin{pmatrix} Y_{[1:n]} & Y_{(n+i)} \end{pmatrix}^T \\ &= A \begin{pmatrix} Y_{[1:n]} & Y_{(n+i)} \end{pmatrix}^T. \end{aligned} \tag{4.8}$$

Since any linear combination of independent univariate normal random variables follows a multivariate normal distribution, it follows that

$$\begin{pmatrix} Y_{[1:n]} \\ Y_{(n+i)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} X\beta \\ X_{(n+i)\bullet} \beta \end{pmatrix}, \sigma^2 I \right). \tag{4.9}$$

Because of Lemma D.0.1 it holds that

$$\begin{aligned} e_{(n+i)} &\sim \mathcal{N}\left(A\begin{pmatrix} X\beta \\ X_{(n+i)\bullet}\beta \end{pmatrix}, A\sigma^2 I A^T\right) \\ &\sim \mathcal{N}\left(0, \sigma^2\left(X_{(n+i)\bullet}(X^TX)^{-1}X_{(n+i)\bullet}^T + 1\right)\right). \end{aligned} \tag{4.10}$$

Elaborate calculations that lead to (4.10) can be found in Van Dalen (2018).

The statistical process control methods that will be used later, assume that the monitored observations are independent. In our case, the observations that will be monitored are the residuals. Therefore, it is important to know how the residuals relate to each other. The predictive residuals are likely to be dependent, since they are all determined using the same regression model. To check if this really is the case, we need to know the joint distribution of $e_{(n+i)}$ and $e_{(n+j)}$ for $i \neq j$. Let us write

$$\begin{pmatrix} e_{(n+i)} \\ e_{(n+j)} \end{pmatrix} = \begin{pmatrix} -P_{(n+i)} & 1 & 0 \\ -P_{(n+j)} & 0 & 1 \end{pmatrix}\begin{pmatrix} Y_{[1:n]} \\ Y_{(n+i)} \\ Y_{(n+j)} \end{pmatrix} = B\begin{pmatrix} Y_{[1:n]} \\ Y_{(n+i)} \\ Y_{(n+j)} \end{pmatrix}. \tag{4.11}$$

Following the same arguments as for the distribution of $\begin{pmatrix} Y_{[1:n]} & Y_{(n+i)} \end{pmatrix}^T$, it holds that

$$\begin{pmatrix} Y_{[1:n]} \\ Y_{(n+i)} \\ Y_{(n+j)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} X\beta \\ X_{(n+i)\bullet}\beta \\ X_{(n+j)\bullet}\beta \end{pmatrix}, \sigma^2 I\right). \tag{4.12}$$

Because of Lemma D.0.1 it follows that

$$\begin{aligned} \begin{pmatrix} e_{(n+i)} \\ e_{(n+j)} \end{pmatrix} &\sim \mathcal{N}\left(B\begin{pmatrix} X\beta \\ X_{(n+i)\bullet}\beta \\ X_{(n+j)\bullet}\beta \end{pmatrix}, B\sigma^2 I B^T\right) \\ &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2\begin{pmatrix} X_{(n+i)\bullet}(X^TX)^{-1}X_{(n+i)\bullet}^T + 1 & X_{(n+i)\bullet}(X^TX)^{-1}X_{(n+j)\bullet}^T \\ X_{(n+i)\bullet}(X^TX)^{-1}X_{(n+j)\bullet}^T & X_{(n+j)\bullet}(X^TX)^{-1}X_{(n+j)\bullet}^T + 1 \end{pmatrix}\right). \end{aligned} \tag{4.13}$$

From this it can be concluded that the predictive residuals are correlated and that it depends on X how large this correlation is. The problem of monitoring correlated variables is that the standard deviation will be underestimated. Because of this, the control limits will be too small and many false alarms will be generated. Also, other computations, such as the computation of the average in-control run length, are no longer valid when the independence assumption is violated.

Until this point, we considered σ to be known, but this is usually not the case. Lemma 4.1.1 gives the distribution of $e_{(n+i)}$, when σ is unknown. But first, we give the definition of the student's t-distribution.

Definition 4.1.2. *Let Z follow a standard normal distribution, $Z \sim \mathcal{N}(0,1)$, and let U be chi-squared distributed with df degrees of freedom, $U \sim \chi^2_{(df)}$. Given that Z and U are independent, the random variable*

$$\frac{Z}{\sqrt{U/df}} \tag{4.14}$$

follows a student's t-distribution with df degrees of freedom.

Lemma 4.1.1. *Let $e_{(n+i)}$ for $(i = 1, 2, \ldots)$ be the predictive residuals as defined in (4.8) estimated by the multiple linear regression model as described in Definition 4.1.1. Then*

$$e_{(n+i)} \sim \left[\frac{\hat{\sigma}\sqrt{n\left(1 + X_{(n+i)\bullet}(X^TX)^{-1}X_{(n+i)\bullet}^T\right)}}{\sqrt{n-p-1}}\right] t_{(n-p-1)}, \tag{4.15}$$

where $\hat{\sigma} = \sqrt{SSE/n}$ and $t_{(n-p-1)}$ is a random variable which follows a student's t-distribution with $n-p-1$ degrees of freedom.

Proof. First, we bring the coefficient before $t_{(n-p-1)}$ to the other side of the equation. Then the left hand side of (4.15) can be rewritten as

$$\begin{aligned}\frac{e_{(n+i)}\sqrt{n-p-1}}{\hat{\sigma}\sqrt{n\left(1+X_{(n+i)\bullet}(X^TX)^{-1}X^T_{(n+i)\bullet}\right)}} &= \frac{e_{(n+i)}}{\sigma\sqrt{1+X_{(n+i)\bullet}(X^TX)^{-1}X^T_{(n+i)\bullet}}}\cdot\frac{\sigma\sqrt{n-p-1}}{\hat{\sigma}\sqrt{n}} \\ &= \frac{\frac{e_{(n+i)}}{\sigma\sqrt{1+X_{(n+i)\bullet}(X^TX)^{-1}X^T_{(n+i)\bullet}}}}{\frac{\hat{\sigma}\sqrt{n}}{\sigma\sqrt{n-p-1}}}. \end{aligned} \tag{4.16}$$

From (4.10) we know that

$$\frac{e_{(n+i)}}{\sigma\sqrt{1+X_{(n+i)\bullet}(X^TX)^{-1}X^T_{(n+i)\bullet}}} \sim \mathcal{N}(0,1). \tag{4.17}$$

Theorem D.0.8 tells us that

$$\frac{\hat{\sigma}\sqrt{n}}{\sigma\sqrt{n-p-1}} \sim \sqrt{\frac{\chi^2_{(n-p-1)}}{n-p-1}},$$

which is the square root of a random variable that follows a chi-squared distribution divided by its degrees of freedom. What rests us, is to show that the numerator and denominator of (4.16) are independent. To this end, it is sufficient to show that the residuals $e_{(n+i)}$ and the SSE are independent. To show that SSE and $e_{(n+i)}$ are independent we will show that $e_{(n+i)}$ is independent of $Y_{[1:n]} - X\hat{\beta}$. Since SSE is a function of $Y_{[1:n]} - X\hat{\beta}$, it then automatically follows that $e_{(n+i)}$ is independent of SSE. The j-th component of $Y_{[1:n]} - X\hat{\beta}$ can be written as $Y_j - X_{j\bullet}\hat{\beta}$. For all $j = \{1,\dots,n\}$ we have that

$$\begin{aligned}\mathrm{Cov}(e_{(n+i)}, Y_j - X_{j\bullet}\hat{\beta}) &= \mathrm{Cov}(Y_{(n+i)}, Y_j - X_{j\bullet}\hat{\beta}) - \mathrm{Cov}(X_{(n+i)\bullet}\hat{\beta}, Y_j - X_{j\bullet}\hat{\beta}) \\ &= \mathrm{Cov}(Y_{(n+i)}, Y_j) - \mathrm{Cov}(Y_{(n+i)}, X_{j\bullet}\hat{\beta}) - \mathrm{Cov}(X_{n+i)\bullet}\hat{\beta}, Y_j) + \mathrm{Cov}(X_{(n+i)\bullet}\hat{\beta}, X_{j\bullet}\hat{\beta}).\end{aligned} \tag{4.18}$$

The first term of this expression is equal to zero, since $j \in \{1,\dots,n\}$. When we substitute our expression for $\hat{\beta}$ in the second and third term we obtain

$$\begin{aligned}\mathrm{Cov}(e_{(n+i)}, Y_j - X_{j\bullet}\hat{\beta}) &= -\mathrm{Cov}(Y_{(n+i)}, X_{j\bullet}(X^TX)^{-1}X^TY_{[1:n]}) - \mathrm{Cov}(X_{(n+i)\bullet}(X^TX)^{-1}X^TY_{[1:n]}, Y_j) \\ &\quad + \mathrm{Cov}(X_{(n+i)\bullet}\hat{\beta}, X_{j\bullet}\hat{\beta}).\end{aligned} \tag{4.19}$$

Now, define

$$a^T = X_{(n+i)\bullet}(X^TX)^{-1}X^T \in \mathbb{R}^n, \quad b^T = X_{j\bullet}(X^TX)^{-1}X^T \in \mathbb{R}^n \tag{4.20}$$

Using this, we obtain

$$\begin{aligned}\mathrm{Cov}(e_{(n+i)}, Y_j - X_{j\bullet}\hat{\beta}) &= -\mathrm{Cov}(Y_{(n+i)}, b^TY_{[1:n]}) - \mathrm{Cov}(a^TY_{[1:n]}, Y_j) + \mathrm{Cov}(X_{(n+i)\bullet}\hat{\beta}, X_{j\bullet}\hat{\beta}) \\ &= -\mathrm{Cov}(Y_{(n+i)}, \textstyle\sum_{k=1}^n b_kY_k) - \mathrm{Cov}(\sum_{k=1}^n a_kY_k, Y_j) + \mathrm{Cov}(X_{(n+i)\bullet}\hat{\beta}, X_{j\bullet}\hat{\beta}).\end{aligned} \tag{4.21}$$

Now, it can be seen that the first term vanishes for the same reasons as before. For the second term, we only keep the jth term. When we apply the properties of the covariance we obtain

$$\begin{aligned}\mathrm{Cov}(e_{(n+i)}, Y_j - X_{j\bullet}\hat{\beta}) &= -a_j\mathrm{Cov}(Y_j, Y_j) + X_{(n+i)\bullet}\mathrm{Cov}(\hat{\beta}, \hat{\beta})X^T_{j\bullet} \\ &= -X_{(n+i)\bullet}(X^TX)^{-1}X^T_{j\bullet}\sigma^2 + X_{(n+i)\bullet}\sigma^2(X^TX)^{-1}X^T_{j\bullet} \\ &= 0.\end{aligned} \tag{4.22}$$

This result holds for all $j \in \{1,\dots,n\}$, so $e_{(n+i)}$ is independent of $Y_{[1:n]} - X\hat{\beta}$ and thus $e_{(n+i)}$ independent of SSE. This completes the proof. □

4.1.3 Recursive residuals

In the previous section the predictive residuals are discussed. Formula (4.13) shows that these residuals are correlated. This means that an important assumption in statistical process control is violated. In Brown et al. (1975), the recursive residuals are introduced. These recursive residuals are uncorrelated with zero mean and constant variance. Therefore, they meet the statistical process control assumption. The recursive residuals are considered to be self-starting which means that a stable (phase I) period is not necessary.

The idea of the recursive residuals is to re-estimate the multiple linear regression model each time a new observation occurs. Then the updated model is used to predict the next observation of the quality characteristic. Next, the predicted outcome is compared to the observed outcome. So, at time n the data up until time $n-1$ is used to fit a multiple linear regression model. This model is used to determine $\hat{Y}_n$. Next, the fitted value can be subtracted from the observed value, $Y_n - \hat{Y}_n$. The standardized version of this difference is what we call the recursive residual. This idea is written down more formally in the following definition.

Definition 4.1.3. *Consider the multiple linear regression model as defined in Definition 4.1.1. Let $\hat{\beta}_{(n)}$ be the least squares estimate of β based on the first n observations, i.e. $\hat{\beta}_{(n)} = (X^T_{[1:n]\bullet}X_{[1:n]\bullet})^{-1}X^T_{[1:n]\bullet}Y_{[1:n]}$, where the matrix $(X^T_{[1:n]\bullet}X_{[1:n]\bullet})$ is assumed to be non-singular. Then the recursive residual of observation n is defined as*

$$w_n = \frac{Y_n - X_{n\bullet}\hat{\beta}_{(n-1)}}{\sqrt{\left(1 + X_{n\bullet}(X^T_{[1:n-1]\bullet}X_{[1:n-1]\bullet})^{-1}X^T_{n\bullet}\right)}}. \tag{4.23}$$

The following lemma gives some properties that make the recursive residuals very interesting.

Lemma 4.1.2. *Consider the multiple linear regression model as defined in Definition 4.1.1. Let $\hat{\beta}_n$ be the least squares estimate of β based on the first n observations, i.e. $\hat{\beta}_n = (X^T_{[1:n]\bullet}X_{[1:n]\bullet})^{-1}X^T_{[1:n]\bullet}Y_{[1:n]}$, where it is assumed that the matrix $(X^T_{[1:n]\bullet}X_{[1:n]\bullet})$ is non-singular. Then the recursive residuals w_n as defined in Formula (4.23) are i.i.d $\mathcal{N}\left(0, \sigma^2\right)$.*

Proof. See Brown et al. (1975). □

So, the recursive residuals as defined in (4.23) are all independent and identically distributed. This makes them much more appropriate to use for statistical process control than the predictive residuals as defined earlier. Although σ is assumed to be known, it is usually unknown. If this is the case, then the recursive residuals are distributed according to Lemma 4.1.3.

Lemma 4.1.3. *Consider the multiple linear regression model as defined in Definition 4.1.1 and let w_n be the recursive residual as defined in (4.23). Then,*

$$w_n \sim \left[\frac{\hat{\sigma}\sqrt{n}}{\sqrt{n-p-1}}\right] t_{(n-p-1)}, \tag{4.24}$$

where $t_{(n-p-1)}$ is a random variable that follows a student's t-distribution with $n-p-1$ degrees of freedom and $\hat{\sigma} = \sqrt{SSE_{n-1}/(n-1)}$. In this lemma SSE_{n-1} is defined as

$$SSE_{n-1} = (Y_{[1:n-1]} - X_{[1:n-1]}\hat{\beta}_{(n-1)})^T(Y_{[1:n]} - X_{[1:n-1]}\hat{\beta}_{(n-1)}). \tag{4.25}$$

Proof. This proof is similar to the proof of Lemma 4.1.1, except for some small details. For completeness we write it out. First observe that

$$\frac{w_n\sqrt{n-p-1}}{\hat{\sigma}\sqrt{n}} = \frac{w_n}{\sigma} \cdot \frac{\sigma\sqrt{n-p-1}}{\hat{\sigma}\sqrt{n}} = \frac{w_n/\sigma}{\frac{\hat{\sigma}\sqrt{n}}{\sigma\sqrt{n-p-1}}}. \tag{4.26}$$

From Lemma 4.1.2 it follows that $w_n/\sigma \sim \mathcal{N}(0,1)$. Just as in the proof of Lemma 4.1.1 we have that

$$\frac{\hat{\sigma}\sqrt{n}}{\sigma\sqrt{n-p-1}} \sim \sqrt{\frac{\chi^2_{(n-p-1)}}{n-p-1}}.$$

What rests is to show that w_n and SSE_{n-1} are independent. This corresponds to showing that $Y_n - X_{n\bullet}\hat{\beta}_{(n-1)}$ and SSE_{n-1} are independent. Since, the functions of two independent random variables are also independent, it is sufficient to show that $Y_n - X_{n\bullet}\hat{\beta}_{(n-1)}$ and $Y_{[1:n-1]} - X_{[1:n-1]}\hat{\beta}_{(n-1)}$ are independent. For $j \in \{1, \ldots, n-1\}$ we have that

$$
\begin{aligned}
&\mathrm{Cov}(Y_n - X_{n\bullet}\hat{\beta}_{(n-1)}, Y_j - X_{j\bullet}\hat{\beta}_{(n-1)}) = \\
&\mathrm{Cov}(Y_n, Y_j) - \mathrm{Cov}(X_{n\bullet}\hat{\beta}_{(n-1)}, Y_j) + \mathrm{Cov}(X_{n\bullet}\hat{\beta}_{(n-1)}, X_{j\bullet}\hat{\beta}_{(n-1)}) - \mathrm{Cov}(Y_n, X_{j\bullet}\hat{\beta}_{(n-1)}) = \\
&\mathrm{Cov}(X_{n\bullet}\hat{\beta}_{(n-1)}, X_{j\bullet}\hat{\beta}_{(n-1)}) - \mathrm{Cov}(X_{n\bullet}(X^T_{[1:n-1]}X_{[1:n-1]})^{-1}X^T_{[1:n-1]}Y_{[1:n-1]}, Y_j) \\
&\quad - \mathrm{Cov}(Y_n, X_{j\bullet}(X^T_{[1:n-1]}X_{[1:n-1]})^{-1}X^T_{[1:n-1]}Y_{[1:n-1]}).
\end{aligned}
$$

Now, define

$$
a^T = X_{n\bullet}(X^T_{[1:n-1]}X_{[1:n-1]})^{-1}X^T_{[1:n-1]} \in \mathbb{R}^{n-1}, \quad b^T = X_{j\bullet}(X^T_{[1:n-1]}X_{[1:n-1]})^{-1}X^T_{[1:n-1]} \in \mathbb{R}^{n-1}.
$$

Then

$$
\begin{aligned}
&\mathrm{Cov}(Y_n - X_{n\bullet}\hat{\beta}_{(n-1)}, Y_j - X_{j\bullet}\hat{\beta}_{(n-1)}) = \\
&\mathrm{Cov}(X_{n\bullet}\hat{\beta}_{(n-1)}, X_{j\bullet}\hat{\beta}_{(n-1)}) - \mathrm{Cov}(a^T Y_{[1:n-1]}, Y_j) - \mathrm{Cov}(Y_n, b^T Y_{[1:n-1]}) = \\
&\mathrm{Cov}(X_{n\bullet}\hat{\beta}_{(n-1)}, X_{j\bullet}\hat{\beta}_{(n-1)}) - \mathrm{Cov}(\textstyle\sum_{k=1}^{n-1} a_k Y_k, Y_j) - \mathrm{Cov}(Y_n, \sum_{k=1}^{n-1} b_k Y_k) = \\
&\mathrm{Cov}(X_{n\bullet}\hat{\beta}_{(n-1)}, X_{j\bullet}\hat{\beta}_{(n-1)}) - \mathrm{Cov}(a_j Y_j, Y_j) = \\
&X_{n\bullet}\mathrm{Cov}(\hat{\beta}_{(n-1)}, \hat{\beta}_{(n-1)})X^T_{j\bullet} - a_j \mathrm{Cov}(Y_j, Y_j) = \\
&X_{n\bullet}(X^T_{[1:n-1]}X_{[1:n-1]})^{-1}X^T_{j\bullet}\sigma^2 - a_j\sigma^2 = \\
&X_{n\bullet}(X^T_{[1:n-1]}X_{[1:n-1]})^{-1}X^T_{j\bullet}\sigma^2 - X_{n\bullet}(X^T_{[1:n-1]}X_{[1:n-1]})^{-1}X^T_{j\bullet}\sigma^2 = 0.
\end{aligned}
$$

From this it follows that $Y_n - X_{n\bullet}\hat{\beta}_{(n-1)}$ and $Y_{[1:n-1]} - X_{[1:n-1]}\hat{\beta}_{(n-1)}$ are independent. And thus, w_n and SSE_{n-1} are independent. Therefore, the expression in (4.26) follows a student's t-distribution with $n-p-1$ degrees of freedom and the result from the lemma follows. □

The proof that Brown gives on the independence of the recursive residuals assumes that σ is known. We suspect that this property still holds when σ is unknown and an estimator of σ is used. However, we did not establish to figure out the exact proof.

4.1.4 Recursive coefficients

While the previous two approaches focus on the residuals of the regression model, it is also possible to monitor the coefficients β of the regression model. This method was introduced by Dufour (1982) and this section is based on his work. The idea is to re-estimate the coefficients of the regression model every time a new observation occurs, using

$$
\hat{\beta}_{(n)} = (X^T_{[1:n]\bullet}X_{[1:n]\bullet})^{-1}X^T_{[1:n]\bullet}Y_{[1:n]}. \tag{4.27}
$$

In this way a sequence of coefficient estimates $(\hat{\beta}_{(n+1)}, \hat{\beta}_{(n+2)}, \ldots)$ can be derived. We are interested in the difference between two successive coefficient estimates $\hat{\beta}_{(i)} - \hat{\beta}_{(i-1)}$. Some useful formulas to compute these differences are given in Brown et al. (1975) and they are summarized in the following lemma.

Lemma 4.1.4. *Consider the multiple linear regression model as defined in Definition 4.1.1. Then* $(X^T_{[1:n]\bullet}X_{[1:n]\bullet})^{-1}$ *can be rewritten as*

$$
(X^T_{[1:n]\bullet}X_{[1:n]\bullet})^{-1} = (X^T_{[1:n-1]\bullet}X_{[1:n-1]\bullet})^{-1} - \frac{(X^T_{[1:n-1]\bullet}X_{[1:n-1]\bullet})^{-1}X^T_{n\bullet}X_{n\bullet}(X^T_{[1:n-1]\bullet}X_{[1:n-1]\bullet})^{-1}}{1 + X_{n\bullet}(X^T_{[1:n-1]\bullet}X_{[1:n-1]\bullet})^{-1}X^T_{n\bullet}} \tag{4.28}
$$

and $\hat{\beta}_{(n)}$ *as described in Definition 4.1.3 can be re-written as*

$$
\hat{\beta}_{(n)} = \hat{\beta}_{(n-1)} + (X^T_{[1:n]\bullet}X_{[1:n]\bullet})^{-1}X^T_{n\bullet}(Y_n - X_{n\bullet}\hat{\beta}_{(n-1)}). \tag{4.29}
$$

Proof. See the proof of Lemma 2 in Brown et al. (1975). □

Let the nth recursive coefficient be defined as

$$b_n = \hat{\beta}_{(n)} - \hat{\beta}_{(n-1)} \tag{4.30}$$

The results from Lemma 4.1.4 can be used to write the recursive coefficients as

$$\begin{aligned} b_n &= \hat{\beta}_{(n)} - \hat{\beta}_{(n-1)} \\ &= (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1} X^T_{n\bullet}(Y_n - X_{n\bullet}\hat{\beta}_{(n-1)}) \\ &= d_n (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1} X^T_{n\bullet} w_n, \end{aligned} \tag{4.31}$$

where w_n is the nth recursive residual as defined in (4.23) and d_n is its denominator. Using Lemma D.0.1 it follows that

$$\mathbb{E}(b_n) = \mathbb{E}(w_n) = 0. \tag{4.32}$$

The covariance of the recursive coefficient can be written as

$$\begin{aligned} \text{Cov}(b_n) = \Sigma_b &= \mathbb{E}((\hat{\beta}_{(n)} - \hat{\beta}_{(n-1)}))(\hat{\beta}_{(n)} - \hat{\beta}_{(n-1)}))^T) \\ &= \mathbb{E}((d_n (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1} X^T_{n\bullet} w_n)(d_n (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1} X^T_{n\bullet} w_n)^T) \\ &= d_n^2 (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1} X^T_{n\bullet} X_{n\bullet} (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1} \mathbb{E}(w_n^2) \\ &= \sigma^2 d_n^2 (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1} X^T_{n\bullet} X_{n\bullet} (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1}. \end{aligned} \tag{4.33}$$

In the previous section it was already shown that the recursive residuals w_n are independent, identically $\mathcal{N}(0, \sigma^2)$ distributed. Using this property and the fact that b_n is a linear combination of univariate normally distributed random variables, it follows that the b_n's are independent, identically $\mathcal{N}_{p+1}(0, \Sigma_b)$ distributed. Note that $(X^T_{n\bullet} X_{n\bullet})$ is the product of a column vector and a row vector, which results in a $(p+1 \times p+1)$-matrix. Because of how the matrix is constructed, the rows of $(X^T_{n\bullet} X_{n\bullet})$ are linearly dependent and thus it follows that $(X^T_{n\bullet} X_{n\bullet})$ is a singular matrix. Let us define the following lemma to understand why the covariance of b_n is singular.

Lemma 4.1.5. *Let $A, B, C \in \mathbb{R}^{n\times n}$. Assume that A and C are non-singular and B singular. Then the matrix product ABC is singular.*

Proof. B is a singular matrix if and only if its determinant is equal to zero. The determinant is a multiplicative function. So, if the determinant of B is zero then the determinant of ABC is also zero. Thus, ABC is also singular. □

Thus, from Lemma 4.1.5 it follows that Σ_b is a singular matrix. There is one exception in the case that there is no intercept included in the model and the number of variables is equal to one. In this case $(X^T_{n\bullet} X_{n\bullet})$ is a scalar and the inverse can be computed if $(X^T_{n\bullet} X_{n\bullet})$ is unequal to zero. The fact that the covariance of b_n is singular, makes b_n unsuitable for multivariate statistical process control via the T^2 chart. In Section 4.4.1 it will become clear why this is the case.

Since one of the requirements for multivariate statistical process control cannot be met, let us study the components of b_n to see if these can be used for univariate statistical process control. Let $\hat{\beta}_{(n)} = (\hat{\beta}_{0,(n)}, \ldots, \hat{\beta}_{p,(n)})^T$ and define

$$A_{(n)} = (X^T_{[1:n]\bullet} X_{[1:n]\bullet})^{-1}, \tag{4.34}$$

where $A_{(n),j}$ is the jth column of $A_{(n)}$. Then the jth component of $\hat{\beta}_{(n)} - \hat{\beta}_{(n-1)}$ can be written as

$$b_{j,(n)} = \hat{\beta}_{j,(n)} - \hat{\beta}_{j,(n-1)} = d_n (X_{n\bullet} A_{(n),j})^T w_n. \tag{4.35}$$

Since we already saw that the recursive residuals w_n are i.i.d $\mathcal{N}(0, \sigma^2)$, we also have that all components of $b_{(n)}$ are independent with

$$\mathbb{E}(b_{j,(n)}) = 0 \tag{4.36}$$

and

$$\text{Var}(b_{j,(n)}) = (d_n(X_{n\bullet}A_{(n),j})^T)^2\sigma^2. \tag{4.37}$$

So, if we assume that $d_n(X_{n\bullet}A_{(n),j})^T \neq 0$ for all j, n, then we can define the scaled components of $b_{(n)}$ as

$$\Delta_{j,(n)} = \frac{b_{j,(n)}}{d_n(X_{n\bullet}A_{(n),j})^T}, \quad n = 2, 3, \ldots, \tag{4.38}$$

where $\Delta_{j,(n)}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Summary and conclusions

We discussed two types of residuals in this section. The predictive residuals are correlated and follow a normal distribution when μ and σ are assumed to be known. If μ and σ are unknown, we proved that the predictive residuals follow a student's t-distribution. Unfortunately, the predictive residuals are correlated, which makes them inappropriate to use for statistical process control. The recursive residuals are independent, identically $\mathcal{N}(0, \sigma^2)$ distributed when μ and σ are assumed to be known. We proved that they follow a student's t-distribution with $n - p - 1$ degrees of freedom, if μ and σ are unknown. While we suspect that the independence property also holds when μ and σ are unknown, a formal proof could not be given. However, we still believe that the recursive residuals are more suitable for monitoring than the predictive residuals.

We also discussed what we call the recursive coefficients. The recursive coefficients are the differences of two successive recursive coefficient estimates. The recursive coefficients follow a $\mathcal{N}_{p+1}(0, \Sigma_b)$ distribution when μ and Σ_b are assumed to be known. However, Σ_b is singular by definition. This makes the recursive coefficients unsuitable for multivariate statistical process control. However, the distributions of the components of the recursive coefficient vectors are also derived. Therefore, the individual components can be monitored using univariate statistical process control approaches.

4.2 Variable selection

In this section, the LASSO variable selection method will be discussed in more detail. Furthermore, cross validation methods will be discussed since these methods are not only used in the LASSO method but also to validate the final regression model.

4.2.1 The LASSO method

The data set we use contains approximately 60 variables. When we would like to include also the squares and second order interactions of these variables, this number soon becomes very large. This large number of variables makes it impossible to perform classical variable selection methods such as stepwise selection. The Least Absolute Shrinkage and Selection Operator (LASSO) method was introduced in Tibshirani (1996). LASSO is a regularization and variable selection method that can be used for a large number of variables. A regularization method is used to reduce the variance of the regression model when the number of variables increases to avoid overfitting the model. LASSO sets constraints to the magnitude of the regression coefficients, which leads to some of the coefficients being shrunk to zero. This is why LASSO also can be used as a variable selection method. Variables that are shrunk to zero, are not selected by the LASSO method.

In LASSO regression the following optimization problem is solved:

$$\text{minimize } ||Y - X\beta||_2^2 \quad \text{subject to } \sum_{j=0}^{p} |\beta_j| < t, \tag{4.39}$$

where $t > 0$ is an upper bound for the 1-norm of the coefficient vector. Setting an upper bound to the norm of the coefficient vector, reduces overfitting. To understand this, think of the case where the number of variables is equal to the number of observations. Then, the coefficients can be chosen such that the model fits the training

data set perfectly. Say that the 1-norm of the coefficient vector in this case is t. Then, if we add the constraint that the 1-norm of the coefficients must be smaller than t, we can no longer have a perfect fit on the training data. This probably improves the fit in the test data and thus the overfitting is reduced. When the number of observations is larger than the number of variables, a perfect fit in the training data can probably not be made, but adding a constraint to the 1-norm of the coefficient vector still reduces overfitting in the same way.

According to Hastie et al. (2016), the optimization problem in (4.39) is equivalent to

$$\min_{\beta} \left\{ ||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}, \tag{4.40}$$

where $\lambda \geq 0$. Insignificant parameters usually have little effect on the sum of squares, but they do have an effect on the 1-norm of the coefficient vector. That is why insignificant parameters are pushed to zero by the LASSO method. An example of the result obtained by performing LASSO can be found in Figure 4.1.

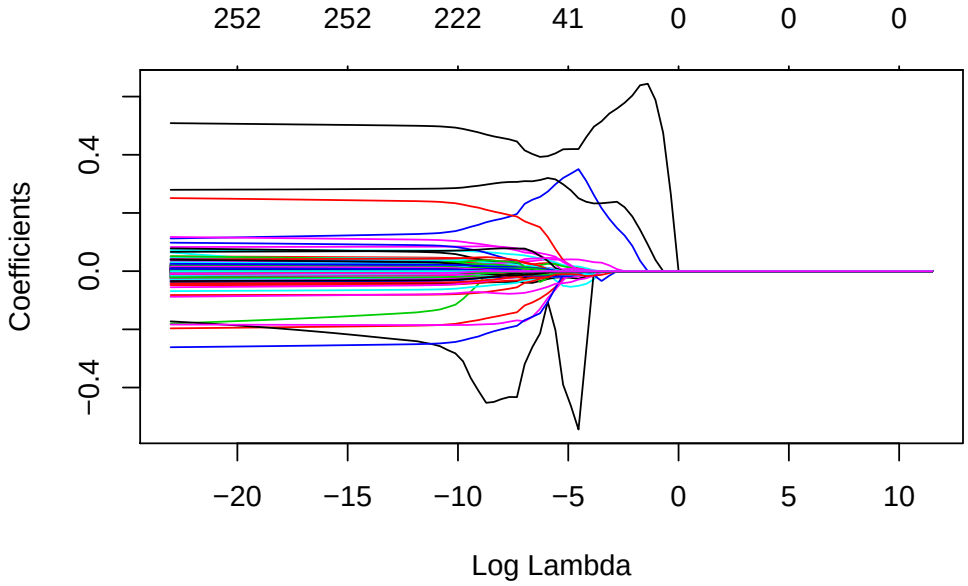


Figure 4.1: Example of LASSO variable selection. The bottom horizontal axis shows the different values of λ and the top horizontal axis shows the number of variables that are included the model for the corresponding values of λ. The coloured lines represent the coefficient estimates. It can be seen that eventually all estimates are shrunk to zero.

The value of λ influences the amount of shrinkage that takes place. In the extreme case where $\lambda = 0$, the optimization problem in (4.40) is equivalent to the ordinary least squares problem. A large value of λ will lead to a large amount of shrinkage, which reduces the number of variables that is included in the model. The "best" choice for λ can be determined via cross validation, which is described in more detail in the next section. Here, the "best" value of λ is the value of λ that leads to a model which can predict the observations of independent test data the most accurately. In this case, we use 10-fold cross validation. An example of the result obtained via this 10-fold cross validation can be found in Figure 4.2. After cross validation is performed for each value of λ, we choose the value of λ for which the error is one standard error away from the minimum Mean Squared Error (MSE). As can be seen in Figure 4.2, the minimal MSE plus one standard deviation is still very small, but the number of parameters is reduced from 188 to 61, which is a lot. This is why we do not take the value of λ that corresponds to the minimal MSE.

4.2.2 Cross validation

Cross validation is a method that can be used to validate a model. The idea of cross validation is to fit the model on a training data set and then validate the model on an independent test data set. According to James et al. (2014) using just one training and test set has two major drawbacks. Firstly, the quality of the predictions

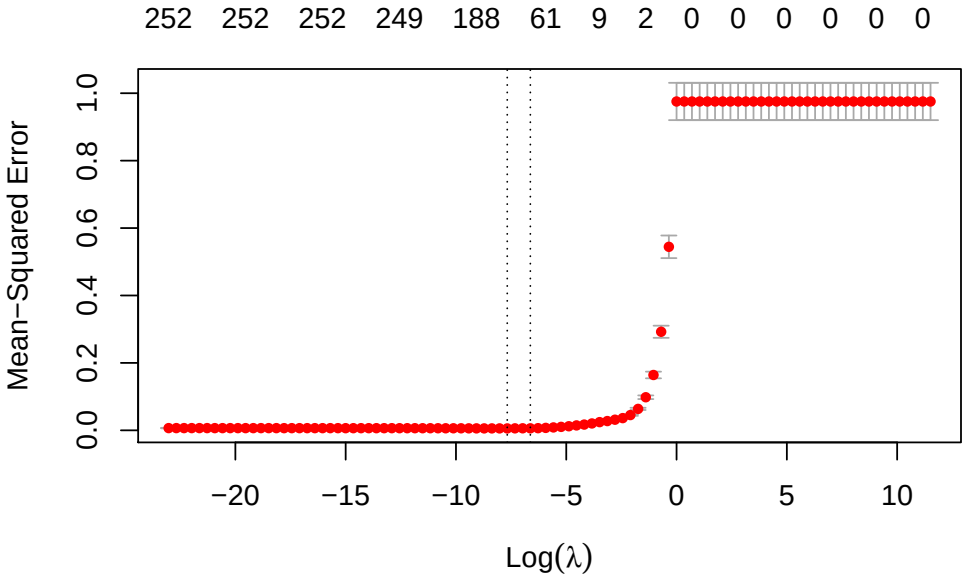


Figure 4.2: The optimal value of λ in LASSO variable selection. The first vertical dotted line is where the mean squared error is minimal. The second vertical dotted line is where the mean squared error is one standard error away from the minimum. We choose the value of λ that corresponds to the second dotted line.

in the test set is highly variable for different test sets. Secondly, the training set only uses a subset of the observations to fit the model. Methods usually perform worse when trained on less data, so this will lead to overestimating the prediction errors. James et al. (2014) suggests using cross validation to overcome these issues. In cross validation, the data is split into equal parts. Each part will be included in the test data set and the remaining observations will be used to train the data. In Leave One Out Cross Validation (LOOVC) only one observation will be used as test data. The advantage of this method is that the training data almost uses all available data, so the effect of overestimating the prediction errors will no longer be present. However, when the data set grows, the computational complexity of this method also grows. Therefore, we will use k-fold cross validation. In k-fold cross validation the data is split into k equal parts. This makes the method more computationally efficient. According to James et al. (2014) one typically uses 5-fold or 10-fold cross validation.

Summary and conclusions

In this section, two statistical methods that will be used in the variable selection process are described. LASSO is a regularization and variable selection method, that is especially useful when the number of variables is large. The model that is obtained via LASSO variable selection must be validated properly. To this end k-fold cross validation will be used. While LOOCV is the "best" method to validate the model, it is computationally infeasible because of the large data sets. Therefore, k-fold cross validation is an appropriate alternative.

4.3 Univariate Statistical Process Control

Statistical Process Control (SPC) is a collection of methods that can be used to monitor the quality of a process. It is important to note that SPC is about monitoring the quality and not about controlling the process as the name might suggest. In this section we start with describing some important concepts in the area of statistical process control. The book of Qiu (2013) is used as our main reference.

An important part of statistical process control is to determine when a stable (In-Control (IC)) process becomes unstable (Out-of-Control (OC)). A process will always have some variability or randomness, which is

called the common cause variation. But, when the system becomes unstable this is often caused by special cause variation. So, it is important to distinguish between those two types of variation. An advantage of using SPC methods is that deviations in the production process can be detected early, which usually prevents larger problems. This is perfectly in line with the goal of the 'Data voor Onderhoud' department at the Royal Netherlands Navy.

SPC is often divided into two phases, phase I and phase II. In phase I the goal is to obtain a process that is in-control. At the beginning of this phase, this might not be the case. Some basic statistical methods can be used to adjust the process until the process runs stably. When the process is stable, an estimation of the involved parameters can be made. In phase II it is assumed that the process runs stably from the beginning. The goal for this phase is to monitor the process and give a signal if the process is no longer in-control. The parameter estimates from the phase I analysis are used to this end. An important difference between the two phases is that the distribution of the quality characteristics is usually unknown in phase I, but in phase II it is assumed to be known (note that the distribution is estimated based on phase I in practice). In this report, the focus will lie on phase II SPC, since we have no influence on the process itself.

Important tools to monitor the process during phase I and phase II are control charts. In a control chart, the charting statistic, an Upper Control Limit (UCL) and a Lower Control Limit (LCL) are plotted over time. Whenever the charting statistic exceeds the UCL or LCL, the chart signals. This signal indicates that the process shows out-of-control behaviour. Before we can discuss these charts in more detail, an important concept in SPC should be discussed: the run length. First, let us consider the situation in which the process is in-control. In this case the run length is the number of observations from the initial time point to the first *false* alarm. Since the observations are all random, the run length is also a random variable. The Average Run Length (ARL) of an in-control process is referred to as the ARL_0. Preferably, we want the ARL_0 to be large, which means that we have few false alarms. The average out-of-control run length is referred to as the ARL_1. This is the number of observations from the time point of a shift in the process to the first signal. Therefore, we want the ARL_1 to be small, so that we can detect shifts in the process fast. Together, the ARL_0 and ARL_1 are measures of the quality of a control chart. Unfortunately, a large ARL_0 typically leads to a large ARL_1 and vice-versa. Usually, the ARL_0 value of a chart is set to a pre-specified value and then the parameters of the chart are chosen such that the ARL_1 is as small as possible.

In this section we will introduce several univariate control charts. We continue to use the book of Qiu (2013) as the main reference.

4.3.1 Shewhart chart

Originally, Shewhart control charts were used to monitor batch data. At each point in time, a sample of quality characteristics is collected and monitored. However, the data from the Royal Netherlands Navy vessels consists of a single observation at each time point. That is why we will discuss the Shewhart individuals chart.

Let $\{X_1, X_2, \ldots\}$ be consecutive observations of the quality characteristic or any other quantity that is being monitored. It is assumed that the observations X_i are independent identically distributed from a normal distribution. In the Shewhart individuals chart, as a charting statistic, the individual observations X_i are used. Furthermore, a Centre Line (CL) and the control limits (LCL, UCL) are plotted in the chart. Let μ be the mean of the quality characteristic and σ^2 the variance of the quality characteristic. Then, the control limits are $\mu \pm 3\sigma$. When an observation exceeds the control limits, the chart will give a signal. Often μ and σ are unknown, in this case, they need to be estimated from the phase I data. As an estimate for μ, $\hat{\mu} = \overline{X}$ is used. Historically, σ is estimated using the sample ranges. In the case of individual observations, sample ranges are unavailable and therefore the moving range $\text{MR}_i = |X_i - X_{i-1}|$ is used. Using these moving ranges an estimate for σ is $\hat{\sigma} = \overline{\text{MR}}/1.128$. Dividing by the constant 1.128 makes $\hat{\sigma}$ and unbiased estimator. This is explained in more detail in Qiu (2013). As an example, a visualization of a Shewhart individuals chart can be found in Figure 4.3.

When the process is in-control (IC), it can still happen that an observation exceeds the control limits, we then speak of a false alarm. When the process is IC, the run length follows a geometric distribution since the number of failures before a success (in our case a false alarm) are counted. Since the quality characteristics are assumed to be independent and identically normal distributed, the probability of a false alarm is equal to the probability that a standard normal random variable takes a value outside the interval $(-3, 3)$. This probability

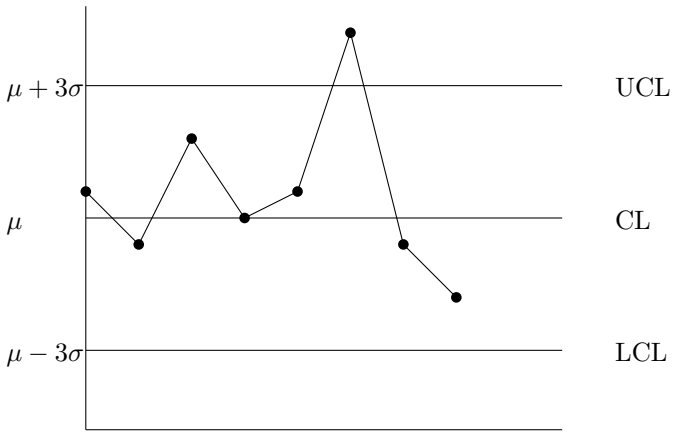


Figure 4.3: Shewhart individuals chart with control limits.

is equal to 0.0027. So, the false alarm rate of the Shewhart chat is 0.0027, which leads to an ARL_0 of $1/0.0027$ which is approximately equal to 370.

When the process is out-of-control (OC), the run length still follows a geometric distribution. Say that at time τ the process mean shifts from μ_0 to $\mu_1 = \mu_0 + k\sigma$. Then the probability that an observation X_i with $i \geq \tau$ exceeds the control limits is

$$\gamma = 1 - \mathbb{P}(\mu_0 - 3\sigma \leq X_i \leq \mu_0 + 3\sigma) = 1 - \mathbb{P}(-k - 3 \leq \psi \leq -k + 3), \tag{4.41}$$

where ψ is a standard normal random variable. This leads to an ARL_1 of $1/\gamma$. From this it can be concluded that the ARL_1 of a Shewhart chart depends on the size of the shift. If the shift is large, the ARL_1 will be small and if the shift is small, the ARL_1 will be large. For example, if $k = 1$ (small shift) this leads to an ARL_1 of approximately 44 and if $k = 3$ (large shift) then this leads to an ARL_1 of 2.

Implementation

In the past, it was common practice to use moving ranges as an estimation for the standard deviation, since they were easy to compute. Nowadays, everyone has access to computers and the sample standard deviation can be estimated via the well-known formula

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}. \tag{4.42}$$

The estimator $\hat{\sigma}$ as defined in (4.42) is used in the rest of this report.

When the predictive residuals are used as input for the Shewhart chart the methods that are described in the paragraphs above can be applied directly. A stable period needs to be selected in which the regression model is fitted. Next, the mean and standard deviation of the residuals in the stable period need to be determined. Using the estimated mean and standard deviations, the control limits can be computed. These limits are used in the periods where the regression model is used to monitor the process.

For the recursive residuals a stable period is not available. This means that the estimations of the mean and standard deviation need to be updated in every iteration. In each iteration, an observation is added to the sample. In iteration $i+1$ the observations $X_1, \ldots, X_i$ are used to fit a regression model. These observations are also used to compute the mean and standard deviation as described above. When a new observation is added to the sample, the mean and standard deviation are re-estimated. As a consequence, the control limits of charts based on recursive residuals will be constantly changing.

In the Shewhart control chart, the 3σ limits are chosen because the $(1 - 0.0027/2)$th quantile of the normal distribution is approximately 3. This leads to an ARL_0 of 370. However, in Sections 4.1.2 and 4.1.3 it was shown that the predictive and recursive residuals follow a student's t-distribution if μ and σ are unknown. In our case μ and σ are indeed unknown and they need to be estimated from the data. Therefore, we will not use the $\mu \pm 3\sigma$ control limits. Instead we use $\hat{\mu} \pm q_t\hat{\sigma}$, where $\hat{\mu}$ and $\hat{\sigma}$ are the estimates for μ and σ and where q_t is the

$(1 - 0.0027/2)$th quantile of the corresponding student's t-distribution. Note that the student's t-distribution approaches a normal distribution as the degrees of freedom increase. For the predictive residuals, the degrees of freedom of the student's t-distribution is approximately as large as the number of observations that is included in the stable period. This is typically a large number of observations and thus changing the control limits does not really make a difference in this case. For the recursive residuals changing the limits could make a difference in the beginning of the monitoring process. While monitoring the first few observations, the number of degrees of freedom of the corresponding student's t-distribution is not very large yet. However, from 30 observations onward the difference between the student's t-distribution and the normal distribution is already very small.

4.3.2 CUSUM chart

It appears that Shewhart charts are not good at detecting small and persisting shifts of the process. We have seen that the ARL_1 gets large in case of a small mean shift in the examples of (4.41). Therefore, they are very useful for phase I SPC, but not so much for phase II SPC. This is because Shewhart charts do not take the history of a process into account. The charting statistic of the Shewhart chart is the current observation and all previous observations are ignored. Cumulative Sum (CUSUM) charts do not only use the current observation, but also all previous observations to compute the charting statistic. Assume that $\{X_1, X_2, \ldots\}$ are independent observations from a univariate process at consecutive time points. Suppose that the in-control distribution is $\mathcal{N}(\mu_0, \sigma^2)$ and that the distribution becomes $\mathcal{N}(\mu_1, \sigma^2)$ when the process is out-of-control, where $\mu_0 \neq \mu_1$. The charting statistic for detecting an upward mean shift for the CUSUM chart is defined as

$$C_n^+ = \max\left\{0, C_{n-1}^+ + (X_n - \mu_0) - k\right\}, \tag{4.43}$$

where $C_0^+ = 0$. The chart signals when $C_n^+ > h$. Formula (4.43) shows the restarting principle of the CUSUM chart. When $C_{n-1}^+ + (X_n - \mu_0) < k$ there is little evidence for an upward mean shift, so the statistic will be reset to zero. For detecting a downward mean shift the charting statistic is defined as

$$C_n^- = \min\left\{0, C_{n-1}^- + (X_n - \mu_0) + k\right\}, \tag{4.44}$$

where $C_0^- = 0$. The chart signals when $C_n^- < -h$.

Before the CUSUM chart can be implemented and used, the selection of the parameters h and k should be discussed. In this discussion it is assumed that the observations $\{X_1, X_2, \ldots\}$ are standardized, so $\mu_0 = 0$ and $\sigma^2 = 1$. In SPC, usually, the ARL_0 is chosen beforehand. Then, when detecting a mean shift of size δ, the chart performs better if it has a smaller ARL_1. For given values of k and h, the ARL_0 can be determined using simulations as described in Qiu (2013). Qiu also writes that when one would like to detect a mean shift of size δ, the optimal value of k is $\delta/2$. It can be hard to determine the size of δ on beforehand. If this is the case, the size of the shift one desires to detect can be used. When the values of k and the ARL_0 are fixed, the corresponding value of h can be determined. This is commonly done using a bisection-like method, where in each iteration the ARL_0 is computed for the fixed k and some value of h. Then the search interval for h is adjusted based on the resulting ARL_0. Some functions to perform these simulations are available in the R-package `spc`.

When k and h are chosen to achieve a specific ARL_0, the ARL_1 can be computed. First, consider the situation where mean shift takes place at the initial time point. Suppose that at the initial time point the process shifts from a $\mathcal{N}(\mu_0, \sigma^2)$ to a $\mathcal{N}(\mu_0 + \delta, \lambda^2\sigma^2)$ distribution. Then, Formula 4.43 can be rewritten as

$$\frac{C_n^+}{\lambda\sigma} = \max\left\{0, \frac{C_{n-1}^+}{\lambda\sigma} + \left(\frac{X_n}{\lambda\sigma} - \frac{\mu_0 + \delta}{\lambda\sigma}\right) - \frac{k - \delta}{\lambda\sigma}\right\}. \tag{4.45}$$

When we define $C_n^* = C_n^+/(\lambda\sigma)$ Formula (4.45) shows a CUSUM charting statistic with $k^* = (k - \delta)/(\lambda\sigma)$ and $h^* = h/(\lambda\sigma)$. So, if the process shifts from a $\mathcal{N}(\mu_0, \sigma^2)$ to a $\mathcal{N}(\mu_0 + \delta, \lambda^2\sigma^2)$ distribution at the initial time point, the ARL_1 of the process is the same as the ARL_0 of a process with $k = k^*$ and $h = h^*$. Since the distribution of C_n^+ depends on n, the true ARL_0 and ARL_1 also depend on n. (Qiu, 2013, Chapter 4) discusses this issue. It can be concluded that the true ARL_0 and ARL_1, might be a little different when the shifts do not occur at the initial time points. But these differences are relatively small. Therefore, we will use the zero-state average run lengths.

So far, we only discussed one-sided CUSUM charts for detecting an upward or a downward mean shift. However, these two charts can be combined to a two-sided CUSUM chart for detecting an arbitrary mean shift. In the two-sided chart, the statistics C_n^+ and C_n^- are both plotted over time. The chart signals if $C_n^+ > h$ or $C_n^- < -h$ or both. Say the ARLs of the one-sided charts are ARL^+ and ARL^-. Then, the ARL of the two-sided chart can be computed as

$$\frac{1}{\text{ARL}} = \frac{1}{\text{ARL}^+} + \frac{1}{\text{ARL}^-}. \tag{4.46}$$

This holds for both the in-control and out-of-control average run lengths. From this formula it also follows that the ARL_1 of a two-sided chart is always smaller than the ARL_1 of the corresponding one-sided charts. This is why it is recommended to use two-sided CUSUM charts. Qiu (2013) computed the corresponding value of h for several commonly used values of k and ARL_0 for the one-sided CUSUM charts. For completeness, we did the same for the two-sided version. The results can be found in Table B.1 in the appendix.

4.3.3 EWMA chart

Just like the CUSUM chart, the Exponentially Weighted Moving Average (EWMA) charts also use all observations including the current observation to compute the charting statistic. However, some optimality properties are proven for the CUSUM chart that are not proven for the EWMA chart. Nonetheless, from many discussions in literature it follows that the performance of EWMA charts is comparable to the performance of CUSUM charts. EWMA charts have the advantage that they are easier to implement and they are more robust against normality. This is why we will also discuss the EWMA charts in this section.

Assume that $\{X_1, X_2, \ldots\}$ are independent observations from a univariate process at consecutive time points. Suppose that the in-control distribution is $\mathcal{N}(\mu_0, \sigma^2)$ and that the distribution becomes $\mathcal{N}(\mu_1, \sigma^2)$ when the process is out-of-control, where $\mu_0 \neq \mu_1$. The charting statistic of the EWMA chart is defined as

$$E_n = \lambda X_n + (1-\lambda)E_{n-1}, \tag{4.47}$$

for $n \geq 1$. Here $\lambda \in (0, 1]$ is called the weighting parameter and $E_0 = \mu_0$. The charting statistic E_n can be rewritten as

$$\begin{aligned} E_n &= \lambda X_n + \lambda(1-\lambda)X_{n-1} + \ldots + \lambda(1-\lambda)^{n-1}X_1 + (1-\lambda)^n\mu_0 \\ &= (1-\lambda)^n\mu_0 + \lambda\sum_{i=1}^{n}(1-\lambda)^{n-i}X_i. \end{aligned} \tag{4.48}$$

It follows that

$$(1-\lambda)^n + \lambda\sum_{i=1}^{n}(1-\lambda)^{n-i} = (1-\lambda)^n + \lambda\sum_{i=0}^{n-1}(1-\lambda)^i = (1-\lambda)^n + \lambda\left(\frac{1-(1-\lambda)^n}{\lambda}\right) = 1. \tag{4.49}$$

So, E_n is a weighted average of the observations $\{X_1, \ldots, X_n\}$ and the in-control mean μ_0. When λ is close to 1 more weight is assigned to the current observation, the present. When λ is close to 0, more weight is assigned to the previous observations, the history. In the special case where $\lambda = 1$, the EWMA chart is equivalent to the Shewhart chart. In order to derive control limits for this charting statistic, the distribution of the charting statistic must be determined. It follows that

$$\begin{aligned} \mathbb{E}(E_n) &= \mathbb{E}\left((1-\lambda)^n\mu_0 + \lambda\sum_{i=1}^{n}(1-\lambda)^{n-i}X_i\right) \\ &= (1-\lambda)^n\mu_0 + \lambda\sum_{i=1}^{n}(1-\lambda)^{n-i}\mu_0 = \mu_0, \end{aligned} \tag{4.50}$$

where the last equality follows from (4.49). Furthermore, it holds that

$$\begin{aligned}\text{Var}(E_n) &= \text{Var}\left((1-\lambda)^n\mu_0 + \lambda\sum_{i=1}^{n}(1-\lambda)^{n-i}X_i\right) = \text{Var}\left(\lambda\sum_{i=1}^{n}(1-\lambda)^{n-i}X_i\right)\\ &= \sum_{i=1}^{n}\text{Var}(\lambda(1-\lambda)^{n-i}X_i) = \sigma^2\lambda^2\sum_{i=1}^{n}((1-\lambda)^{n-i})^2\\ &= \sigma^2\lambda^2\sum_{i=0}^{n-1}((1-\lambda)^2)^i = \sigma^2\lambda^2\left(\frac{1-((1-\lambda)^2)^n}{1-(1-\lambda)^2}\right)\\ &= \frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2n})\sigma^2. \end{aligned} \tag{4.51}$$

So, the in-control distribution of the EWMA statistic is

$$E_n \sim \mathcal{N}\left(\mu_0, \frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2n})\sigma^2\right). \tag{4.52}$$

Now, assume that the process mean shits from μ_0 to μ_1 at time $1 \leq \tau \leq n$. The EWMA charting statistic is denoted as $E_{n,(OC)}$ in this case. The computations that lead to (4.51) are still valid, so the variance of $E_{n,(OC)}$ is the same as the variance of E_n. However, the mean of $E_{n,(OC)}$ becomes

$$\begin{aligned}\mathbb{E}(E_{n,(OC)}) &= (1-\lambda)^n\mu_0 + \lambda\mu_0\sum_{i=1}^{\tau-1}(1-\lambda)^{n-i} + \lambda\mu_1\sum_{i=\tau}^{n}(1-\lambda)^{n-i}\\ &= (1-\lambda)^n\mu_0 + \lambda\mu_0\left(\frac{1-(1-\lambda)^n}{\lambda} - \frac{1-(1-\lambda)^{n-\tau+1}}{\lambda}\right) + \lambda\mu_1\frac{1-(1-\lambda)^{n-\tau+1}}{\lambda}\\ &= (1-\lambda)^{n-\tau+1}\mu_0 + (1-(1-\lambda)^{n-\tau+1})\mu_1. \end{aligned} \tag{4.53}$$

Hence, the expected value of $E_{n,(OC)}$ is a weighted average of the in-control mean μ_0 and the out-of-control mean μ_1. This makes that the charting statistic E_n, indeed caries useful information about the mean shift in the process. Using the in-control distribution of E_n, the in-control region for the EWMA statistic can be defined as

$$\mu_0 \pm \rho\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2n})}\sigma, \tag{4.54}$$

with $\rho > 0$. So, when an EWMA chart is used to monitor a process, E_n should be computed each time a new observation X_n is registered. Next, it should be checked if E_n is within the in-control region as defined in (4.54). If this is the case, it can be concluded that the process is in-control up to time point n. When E_n exceeds one of the control limits this is a signal that the process shows out-of-control behaviour.

Before the chart can be used in practice, the values for λ and ρ should be determined. Usually, the value of λ is specified beforehand. A small value of λ should be chosen when the target mean shift is small and a large λ should be chosen when the target mean shift is relatively large. Then, the value of ρ if computed to obtain the pre-specified ARL_0. The computation of the ARL_0 and corresponding value of ρ is similar to the computations of the ARL_0 and the parameter h of the CUSUM chart. Functions that perform these simulations are available in the R-package `spc`. In Chapter 5 of Qiu (2013), the corresponding values of ρ for several values of λ and ARL_0 are computed. For completeness, these results are also displayed in Table B.2 in the appendix.

Implementation

Since $\lambda \in (0, 1]$, the variance of E_n approaches

$$\frac{\lambda}{2-\lambda}\sigma^2, \tag{4.55}$$

for large n (for $\lambda = 0.1$ this is already true for $n > 15$). Therefore, in practice the control limits are often set to

$$\mu_0 \pm \rho\sqrt{\frac{\lambda}{2-\lambda}}\sigma. \tag{4.56}$$

These control limits will also be used in the continuation of this report.

When the predictive residuals are used as input for the EWMA charts, the procedures as described above can applied directly. A stable period is used to fit the regression model. The residuals from this regression model are used to compute the EWMA statistic. The stable period is also used to compute estimations of the in-control mean and standard deviation. These are used to determine the control limits. If the recursive residuals are used as input for the EWMA chart, the estimators for the mean and standard deviation need to be updated in every iteration, just was described in Section 4.3.1 for the Shewhart charts. Every time an observation is added to the sample, the sample mean and standard deviation are re-estimated. This results in control limits that are constantly updated. Therefore, the EWMA control limits will not be constant.

Summary and conclusions

In this section, various charts for univariate statistical process control are described. While the Shewhart chart is easy to interpret, it does not take the history of a process into account. As a consequence, small but persistent mean shifts are hard to detect using a Shewhart chart. Two alternative charts are the CUSUM and EWMA charts. These two charts do take the history of the process into consideration, which makes them better at detecting small and persistent shifts in mean. A disadvantage of the CUSUM and EWMA charts is that parameter values need to be chosen. The optimal values for these parameters depend on the size of the mean shift, which is often unknown.

4.4 Multivariate Statistical Process Control

Until now, we discussed univariate statistical process control. In univariate SPC one quality characteristic is monitored at a time. But what if multiple quality characteristics together describe the quality of a process best? One option is to monitor each of the quality characteristics individually with a univariate monitoring approach, but this has some drawbacks. Firstly, when the number of variables increases, so does the number of parameters that needs to be determined. For example, if the variables are monitored using CUSUM charts, the value of k needs to be determined for all the variables and the corresponding h values have to be computed. It is often hard to determine the "correct" value of k, so this will be a time-consuming task. Secondly, when p quality characteristics are monitored at the same time, the true false alarm rate is higher than the false alarm rate of the individual charts. Assume that the p quality characteristics are independent. Then the combined false alarm rate α_c is $1-(1-\alpha)^p$, where α is the false alarm rate of the individual charts. For $p = 3$ and $\alpha = 0.0027$ this leads to $\alpha_c = 0.0081$ and for $p = 7$ and $\alpha = 0.0027$ this leads to $\alpha_c = 0.019$. When the p quality characteristics are positively correlated, which is the case in this research, the combined false alarm rate even exceeds the numbers from the examples. Another major drawback of univariate monitoring of multiple quality characteristics, is that the correlational structure of the characteristics is ignored. It can happen that all variables individually seem to be in-control, but correlational structure is different than one might expect. This is also a sign that something in the process has changed and that the process showing out-of-control behaviour. Multivariate SPC does take the correlational structure of the variables into account.

In this section we will discuss several methods for multivariate SPC. The main references for this section are the book of Mason and Young (2002) and the article of Hawkins (1991).

4.4.1 T^2 chart

The T^2 chart is probably the most well-known multivariate control chart. When monitoring a multivariate quality characteristic, there are some important aspects one should keep in mind. First of all, the various components of the multivariate quality characteristic could all contain different amounts of variation, so they should be standardized. Furthermore, the multiple components are often correlated. The correlational structure between the different components cannot be ignored. A statistic that captures both of these aspects is the T^2 statistic. In the T^2 chart, this statistic is plotted over time.

Let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$ be multivariate independent and identically distributed observations from an $\mathcal{N}_p(\mu, \Sigma)$ distribution. Here $\mathcal{N}_p(\mu, \Sigma)$ denotes p-variate normal distribution with mean μ and covariance Σ. Then the T^2

statistic at time point i can be computed as

$$T_i^2 = (\mathbf{X}_i - \mu)^T \Sigma^{-1} (\mathbf{X}_i - \mu). \tag{4.57}$$

From this formula it follows that the T^2 statistic is non-negative. Therefore, only an upper control limit is needed to monitor this statistic. To determine the upper control limit, we need to know the distribution of the T^2 statistic. The distribution of the T^2 statistic depends on whether the parameters μ and Σ are known. In the case that μ and Σ are unknown they can be estimated by

$$\hat{\mu} = \overline{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i, \quad \hat{\Sigma} = S = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})^T S^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}). \tag{4.58}$$

When these estimators are used to compute the T^2 statistic, the T^2 statistic is defined as

$$T_i^2 = (\mathbf{X}_i - \overline{\mathbf{X}})^T S^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}). \tag{4.59}$$

The distribution of the T^2 statistic is given for three different scenarios.

1. The parameters μ and Σ are known. In this case it holds that

$$T_i^2 = (\mathbf{X}_i - \mu)^T \Sigma^{-1} (\mathbf{X}_i - \mu) \sim \chi^2_{(p)}, \tag{4.60}$$

   where $\chi^2_{(p)}$ is a chi-squared distribution with p degrees of freedom. For the proof we refer to Hotelling (1947).

2. The parameters μ and Σ are unknown and their estimates $\overline{\mathbf{X}}$ and S are determined independent of the current observation $\mathbf{X}_i$. In this case we have

$$T_i^2 = (\mathbf{X}_i - \overline{\mathbf{X}})^T S^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}) \sim \left[\frac{p(n+1)(n-1)}{n(n-p)} \right] F_{(p,n-p)}, \tag{4.61}$$

   where $F_{(p,n-p)}$ is an F distribution with p and $n-p$ degrees of freedom. For the proof we refer to Tracy et al. (1992).

3. The parameters μ and Σ are unknown and the current observation $\mathbf{X}_i$ is used (among other observations) to determine $\overline{\mathbf{X}}$ and S. In this case we have

$$T_i^2 = (\mathbf{X}_i - \overline{\mathbf{X}})^T S^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}) \sim \left[\frac{(n-1)^2}{n} \right] B_{(p/2,(n-p-1)/2)}, \tag{4.62}$$

   where $B_{(p/2,(n-p-1)/2)}$ is a beta distribution with $p/2$ and $(n-p-1)/2$ degrees of freedom. For the proof we refer to Tracy et al. (1992).

The upper control limit of the T^2 chart is obtained by taking the $(1-\alpha)$th quantile of the correct distribution. Here, α corresponds to the false alarm rate, which is usually chosen to be 0.0027 because this leads to an ARL_0 of 370.

4.4.2 MEWMA chart

The Multivariate Exponentially Weighted Moving Average (MEWMA) chart is introduced in Lowry et al. (1992) as a multivariate extension of the EWMA chart that is described in Section 4.3.3. Again, let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$ be multivariate identically distributed observations from a $\mathcal{N}_p(\mu_0, \Sigma)$ distribution, where μ_0 and Σ are assumed to be known. Then, a natural generalization of the EWMA chart is

$$\mathbf{E}_n = \Lambda(\mathbf{X}_n - \mu_0) + (I - \Lambda)\mathbf{E}_{n-1}, \tag{4.63}$$

where I is the $(p+1) \times (p+1)$-identity matrix, $\mathbf{E}_0 = 0$, $\Lambda = diag(\lambda_1, \ldots, \lambda_{p+1})$ and $\lambda_j \in (0, 1]$ the weighting parameter for the jth component of $\mathbf{X}$. When there is no specific reason to choose different values of λ for the

different components of $\mathbf{X}$, one should choose $\lambda_1 = \ldots = \lambda_{p+1} = \lambda$. In this case, the MEWMA statistic is computed as

$$\mathbf{E}_n = \lambda(\mathbf{X}_n - \mu_0) + (1-\lambda)\mathbf{E}_{n-1}. \tag{4.64}$$

The MEWMA chart signals when

$$V_n^2 = \mathbf{E}_n^T \Sigma_{\mathbf{E}_n}^{-1} \mathbf{E}_n > h, \tag{4.65}$$

where $\Sigma_{\mathbf{E}_n}$ is the covariance matrix of $\mathbf{E}_n$ and $h > 0$ is the control limit chosen to achieve some pre-defined ARL_0. When the expression from (4.64) is used, it follows that

$$\begin{aligned}
\mathbf{E}_n &= \lambda(\mathbf{X}_n - \mu_0) + (1-\lambda)\mathbf{E}_{n-1} \\
&= \lambda(\mathbf{X}_n - \mu_0) + \lambda(1-\lambda)(\mathbf{X}_{n-1} - \mu_0) + \ldots + \lambda(1-\lambda)^{n-1}(\mathbf{X}_1 - \mu_0) + (1-\lambda)^n \mathbf{E}_0 \\
&= \lambda \textstyle\sum_{i=1}^{n} (1-\lambda)^{n-i}(\mathbf{X}_i - \mu_0).
\end{aligned} \tag{4.66}$$

If the process is in-control until observation n, it is easy to see that $\mathbb{E}(\mathbf{E}_n) = 0$. The covariance of $\mathbf{E}_n$ can be derived as

$$\begin{aligned}
\text{Cov}(\mathbf{E}_n) = \Sigma_{\mathbf{E}_n} &= \text{Cov}(\lambda \sum_{i=1}^{n} (1-\lambda)^{n-i}(\mathbf{X}_i - \mu_0)) = \sum_{i=1}^{n} \text{Cov}(\lambda(1-\lambda)^{n-i}(\mathbf{X}_i - \mu_0)) \\
&= \sum_{i=1}^{n} \lambda^2 (1-\lambda)^{2(n-i)} \text{Cov}(\mathbf{X}_i - \mu_0) = \sum_{i=1}^{n} \lambda^2 (1-\lambda)^{2(n-i)} \Sigma \\
&= \lambda^2 \sum_{i=0}^{n-1} ((1-\lambda)^2)^i \Sigma = \lambda^2 \left(\frac{1-(1-\lambda)^{2n}}{1-(1-\lambda)^2} \right) \Sigma \\
&= \left(\frac{\lambda}{2-\lambda} \right) (1-(1-\lambda)^{2n}) \Sigma.
\end{aligned} \tag{4.67}$$

Thus, the in-control distribution of $\mathbf{E}_n$ is

$$\mathbf{E}_n \sim \mathcal{N}_p(0, \Sigma_{\mathbf{E}_n}). \tag{4.68}$$

This makes the charting statistic of the MEWMA chart, V_n^2, a normal quadratic form. Therefore, $V_n^2 \sim \chi^2_{(p+1)}$ just as the T^2 statistic follows a chi-squared distribution. Note that just as in the univariate case, the covariance of $\mathbf{E}_n$ will approach

$$\tilde{\Sigma}_{E_n} = \frac{\lambda}{2-\lambda} \Sigma \tag{4.69}$$

for n large enough.

Now, assume that the process mean shifts from μ_0 to μ_1 at time τ, with $1 \leq \tau \leq n$. In this case, the MEWMA charting statistic is denoted as $\mathbf{E}_{n,(OC)}$. The derivation of the covariance of the MEWMA statistic as described in (4.67) is still valid. However, the expectation of the out-of-control MEWMA statistic will become

$$\begin{aligned}
\mathbb{E}(\mathbf{E}_{i,(OC)}) &= \mathbb{E}(\lambda(\mathbf{X}_n - \mu_0) + \lambda(1-\lambda)(\mathbf{X}_{n-1} - \mu_0) + \ldots + \lambda(1-\lambda)^{n-\tau}(\mathbf{X}_\tau - \mu_0) + (1-\lambda)^{n-\tau-1}\mathbf{E}_{\tau-1}) \\
&= \mathbb{E}(\lambda \textstyle\sum_{i=\tau}^{n} (1-\lambda)^{n-i}(\mathbf{X}_i - \mu_0) + \lambda(1-\lambda)^{n-\tau-1} \sum_{i=1}^{\tau-1} (1-\lambda)^{n-i}(X_i - \mu_0)) \\
&= \lambda \textstyle\sum_{i=\tau}^{n} (1-\lambda)^{n-i} \mathbb{E}(X_i - \mu_0) + \lambda(1-\lambda)^{n-\tau-1} \sum_{i=1}^{\tau-1} (1-\lambda)^{n-i} \mathbb{E}(X_i - \mu_0) \\
&= \lambda \textstyle\sum_{i=\tau}^{n} (1-\lambda)^{n-i}(\mu_1 - \mu_0).
\end{aligned} \tag{4.70}$$

So, the distribution of $\mathbf{E}_{n,(OC)}$ is

$$\mathbf{E}_{n,(OC)} \sim \mathcal{N}_p \left(\lambda \sum_{i=\tau}^{n} (1-\lambda)^{n-i}(\mu_1 - \mu_0), \Sigma_{\mathbf{E}_n} \right). \tag{4.71}$$

Like we saw in Section 4.3.3 for the univariate EWMA statistic, the expectation of $\mathbf{E}_{n,(OC)}$ is a weighted average of the in-control and out-of-control means. This makes that the MEWMA statistic $\mathbf{E}_n$ contains useful information on the mean shift in the process.

We now have seen that the IC distribution of the charting statistic V_n^2 is a chi square distribution. However, we cannot simply take the $(1-\alpha)$th quantile of the IC distribution to be the upper control limit, since the

V_n^2 are correlated over time. Monte Carlo simulations can be used to derive the upper control limit, just as in the univariate case. Note that in the univariate case tables can be provided if the variable of interest is standardized. In the multivariate case the correlational structure between the quality characteristics varies in every case study. Therefore, tables cannot be provided and the upper control limit must be determined via simulation every time the MEWMA chart is applied to a specific situation. This makes the MEWMA chart less appealing compared to the T^2 chart. The pseudo code that can be used to determine the in-control average run length and upper control limit for the MEWMA chart can be found in the boxes below.

Pseudo-code to determine the ARL_0 for the MEWMA chart

Let Σ be the covariance matrix that describes the dependencies between the variables of interest. Let M be the number of iterations and let $\text{RL}_{\max}$ be the maximum allowed run length in each iteration.

Step 1 Compute the covariance of the MEWMA statistic using Formula (4.69).
Step 2 Generate the MEWMA statistics by simulating $\text{RL}_{\max}$ observations from a multivariate normal distribution with zero mean and covariance Σ_{E_n}.
Step 3 Compute the V^2 statistic using the MEWMA statistics and Σ_{E_n} from the previous steps.
Step 4 Determine when V^2 first exceeds the control limit h. This is the run length for the current iteration. If V^2 does not exceed the control limit, define the run length to be $\text{RL}_{\max}$.
Step 5 Take the average of all the run lengths for all iterations. This is approximation of the true average in-control run length.

Pseudo-code to determine the upper control limit for the MEWMA chart

Let ARL_0 be the pre-specified in-control average run length. Let (h_l, h_u) be the search interval for h. Let ρ be a small number that determines the accuracy of the result and let M be the maximum number of iterations.

Step 1 At the start of each iteration set $h = (h_l + h_u)/2$
Step 2 Approximate the in-control average run length using the pseudo code from the previous box for the value of h from the previous step.
Step 3 If the average run length from step 2 falls in the interval $(\text{ARL}_0 - \rho, \text{ARL}_0 + \rho)$ stop the search and return the current value of h.
Step 4 If the average run length from step 2 is larger than ARL_0, define $h_u = (h_u + h_l)/2$. If the average run length from step 2 is smaller than ARL_0, define $h_l = (h_u + h_l)/2$.

4.4.3 Regression adjusted variables

While the T^2 chart is very useful, it also has a disadvantage. The T^2 chart makes the implicit assumption that one is equally interested in all quality characteristics that need to be monitored. But it could be more likely that only a few of the characteristics are influenced by a change in the process. This issue is addressed in Hawkins (1991) in which an alternative method for multivariate SPC is introduced. The Regression Adjusted Variables (RAV) method described by Hawkins takes the correlational structure between the quality characteristics into consideration by regressing each characteristic on all the others.

Let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$ be p-component vectors that contain the measurements of the quality characteristics that one would like to monitor. We will use the notation $X_{i(j)}$ for the measurement at time i for the jth quality characteristic. The following assumptions are made for $\mathbf{X}$:

1. $\mathbf{X}$ follows a multivariate normal distribution $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$, where $\mu = \mu_0$ and $\Sigma = \Sigma_0$ when the process is in-control.

2. The in-control mean and covariance μ_0 and Σ_0 are known. Usually this is not true, but in this case it is assumed that the sample is sufficiently large, so that the estimates are very accurate.

3. Successive measurements $\mathbf{X}_i, \mathbf{X}_{i+1}$ are independent for all i.

4. The components of $\mathbf{X}$ are scaled to unit standard deviation, which implies that the diagonal elements of Σ_0 are equal to one.

Note that these assumptions require that the components of $\mathbf{X}$ are independent over time, but the p components at each time point are allowed to have nonzero correlation. To deal with this correlation Hawkins (1991) considered multiple linear regression of $X_{(j)}$ on the other $p-1$ components of $\mathbf{X}$. This regression can be written as

$$(X_{(j)} - \mu_{(j)}) = \sum_{i \neq j} \beta_{ji}(X_{(i)} - \mu_{(i)}) + \varepsilon_{(j)}, \tag{4.72}$$

where $\mu_{(j)}$ is the jth component of μ and $\varepsilon_{(j)}$ follows a normal distribution with zero mean and standard deviation $\sigma_{(j)}$. Assume that when the process becomes out-of-control, only a single variable is affected by this. The distribution of the process then becomes $\mathcal{N}_p(\mu_0 + \delta e_j, \Sigma_0)$, where δ is the size of the shift and e_j is a column vector with a 1 in the jth row and all other rows equal to zero. It is interesting to look at the following quantity

$$Z_{(j)} = \left[(X_{(j)} - \mu_{(j)}) - \sum_{i \neq j} \beta_{ji}(X_{(i)} - \mu_{(i)}) \right] / \sigma_{(j)}. \tag{4.73}$$

The scalar $Z_{(j)}$ is the scaled residual that is obtained by regressing $X_{(j)}$ on all other components of $\mathbf{X}$. However, it is not always known which components are affected when the process becomes out-of-control. Therefore, it useful to compute the statistic $Z_{(j)}$ for all the components of $\mathbf{X}$. The Z statistics can then be monitored using p individual charts. Alternatively, a composite measure for the p statistics can be determined, so that only one chart is needed to monitor the process. Both methods and their advantages and disadvantages will be discussed in the remainder of this section.

Charts for the individual components

For monitoring the individual components of Z, Hawkins (1991) suggests using CUSUM charts. This proceeds as follows:

$$\begin{aligned} \text{Define } & L^+_{0,(i)} = 0, L^-_{0,(i)} = 0, \\ & L^+_{n,(i)} = \max(0, L^+_{n-1,(i)} + Z_{n,(i)} - k) \\ & L^-_{n,(i)} = \min(0, L^-_{n-1,(i)} + Z_{n,(i)} + k). \end{aligned} \tag{4.74}$$

Since the $Z_{n,(i)}$ are scaled to a standard normal distribution, the control limit h can be determined in the same way as described in Section 4.3.2.

The disadvantage of using p separate charts to monitor the process is the issue with the increased false alarm rate as described in Section 4.4. Assume that the individual charts have a false alarm rate of α. Then, when the quality characteristics are all positively correlated, this false alarm rate exceeds $(1 - (1 - \alpha)^p)$. Therefore, we conclude that we are not sure what the false alarm rate exactly will be. This should be kept in mind when using this approach.

An advantage of using the p separate charts is that we can immediately see which characteristic is affected. For example, when monitoring the seven main bearings of the diesel engine. If a signal is given in the chart of bearing 1, this bearing should be inspected and maybe even replaced. If the charts from the other bearings do not signal, these bearings do not have to be inspected and time can be saved. In other words, the individual charts can be used to perform a root cause analysis.

Grouped charts

Hawkins suggests two diagnostics to combine the p individual Z statistics. These are defined by

$$\text{MCZ} = \max_{i=1,\ldots,p} \left\{ \max \left(L^+_{n,(i)}, -L^-_{n,(i)} \right) \right\} \tag{4.75}$$

and

$$\text{ZNO} = \sum_{i=1}^{p} \left(L^{+}_{n,(i)} + L^{-}_{n,(i)} \right)^2 . \tag{4.76}$$

The MCZ and ZNO statistics will be monitored using CUSUM control charts. Because the $Z_{n,(i)}$ are correlated for different i, the control limits cannot be determined theoretically, but only via simulation. Since the limits do depend on the correlational structure, the control limits need to be determined for each specific case individually. A general table can therefore not be provided. The pseudo-code to determine the in-control average run length of the MCZ chart can be found in the box below. To compute the average run length for the ZNO chart, the same steps can be used, only in Step 3 ZNO statistic should be computed using (4.76). The upper control limit for the MCZ and ZNO charts can be determined using the same pseudo-code as for the MEWMA chart.

The advantage of using a grouped chart over using p individual charts is that a grouped chart gives a quick overview, which is more convenient to use. Furthermore, the false alarm rate of the grouped chart is known while this is unknown for the combination of individual charts. However, using the grouped chart it cannot be seen which variables are affected by the fact that the process becomes out-of-control. This is very useful information, so Hawkins (1991) suggest using a combination of the grouped and individual charts. In principle the grouped chart is used to monitor the process, but as soon as a signal occurs in the grouped chart, the individual charts can be used to get more insight in which variables are causing the signal.

Pseudo-code to determine the ARL_0 for the MCZ chart

Let Σ be the covariance matrix that describes the dependencies between the Z-statistics of the variables of interest. Let M be the number of iterations and let $\text{RL}_{\max}$ be the maximum allowed run length in each iteration.

Step 1 Generate the Z statistics by simulating $\text{RL}_{\max}$ observations from a multivariate normal distribution with zero mean and covariance Σ.
Step 2 Compute the CUSUM statistics as described in Formula (4.74).
Step 3 Compute the MCZ statistics as described in Formula (4.75).
Step 4 Determine when the MCZ statistic first exceeds the control limit. This is the run length for the current iteration. If the MCZ statistic does not exceed the control limit, define the run length to be $\text{RL}_{\max}$.
Step 5 Take the average of all the run lengths for all iterations. This is approximation of the true average in-control run length.

Summary and conclusions

In this section, several approaches for multivariate statistical process control are described. When the quality of a process is determined by multiple quality characteristics, it is better to use a multivariate approach than to use multiple univariate approaches. This is mainly because the correlational structure of the quality characteristics is ignored by a univariate approach. We discussed the T^2 and MEWMA charts which can be seen as multivariate extensions of the Shewhart and EWMA charts.

Moreover, the regression adjusted variables method was described. In this method, each quality characteristic is regressed on the others. The scaled residuals are then monitored via individual or grouped CUSUM charts. The methods based on predictive or recursive residuals are designed to detect an upward or downward trend in the quality characteristic. The regression adjusted variables method is designed to detect if the behaviour of one quality characteristic is different than what is expected given the values of the other quality characteristics. The regression adjusted variables method is therefore a useful extra tool to use in multivariate process control.

Chapter 5

Results

In this chapter the statistical methods as described in Chapter 4 will be applied to the data that was made available by the Royal Netherlands Navy. In the first section, variables are selected via the LASSO method and an MLR model including these variables is fitted and validated. Next, the univariate statistical process control methods will be applied to the predictive, recursive and hybrid residuals. Finally, the multivariate regression adjusted variables method will be applied and we discuss the recursive coefficients method.

5.1 MLR model

This section consists of two parts. In the first part, the variable selection process is described. During this process two main issues were encountered, namely highly correlated variables and a very large sample size. In the first part it is discussed elaborately how we dealt with these issues. In the second part the multiple linear regression model will be fitted and validated on the data sets.

5.1.1 Variable selection

Dealing with high correlations

In the exploratory data analysis, it was shown that many variables are highly correlated. This influences the variable selection process for several reasons. In the extreme case that two variables have a correlation of 1.00, it makes no sense to include both of them in a model, since one variable can be exactly derived from the other variable. What about the less extreme case where the correlation is high, but not exactly one? Firstly, you can ask yourself if adding a variable that is highly correlated with one of the already included variables adds a lot of information to the model. But it can also lead to more serious problems. Say we have a group of highly correlated variables. Then it can happen that none of the variables are considered to be a significant contribution to the model, while the group as a whole is in fact significant. If this happens, we delete variables that contain important information, which is of course not what we want.

So, having a data set that contains multiple variables that are highly correlated has an impact on the variable selection process. But what if highly correlated variables are selected anyway and end up in the model? When two or more predictors in a model are highly correlated, this is called multicollinearity. According to Alin (2010) the most serious effect of multicollinearity is that the standard errors of the coefficient estimates will be high. When we want to monitor the recursive coefficients as described in Section 4.1.4, this leads to a low signal-to-noise ratio. It then becomes very hard to see the difference between the in-control and out-of-control situations.

For these reasons, we are interested in finding clusters of variables so that within each cluster all variables have a high pairwise correlation. If we find such clusters, only one variable needs to be included in the data set to represent the entire cluster of variables. In this way we reduce the number of highly correlated variables in our data sets. To find these clusters, a common approach from graph theory is used. First, we define what we consider to be a high correlation. This threshold is set to 0.95. Then, the covariance matrix of the predictors is transformed to an adjacency matrix, by setting all covariances higher than the threshold to one and all covariances lower than the threshold to zero. This adjacency matrix then will be transformed to an undirected

graph. The final step is to find all maximal sized cliques of this graph. These maximal cliques are then the clusters with high pairwise correlations. We use this approach since there exists a collection of functions in R that is able to perform these graph theory algorithms automatically. The clusters that the algorithm returned can be found in the appendix in Table B.4 for the transit mode and in Table B.5 for the manoeuvre mode.

In Table B.4 it can be seen that in the transit mode multiple clusters are created for the oil temperatures of the various compartments. This is partly due to the correlation between the oil temperature of compartment 1 and compartment 2. These two compartments have a correlation of 0.9477, which is just below the threshold we defined. It also should be noticed that the return oil temperature is only highly correlated with compartments 4 and 6 and not with the oil temperature of the other compartments in the transit mode. However, the correlation with the other compartments is around 0.92. In the manoeuvre mode, all compartments have a correlation around 0.97 with the return temperature. We decide to adjust the threshold for highly correlated variables to 0.947. In this way, clusters 4 and 5 of the transit mode are merged, but there still is a separate cluster for the return oil temperature. For completeness an overview of the new clusters can be found in Table B.6. The clusters for the manoeuvre mode do not change when the threshold is adjusted. Changing the threshold to 0.94 introduces two extra clusters, that do not have a clear interpretation. That is why we choose to work with a threshold of 0.947 instead of 0.94.

We chose the threshold of 0.947 for highly correlated variables. However, we performed the same analysis with a threshold of 0.90. As an example, the clusters for the transit mode using this threshold are given in Table B.8. The first thing that can be seen is that there are now more clusters. Another result is that several variables are in more than one cluster. This means that the clusters overlap. When we set the threshold to 0.947 there was less overlap. Eventually, we chose to use the threshold of 0.947 because we are only interested in very strongly correlated variables. Since we delete some variables after the process, we want to make sure that we do not delete too many variables.

The next step is to select one variable per cluster. This variable then represents the cluster. The variables that were selected for each cluster can be found in Table 5.1 for the transit mode and in Table 5.2 for the manoeuvre mode. The remaining variables of each cluster are removed. Because not all variables were included in clusters in the first place, we have 21 predictors left for the transit mode, and 23 predictors for the manoeuvre mode. The 21 variables that are left for the transit mode are included in the 23 variables for the manoeuvre mode. Table B.7 in the appendix shows which variables these are. Later, 10 of these variables will be chosen to be included in the model.

Table 5.1: Variables that are selected to represent each cluster in the transit mode.

Cluster	**Variables**
1	TEMP.LT.ZOKOW.RETOUR.HVD.BB
2	TEMPERATUUR.AFVOERGASSEN.NA.T.C.A.HVD.BB
3, 4	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
5	TOERENTAL.HVD.BB
6	TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB

Table 5.2: Variables that are selected to represent each cluster in the manoeuvre mode.

Cluster	**Variables**
1	TEMPERATUUR.AFVOERGASSEN.NA.T.C.A.HVD.BB
2	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
3	TOERENTAL.HVD.BB
4	TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB

We chose to select the variables that are easy to interpret. For the first cluster in the transit mode, the return temperature of the freshwater cooling is selected. This is because the supply temperature is controlled by the engine and is therefore more difficult to interpret. The temperature of the exhaust gases after turbo charger A was selected for the second cluster of the transit mode and the first cluster of the manoeuvre mode. For clusters 3 and 4 of the transit mode and cluster 2 of the manoeuvre mode, the oil temperature of compartment 6 were included. We chose to include compartment 6 to 'merge' the clusters for the transit mode. Choosing compartment 4 would have had the same effect. However, we do not have a reason to believe one choice would

be better than the other. Therefore, we chose to include compartment 6. Next, the RPM was added for the second last clusters for both propulsion modes. According to the engineers, compared to the other parameters in this cluster, the RPM of the turbocharger has the most influence on the temperature of the main bearings. However, because of the large correlation we chose to include the regular RPM to make the model easier to interpret. Finally, the average temperature of the exhaust gases of the cylinders connected to turbo charger A was added, because adding the average is more convenient than adding the temperature of a specific cylinder. The average of turbo charger A was used, because we also used turbo charger A in the previous clusters.

Large sample size problem

Both data sets contain over hundred thousand observations. These are very large sample sizes, which means that the standard methods for variable selection should be applied with caution. This issue is also addressed in Lin et al. (2013). When the sample size is very large, even the smallest effects can become significant. That is statistically significant, which is not always the same as practically significant. Small effects become significant, because the standard error decreases when the sample size increases. This means that the distance between the test statistic and the null hypothesis can easily get large, since this distance is measured in units of standard errors.

This behaviour is also present in our data set. For example, when we use the data set of the transit mode (approximately 200,000 observations) to fit a multiple regression model including all predictors and their squares, all the regression coefficients are statistically significant at a significance level of $\alpha = 0.05$. While this is statistically true, this does not help us selecting variables that should be included in the regression model. The model will be used by the engineers that do the actual maintenance on the diesel engines. This means they should be able to interpret the model. In this specific case the model would include 42 variables, which makes it very hard to interpret.

Suggestions to deal with this problem have been made by Lin et al. (2013). One suggestion is to adjust the significance level. Instead of $\alpha = 0.05$, one could use $\alpha = 0.01$, $\alpha = 0.001$ or an even smaller value of α. However, it might be hard to determine the correct value of α. Another suggestion that is made is to draw smaller samples that yield 'normal' significance levels of $\alpha = 0.05$. Of course, this should be repeated multiple times to gain information about the variation in the results. This is the approach that we will use to select the variables for the regression model. First, we should determine how large the size of the smaller samples should be. Following the rule of thumb from Harrell (2015), on average a sample size of 15 times the number of predictors should give a reliable model. We will apply this rule to determine the sample size of the sub-samples.

So, sub-samples are drawn multiple times and within each sub-sample the variable selection method will be performed. In this thesis the LASSO method is used to select variables within the sub-samples. How many times this procedure will be repeated and how to combine the results from different samples will be discussed in the remainder of this section.

As explained in Chapter 2, we would like to monitor the temperatures of the main bearings. Therefore, the LASSO method will be performed on each of the bearings. However, this could result in 14 different models (one model per bearing, per propulsion mode). This is not desirable, since this makes it hard for the engineers to use the models in practice. Therefore, we will combine the results of all the bearings in the same propulsion mode. This will lead to a maximum of two models, which is much better than 14 models. There might be a difference between the two propulsion modes, since the engine is used in different ways. This could lead to different variables being selected as important. Between the seven bearings we do not expect such differences, that is why the results will be combined. The variable selection procedure will be performed as follows:

1. For all the bearings in the transit mode and the manoeuvre mode, select 500 random sub-samples. Each sub-sample should have size 15 times the number of variables that is included in the data set. The variables, their squares and all second order interactions are included. So, for the transit mode this leads to sub-samples of size 3780 ($15 \cdot 252$).

2. Perform the LASSO method within each of the sub-samples. Each variable will be given a score which is equal to the number of times that variable is selected. So, for each bearing in each propulsion mode, the score will be between 0 and 500.

3. Add the scores of the variables for all bearings in the same propulsion mode. This will lead to a score between 0 and 3500 for each variable.

4. For each propulsion mode, select the ten variables with the highest score. This results in two models with ten variables each.

5.1.2 Results and validation

The methods that are described in the first part of this section are applied to the transit and manoeuvre data sets. It turns out that the selected variables are exactly the same for the transit and manoeuvre modes. This is quite convenient, since we now have one model that can be used in all situations. The variables that are selected are displayed in Table 5.3.

Table 5.3: Variables included in the multiple regression model for the transit and manoeuvre mode.

Variables
TOERENTAL.HVD.BB
DRUK.REGELLUCHT.HVD.BB
DRUK.LU.NOODSTOP.HVD.BB
WATER.IN.SMEEROLIE.HVD.BB
DRUK.CARTER.HVD.BB
TEMPERATUUR.SMO.TOEVOER.HVD.BB
DRUK.SMO.HVD.BB
TEMPERATUUR.HT.ZOKOW.RETOUR.HVD.BB
TEMPERATUUR.HT.ZOKOW.TOEVOER.HVD.BB
DRUK.VULLUCHT.HVD.BB

Since the selected variables are exactly the same for the transit and manoeuvre mode, we can combine the models for the two propulsion modes. This has the advantage that we do not have to make separate control charts for each propulsion mode and this reduces the output that the engineers have to process. While the predictors that are selected for both modes are equal, the coefficients of these predictors in the regression model could still be very different. Therefore, we will investigate if the propulsion mode should also be added to the regression model.

First, the data sets of the transit and manoeuvre mode are merged to one combined data set. In this combined data set an indicator variable (I) is added. The indicator variable equals zero when the ship sails in the transit mode and one when the ship sails in the manoeuvre mode. Adding the indicator variable to the regression model can lead to a change in mean for the manoeuvre mode with respect to the transit mode. However, we also want to add the effect that that the propulsion mode has on the other predictors. Therefore, not only the indicator itself but also the interaction terms of the indicator with all of the predictors from Table 5.3 will be added to the regression model. Let $X_1, \ldots, X_{10}$ be the predictors from Table 5.3, and let Y_i be the bearing temperature of bearing i for $i = 1, \ldots, 7$. We can now define two regression models.

$$\text{model 1:} \quad Y_i = \beta_0 + \beta_1 X_1 + \ldots \beta_{10} X_{10} \tag{5.1}$$

$$\text{model 2:} \quad Y_i = \beta_0 + \beta_1 X_1 + \ldots \beta_{10} X_{10} + \beta_{11} I + \beta_{12} X_1 * I + \ldots \beta_{21} X_{10} * I \tag{5.2}$$

To determine if the propulsion mode should be added to the regression model, we perform a likelihood ratio test. The likelihood ratio test is designed for nested models. From the formulas above it can be clearly seen that model 1 is nested within model 2. The likelihood ratio test is performed on all seven bearings. This resulted in p-values that are all equal to zero. From this we conclude that adding the extra terms for the propulsion mode makes the model significantly better. Therefore, model 2 is the final model.

Some of the choices we made during the variable selection process are not based on literature. For example, the number of sub-samples (500) and the number of variables to include (10). This makes it extra important to validate the model we end up with. The variable selection method makes use of all the available data, including data where the bearings are already severely damaged. This is because we want to know which variables have the most influence on the bearing temperatures in all states. However, model validation should be done in a

stable period. In a stable period, the residuals of the model should be normally distributed and no large outliers should be present. On top of that, the model should be able to predict independent data from the stable period with high accuracy. Outside the stable period, some large prediction errors and outliers might or should be present. These outliers make it possible to detect failures in the bearings, which is the final goal. Therefore, the validation of the model should be performed on a stable period.

To find a stable period, we first fit the model on the complete data set. The stable period should not be close to maintenance because before maintenance actions take place, there probably is unstable behaviour caused by the degradation of the bearings or other parts of the engine. After maintenance, new parts have to wear in which also might cause unstable behaviour. Since the residual plots are quite noisy, it is hard to select a stable period from these plots with the bare eye. That is why we will look at the rolling means of the residuals. In Figure 5.1 the rolling mean of the residuals from the first bearing is displayed. For the stable period, we look for periods where the rolling mean is approximately constant. In this figure, the dashed lines indicate the candidate stable period. To verify this choice for a stable period, we also look at the rolling standard deviation which is displayed in Figure 5.2. In a stable period, a small and constant standard deviation is expected. The standard deviation is not constant in the selected period, but it can also be seen that this is the best we find in the data set.

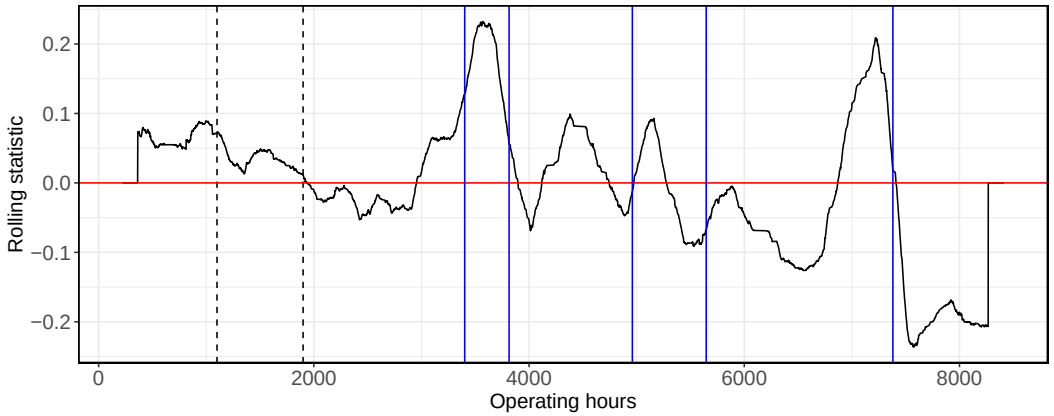


Figure 5.1: Rolling mean of the residuals of bearing 1. The dashed lines indicate the candidate stable period. The blue lines indicate when maintenance took place.

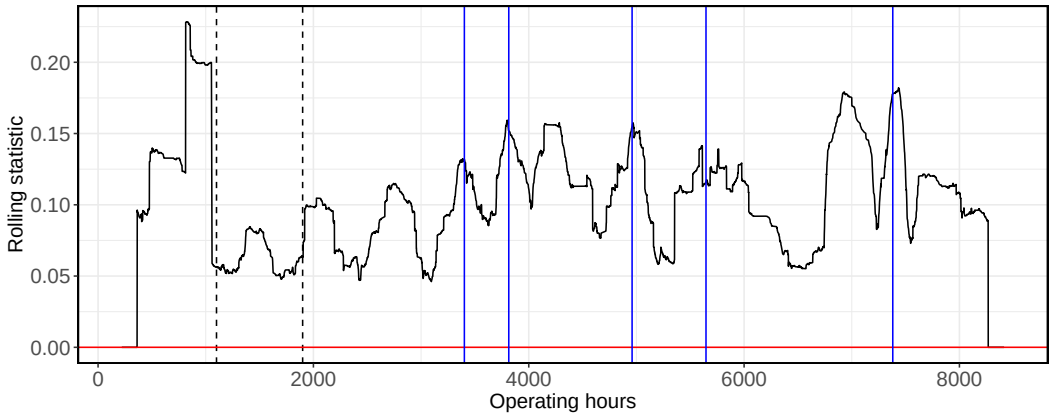


Figure 5.2: Rolling standard deviation of the residuals of bearing 1. The dashed lines indicate the candidate stable period. The blue lines indicate when maintenance took place.

Now that the stable period is fixed, we can validate the regression model. The first step of the validation process is to refit the model on the stable period. An important assumption for the statistical process control methods that will be applied later, is that the observations are independent over time. First, we will check if this assumption holds. Figure 5.3 shows the autocorrelation plot of the residuals from the first bearing. Significant positive autocorrelations are obtained for all time lags. Therefore, the assumption of independent observations is violated. It should be discussed how realistic this assumption is in the context of this research. In this

research, a time series of a physical process is studied. By assuming that the observations are independent over time, it is supposed that the process has no memory. This is very unrealistic since we are looking at a degradation process which depends on time among other factors.

The discussion above suggest that a trade-off needs to be made between meeting the assumptions of a mathematical model and keeping the characteristic properties of a physical process. We chose to downsample the data to one observation per hour. The resulting Autocorrelation Function (ACF) can be viewed in Figure 5.4. We can still see some positive significant peaks, but downsampling the data even further will make it harder to detect the degradation process on time. Therefore, we accept that the data is not perfectly independent over time and the data set that is downsampled to one observation per hour will be used in the continuation of this research.

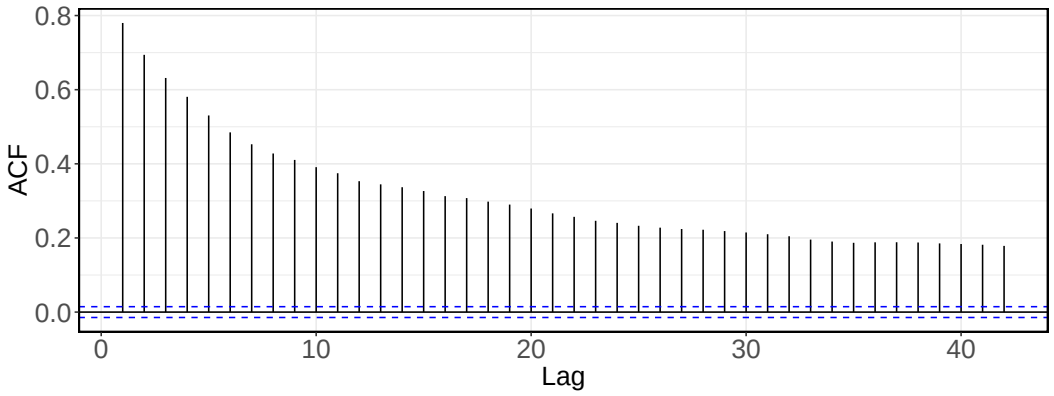


Figure 5.3: ACF for the residuals of the first bearing.

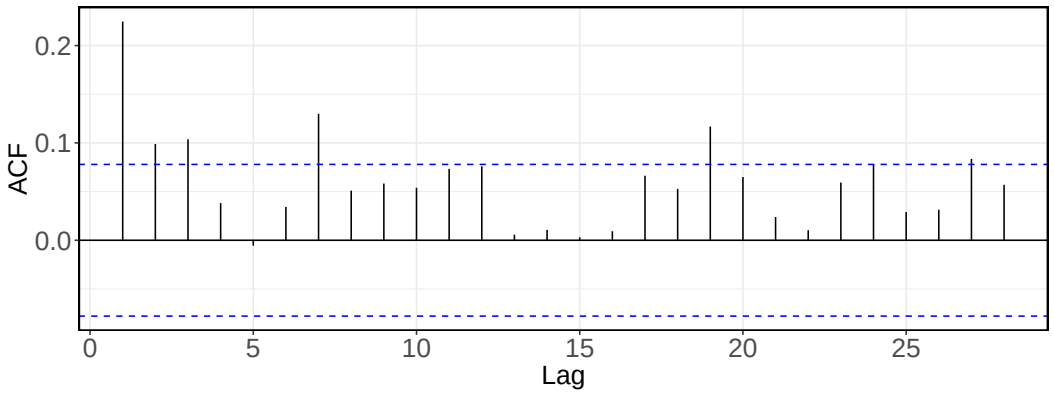


Figure 5.4: ACF for the residuals of the first bearing when the data is resampled to one observation per hour.

The densities of the residuals are displayed in Figure A.2. It can be seen that for all bearings the densities are centred around zero, which is as expected. However, the tails seem to be heavier compared to a normal distribution. In the QQ-plots from Figure A.3 we see clear deviations at the lower tails for all the bearings. From this it can be concluded that the residuals are not perfectly normally distributed, but if the model is able to predict independent data, this is acceptable.

To assess the prediction qualities of the model, 10-fold cross validation will be used. The results are displayed in Table 5.4. For each bearing we performed 10-fold cross validation and computed the average of the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R^2 values over the folds. The RMSE varies between 0.03 and 0.06. The RMSE has the same measure of unit as the dependent variable. Since all variables were standardized the standard deviation of the bearing temperatures is one. The RMSE values are very small compared to one, which means that the prediction error is relatively small. The MAE values vary between 0.02 and 0.05. Which is also relatively small. Furthermore, it can be seen that the RMSE and MAE are really close. The difference between the RMSE and the MAE is that the RMSE is a quadratic measure and the MAE is a

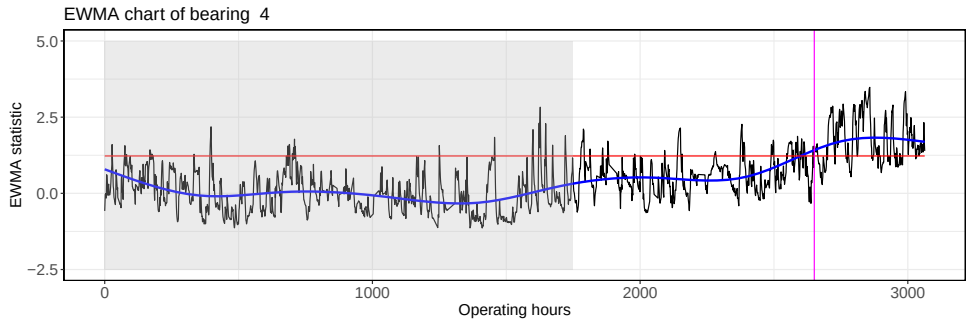
linear measure. A quadratic measure assigns more weight to large deviations. That the RMSE and MAE are close means that all prediction errors are roughly of the same magnitude.

Table 5.4: Results of performing 10-fold cross validation in the stable period for each bearing.

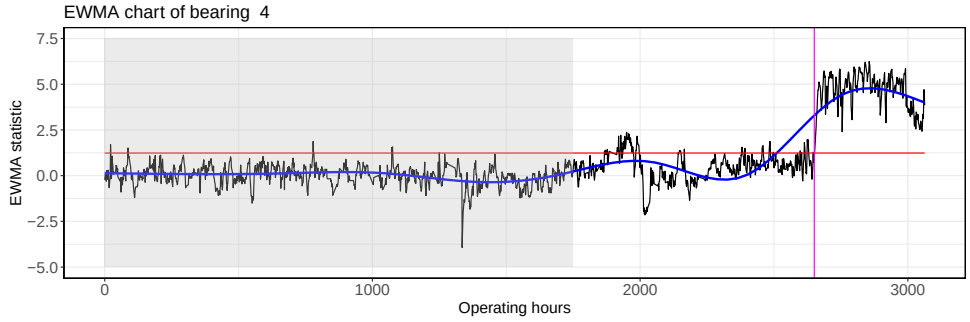
Bearing	**RMSE**	**MAE**	**R^2**
1	0.046	0.034	0.996
2	0.038	0.028	0.997
3	0.033	0.025	0.998
4	0.031	0.024	0.998
5	0.030	0.023	0.998
6	0.034	0.025	0.998
7	0.058	0.043	0.995

The R^2 value gives the amount of variation in the dependent variable that is explained by the predictors. In Table 5.4 it can be seen that the R^2 values are higher than 0.99 for the bearings in each propulsion mode. Such a high R^2 value could be an indication of overfitting when it is computed on the complete data set. But we computed the R^2 value on independent test data, which means the model almost explains all variation in the bearing temperatures.

To finish this section, it is interesting to compare the results of the regression model we came up with in this research with the regression model proposed in Heek (2021). Figure 5.5 shows the predictive residuals of both regression models when they are fit on the same initialization period. It can be seen that the signal-to-noise ratio of the model in Figure 5.5b is higher than the signal to noise ratio in Figure 5.5a. A large signal-to-noise ratio is desired. Therefore, we conclude that regarding this aspect the model we designed is an improvement over the model of Heek.



(a) Predictive residuals of regression model proposed in Heek (2021).



(b) Predictive residuals of regression model proposed in this research.

Figure 5.5: Comparison between the regression model of this research and the model proposed in Heek (2021).

Summary and conclusions

To reduce the number of highly correlated variables in the data, clusters of pairwise highly correlated variables were determined. For each cluster one variable was selected to represent the cluster in the data set. Next, LASSO variable selection was performed on multiple sub-samples to overcome the large sample problem. Since the selected variables are the same for the transit and manoeuvre mode, the data sets can be combined. Adding the propulsion mode to the regression model improved the model significantly. To validate the final model, a stable period was selected. The residuals that are obtained using the model were correlated over time. Therefore, the data was resampled to one observation per hour. Finally, 10-fold cross validation was used to investigate the predictive quality of the model. From the RMSE, MAE and R^2 values we can conclude that the final model can adequately predict independent test data in the stable period. This makes the model appropriate for monitoring the main bearings. Regarding the signal-to-noise ratio, this model is an improvement on the model proposed in Heek (2021).

5.2 Univariate SPC

In this section the results from the univariate statistical process control methods will be discussed. We first discuss the predictive residuals and then the recursive residuals. Finally, a hybrid method where we combine the recursive and predictive residuals will be explained and reviewed. To assess the quality of the various methods we must look at two criteria. Firstly, the chart must signal in case a failure occurs. Secondly, the chart should not signal when there is stable behaviour. To check these criteria, five periods are defined. They can be found in Table 5.5. In an ideal situation there would be no signals in periods $1, 3$ and 4 and multiple signals towards the end of periods 2 and 5.

Table 5.5: Definition of the five monitoring periods.

Period	Description
1	Stable period as defined in Section 5.1
2	Period between maintenance and a failure
3	Period between two maintenance actions
4	Period between two maintenance actions
5	Period between maintenance and a failure

The methods for statistical process control that are described in Section 4.3 and Section 4.4 use lower and upper control limits to determine whether a process is in control. In the degradation process of the main bearings, the temperature starts to rise if the quality of the bearings decreases. To detect these changes monitoring with an upper control limit is sufficient. This was proposed to the engineers and they confirmed that disregarding the lower control limit is a good choice.

5.2.1 Predictive residuals method

The predictive residuals require a stable period on which the model can be fitted. We will use period 1 as defined in Table 5.5 as the stable period. The model that is fitted in period 1 will be used to make predictions in periods 2 to 5. The predictive residuals are monitored using Shewhart and EWMA charts. In this section the results from the Shewhart charts will be discussed first, followed by the results of the EWMA charts.

Figure 5.6 shows the predictive residuals of the first period. Notice that this first period is also used to fit the model. As can be seen in the figure, the chart gives one signal. This signal is considered a false alarm, since we already know that there was stable behaviour in this period. In total, 633 observations are registered during this period, from which one observation is a false alarm. This false alarm rate is lower than the theoretical false alarm rate of 1 false alarm every 370 observations. However, there are also bearings that show three false alarms during the first period, so overall this is in line with the theoretical false alarm rate.

Figure 5.7 and Figure 5.8 show the Shewhart charts of the predictive residuals in the second period. These two figures immediately show the disadvantage of the predictive method. After maintenance actions are performed the model from the stable period becomes biased. Figure 5.7 shows that the residuals are centred around 2.5,

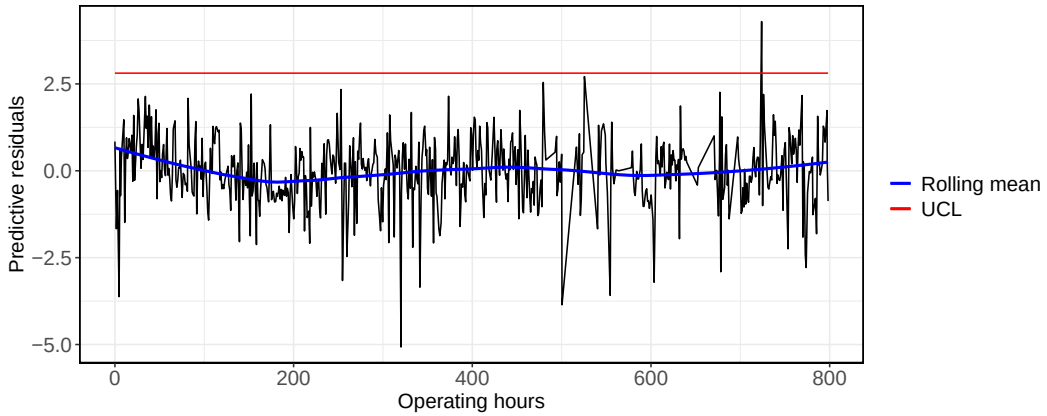


Figure 5.6: Shewhart chart of predictive residuals in period 1 (stable period) for bearing 1.

instead of zero as should be the case when a valid model is used. This effect is even larger in Figure 5.8. Here, the residuals are centred around 5 for the first 300 operating hours. This clearly shows why the model should be refitted after maintenance, as Heek (2021) already concluded in his research.

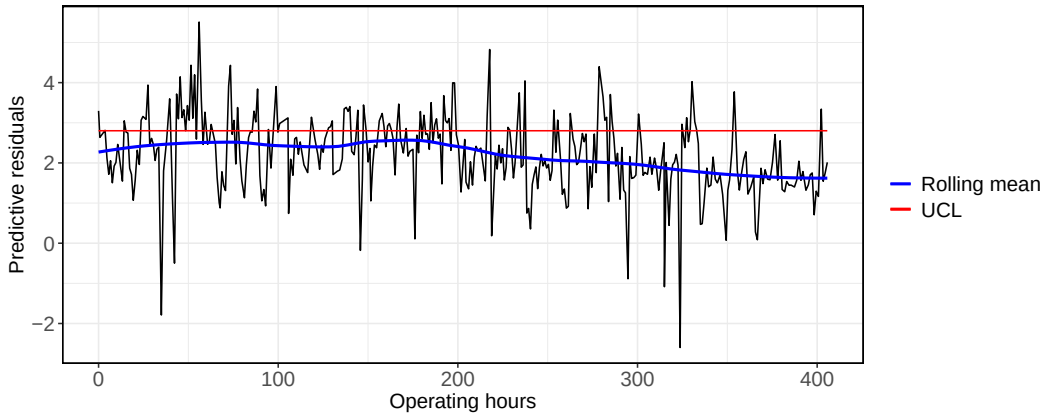


Figure 5.7: Shewhart chart of predictive residuals in period 2 for the non-failing bearing 2.

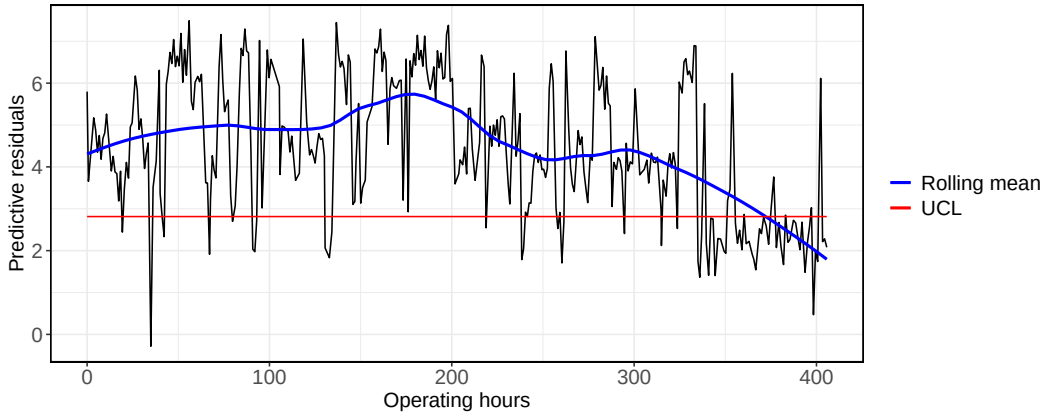


Figure 5.8: Shewhart chart of predictive residuals in period 2 for the failing bearing 4.

While the predictive residuals are not the method we will recommend in this thesis, these charts still show some interesting patterns that we will discuss. In Figure 5.8, it can be seen that the residuals of the temperature of the failing bearing 4 are roughly two and a half degrees Celsius higher than the temperature of the other non-failing bearing 2, starting immediately after maintenance. This phenomenon was discussed with the engineers and they came up with the following explanation. During maintenance the crankshaft is rotated without lubrication oil, which could damage the bearings. It is very plausible that the fourth bearing

was already damaged (but to such a level that the temperature not yet increased) and that the existing damage was aggravated by the rotations without lubrication. This could result in the higher temperatures for the fourth bearing. This suggested specific degradation process could not be verified. However, it is clear that during maintenance something happened to the fourth bearing, which could have been detected from the control chart.

Another pattern that is visible in the charts for all the bearings, is that after approximately 300 operating hours, the residuals start to decrease. However, we expect that the residuals of the bearing that fails at the end of this period would increase. A possible explanation for this behaviour given by the engineers is that the damage that is caused by the rotations of the crankshaft without lubrication is evened out after 300 operating hours. This would be possible, since the top layer of the bearings is made of a relatively soft and ductile material. However, the same decreasing trend is also observed for the other (undamaged) bearings, which suggests there might be another (common) cause for this behaviour.

The behaviour in the third period is quite similar to the behaviour in the first period, except for the fact that the residuals are biased in the third period. In this period, we also see an oscillating pattern in the residuals that repeats approximately every 250 operating hours. This is addressed in the discussion of the charts of the fifth period, since these charts show a similar behaviour. Since no further insights can be gained from the charts in the third period, they are omitted in this report.

Figure 5.9 shows the Shewhart chart of the predictive residuals in period 4. During this period, we expect stable behaviour, but between 500 and 1000 operating hours the residuals are extremely low and the variation is relatively high. This was also discussed with the engineers and they reported an electrical problem. During this period, there were problems with the measuring and monitoring systems of the ships, which resulted in measurements that largely deviated from reality. Therefore, the fourth period should no longer be considered to validate the SPC methods we developed.

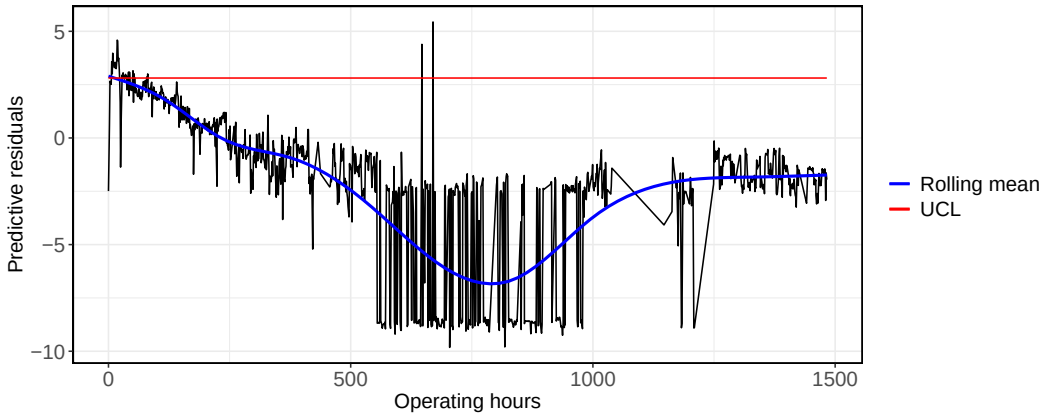


Figure 5.9: Shewhart chart of predictive residuals in period 4 (between two maintenance actions) for bearing 1.

Figure 5.10 shows the predictive residuals of the failing bearing 3. In this figure, no signals are visible, which means that this chart does not give useful information about the quality of the bearing. However, we see an oscillating pattern with a wavelength of approximately 250 operating hours. This pattern was also visible in the charts of period 3. The pattern is probably caused by small maintenance called 'readjusting the cylinder valves'. This action is performed every 250 operating hours and influences the thermodynamic behaviour of the engine, and therefore slightly affects the temperatures in the bearings.

Until this point, we only discussed the Shewhart chart for the predictive residuals. From the Shewhart charts we can already conclude that the method based on solely predictive residuals is not suitable to solve our problem. The method results in a biased model after a maintenance action is performed. Earlier, we argued that we should also look at EWMA charts because these charts take the history of the process into account. However, in this situation, the EWMA charts will enhance the effect of the biased model even more than the Shewhart charts. Therefore, they will be omitted in this section. We will come back to these charts when we found a method that deals with the bias after maintenance actions.

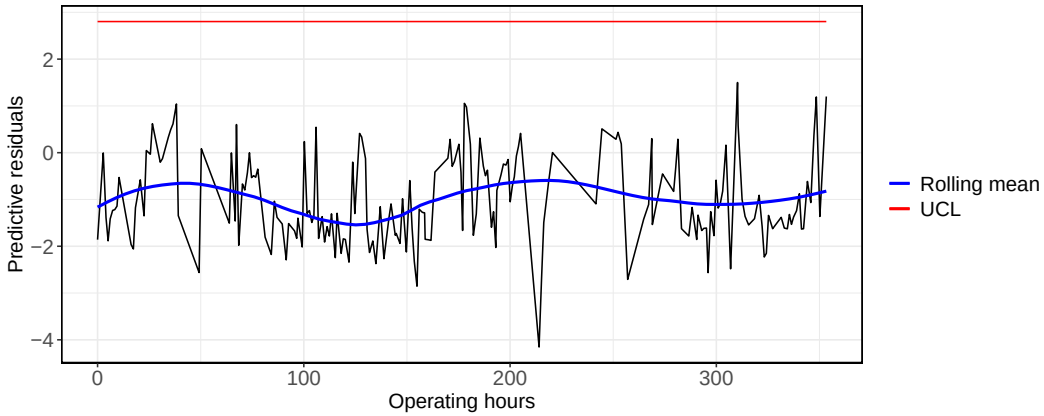


Figure 5.10: Shewhart chart of predictive residuals in period 5 for the failing bearing 3.

Summary and conclusions

In this section, we showed the results of the predictive residuals applied to the five periods of interest. The Shewhart charts confirmed what we described in the Exploratory Data Analysis. Namely, that the model fitted in the stable period is no longer valid after maintenance is performed on the engines. While this method proved to be inappropriate for predictive maintenance, we did retrieve some interesting results from this analysis. The bias after maintenance is different for the various bearings. A larger bias could be an indication that the bearing is damaged during maintenance. The engineers should then keep an eye on this bearing, since it is at risk. Furthermore, an oscillating pattern is visible in the data. This pattern is probably caused by readjusting the cylinder valves of the engine.

5.2.2 Recursive residuals method

In the previous section, the predictive method is discussed in great detail. In this section we will discuss the recursive method. The focus will lie on the advantages and disadvantages of the recursive method compared to the predictive method.

Figure 5.11 shows the recursive residuals during period 3 of bearing 4. The third period was a period between two maintenance actions where no failures were reported. Therefore, we expect stable behaviour during this period. In the Shewhart chart no signals are generated. Therefore, we can conclude that this chart is confirming the stable behaviour during the third period.

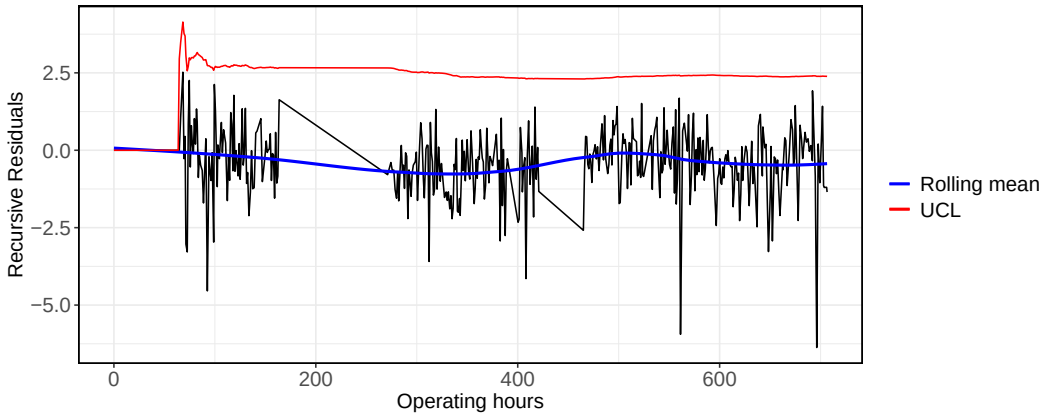


Figure 5.11: Shewhart chart of recursive residuals in period 3 (between two maintenance actions) for bearing 4.

A difference with the predictive residuals is that the recursive residuals are centred around zero. The bias that was present in the predictive residuals is no longer visible in the recursive residuals. This is because in the recursive method the regression model is reset after maintenance on the diesel engines is performed. When

the model is reset, it uses an initialization period of 20 observations to restart. After these 20 observations, the quality of the bearings can be monitored. The ability to restart after maintenance makes the recursive residuals appealing compared to the predictive residuals. Heek suggested in his research that an initialization period of 800 hours is necessary to refit the model after maintenance when the predictive method is used. The recursive method can reduce this initialization period to only 20 hours.

Figure 5.12 shows the EWMA chart of the same bearing and in the same period as Figure 5.11. The EWMA chart generates one signal during the third period. This signal should be interpreted as a false alarm. Keeping in mind that these charts are designed to have a theoretical false alarm rate of 1/370, this false alarm is not unexpected. Another aspect of the recursive residuals, that can be clearly seen in the EWMA chart, is that the control limits are constantly moving. This is because the control limits depend on the mean and standard deviation of the residuals in the regression model. Since the regression model is recursively estimated, the mean and standard deviation will be updated in every iteration. In the beginning, there are only a few observations, so the standard deviation will be large. When more observations follow, the standard deviation of the recursive residuals will converge to one. Since the standard deviation of the EWMA-statistic depends on the standard deviation of the residuals, this standard deviation will also converge.

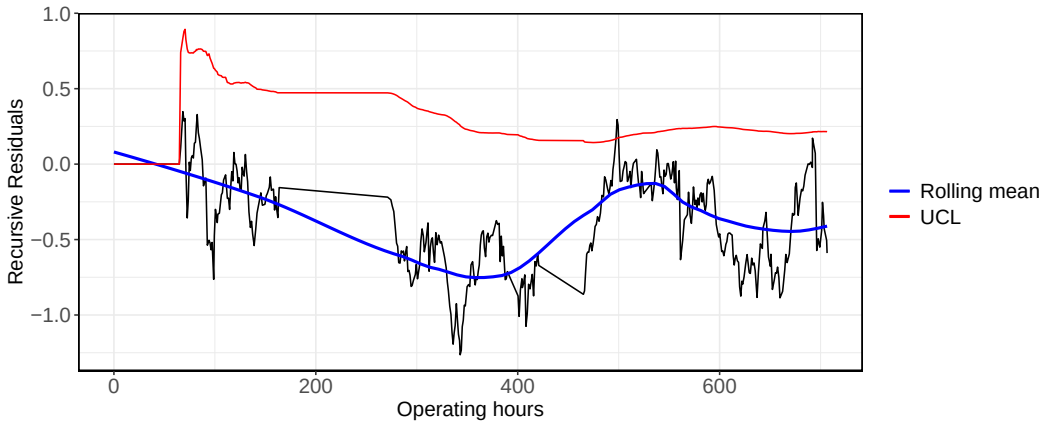


Figure 5.12: EWMA chart of recursive residuals in period 3 (between two maintenance actions) for bearing 4.

In Figure 5.13 the Shewhart chart of the third bearing in period 5 is shown. This bearing fails at the end of this period. Therefore, we would expect many signals towards the end. In the Shewhart chart, a weak upward trend in mean is visible, but this trend is not strong enough to cause multiple signals. The signal around 310 operating hours could easily be interpreted as a false alarm. Therefore, we conclude that the Shewhart chart based on the recursive residuals is not able to detect the failure at the end of the fifth period in advance.

Figure 5.14 shows the same bearing during the same period, but now an EWMA chart is used to monitor the recursive residuals. The upward trend in the bearing temperatures is clearly visible in the EWMA chart and some signals are generated towards the end. At this point, we conclude that the EWMA charts are preferred over the Shewhart charts. This is because the Shewhart charts have a low signal-to-noise ratio compared to the EWMA charts. The upward trend in temperature is more clearly visible in the EWMA chart than in the Shewhart chart. This conclusion is not only based on the charts that are shown in this report, but also on all the other charts of all bearings in all periods. These charts show the same characteristics as we just discussed for the chart of bearing 3 during period 5. Therefore, from now on, the Shewhart charts will be omitted and we only focus on the EWMA charts.

The EWMA chart performs better than the Shewhart chart during period 5, but this chart is still not very convincing. The two signals around 310 operating hours could be interpreted as false alarms. The final signal might convince the engineers of a problem, but this is too late, since the bearing fails after one operating hour. In the recursive residuals method, the regression model is constantly updated. This implies that the model learns a trend if this trend is present for a longer time. This behaviour is clearly visible in Figure 5.14. The upper control limit follows the upward trend towards the end of the period. This means that the recursive method adapts to trends in the data, which is an undesirable property because we want to detect these trends. In this example the trend is strong enough to cause some signals, but these signals are not very convincing.

Another disadvantage of the recursive method is that the method does not directly give the root cause of

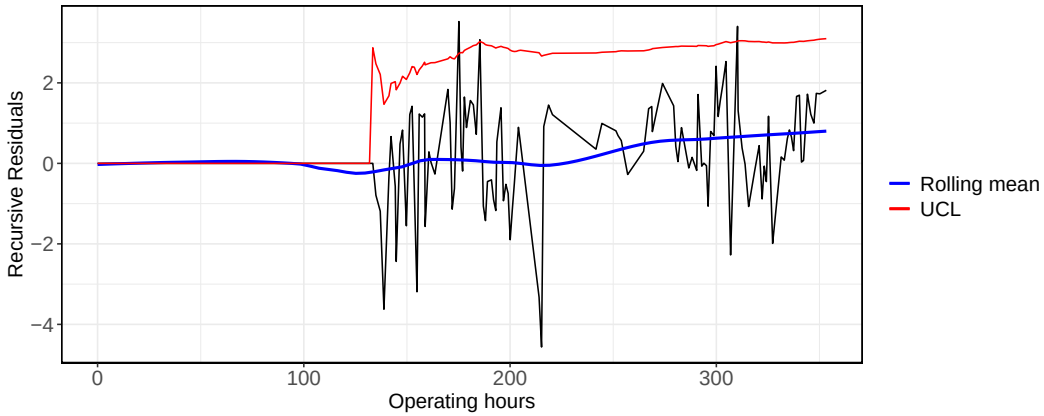


Figure 5.13: Shewhart chart of recursive residuals in period 5 for the failing bearing 3.

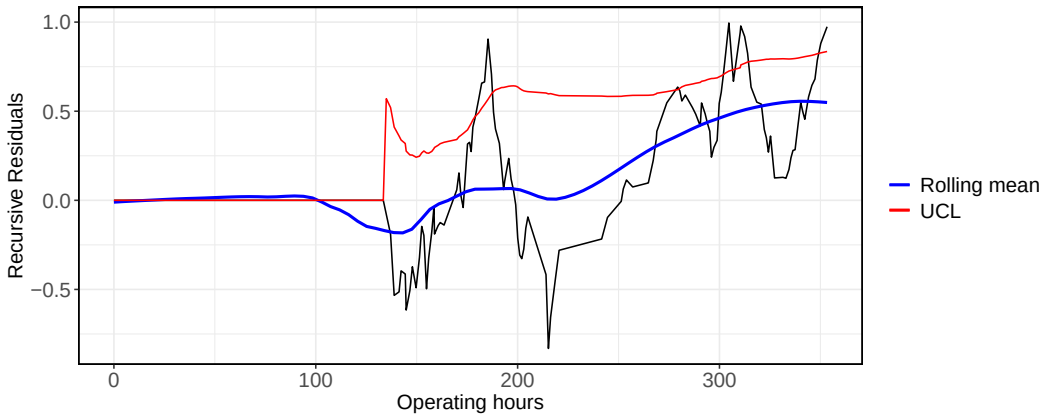


Figure 5.14: EWMA chart of recursive residuals in period 5 for the failing bearing 3.

the signals that are produced. In Figure A.4 in the appendix, all EWMA charts for the recursive residuals in the fifth period are presented. These charts all look very similar and various charts start signalling around the same time. Looking at these charts, one cannot predict which bearing is going to cause a failure.

Note that we did not discuss the results of the recursive residuals method during the second period. The EWMA charts (and the Shewhart charts) do not give multiple signals during this period. This is due to the fact that the temperature of the failing bearing decreases at the end of this period instead of increases. This issue will be discussed in more detail in the next section for the hybrid method. Since the charts of the recursive method in the second period do not give any further insights, they are omitted from the report.

Finally, we want to discuss another issue that becomes visible when you look at the charts of the recursive residuals. For the recursive residuals in the fifth period, it can be seen that the chart only starts after approximately 130 operating hours. This means that during the first 130 operating hours, the process cannot be monitored properly. This is caused by the combined regression model. The model cannot be fit until 11 observations in both the transit and manoeuvre mode are collected. The charts of period 5 show that this might take a while. Using the predictive method this does not cause a problem as long as enough observations in both modes are present in the stable period.

Table 5.6 shows how long it takes before the combined model can be used compared to the total number of hours in each monitoring period. The model minimally needs 11 observations in each propulsion mode to estimate the parameters. However, the parameter estimation of eleven parameters based on eleven observations is unreliable. Therefore, we demand a minimum of 20 observations in each mode before the combined model can be applied. The number of operating hours it takes to reach this number of observations is presented in the first column of the table. The last column of Table 5.6 shows the total number of operating hours in each period. Table 5.6 displays that there is a lot of variation in the time that passes until the combined model can be applied. In the first period, only 55 observations are needed before the combined model can be used. It might be acceptable if the process cannot be monitored during these 55 hours. However, it is not acceptable

if the process cannot be monitored for 312 hours as is the case in the second period. A simple, but practical solution can be implemented to overcome this issue. Until there are enough observations in both propulsion modes, the regression model of the mode that is sailed in the most should be used to construct the control charts. In case of period 5 this would mean that until 130 operating hours, one should use the model for the transit mode and disregard the observations that come from the manoeuvre mode. Of course, we still need 20 observations to initialize the model, but in this way the process can be monitored during the 110 extra hours it takes to collect a sufficient number of observations in the manoeuvre mode. From the moment there are enough observations in both modes, the switch can be made to the combined model and all observations can be used to monitor the process.

Table 5.6: Operating hours needed to estimate the parameters of the combined model.

Period	Hours needed	Total hours
1	55	789
2	343	405
3	64	706
4	312	1482
5	133	353

Summary and conclusions

In this section, the results from the recursive residuals are compared to the predictive residuals. An important advantage of the recursive method is that this method reduces the initialization period after maintenance from 800 hours to only 20 hours. A disadvantage of the recursive method is that persistent trends are learned by the model. Therefore, less signals than expected are visible before a bearing fails and thus the failure is not detected in advance. It was concluded that the EWMA charts are preferred over the Shewhart charts, since the EWMA charts have a higher signal-to-noise ratio and the upward temperature trend is more clearly visible in these charts.

Using the recursive residuals, a disadvantage of the combined model became visible. There need to be 20 observations in both propulsion modes before the model can be used. It appears that the time it takes to collect these observations varies a lot. The historical data shows that this could take up to 343 hours. As an alternative it was proposed to monitor the bearings in the mode that is sailed in the most until enough observations for the combined model are collected.

5.2.3 Hybrid residuals method

The hybrid method was designed to benefit from the advantages of both the recursive and the predictive methods. The main advantage of the recursive method is that these residuals shorten the initialization period of the model. This makes it possible to restart the model after maintenance is performed on the engines. However, the recursive method will adapt to a trend if this trend is persistent and not too strong. To overcome this issue, we switch to the predictive method once the model obtained by the recursive method is "good enough". Once this switch is made the model and the control limits will be fixed.

Determination of the changepoint

Of course, the following question now should be answered: How do we determine if the model that is estimated by the recursive method is good enough? To answer this question, we looked at various characteristics to assess the quality of a regression model. These characteristics were studied in periods $2, 3$ and 5. Period 4 is dismissed because of the electrical issues leading to unreliable sensor measurements. Period 1 is also dismissed, since this period does not start directly after maintenance.

From this moment on τ is defined as the moment when the switch from the recursive to the predictive method is made. To assess the quality of the recursively estimated regression model, we looked at the adjusted R^2, RMSE, p-value of the F-test and AIC. These characteristics were determined for all bearings in every iteration.

Since it is convenient if the changepoint is the same for all bearings, these characteristics were combined. For the adjusted R^2 value, the minimum over all bearings is used. For the RMSE, p-value of the F-test and AIC the maximum over all bearings is used. The behaviour of these characteristics during the third period is displayed in Figure A.6 in the appendix. The figures for the other periods are omitted, since the behaviour of the characteristics is similar. It can be seen that the adjusted R^2 and the p-value of the F-test are approximately constant. Therefore, these characteristics cannot be used to determine τ. The AIC is not constant, but follows a linear, downward trend. One could set a threshold to the AIC and consider the model to be good enough, when the AIC drops below the threshold. However, determining this threshold is rather arbitrary since the range of the AIC is varies between the different periods. So, in this case study, the AIC also cannot be used to determine τ.

The characteristic that was not discussed yet is the RMSE. Figure A.6 shows that the RMSE increases during the first 40. From 40 hours onward the RMSE shows stable behaviour. Note that the RMSE cannot be used to determine when the model is "good enough", since the RMSE increases. However, the RMSE could indicate whether the quality of the model is stable. If the model shows stable behaviour, it is not useful to continue with the recursive method. Therefore, this is another criterion to switch from the recursive to the predictive method. Since the RMSE seems to be the most promising option, we will determine the changepoint τ based on the RMSE.

To determine when the RMSE starts behaving stable the following approach is used. In iteration i, the mean RMSE of models $(i-9,\ldots,i)$ is compared with the mean RMSE of models $(i-19,\ldots,i-10)$ for $i \geq 20$. If the difference in mean is less than five percent, the RMSE is considered to be stable. The changepoint τ is chosen as the iteration in which the difference in mean is less than five percent for the first time.

Application to historical data

In this section, the results of the hybrid method applied to periods 2, 3 and 5 will be discussed. First, we discuss the results in period 3, which are presented in Figure 5.15. The changepoint τ is visualized by the vertical dashed line. The EWMA chart of the first bearing is presented, since the charts of the other bearings show similar behaviour. Note that period 3 is a period between two maintenance actions in which no failures were reported. As can be seen in the chart, the signal does not cross the upper control limit, which means there are no false alarms. From this it can be concluded that the hybrid method performs as desired during the third period.

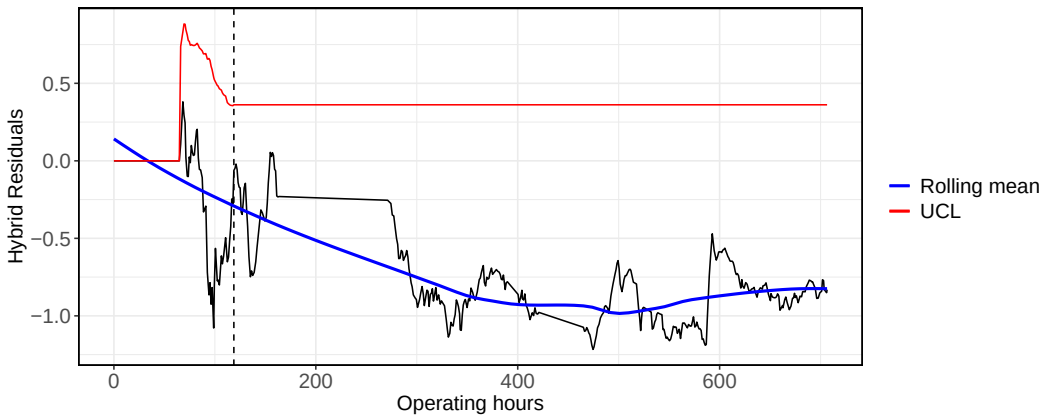


Figure 5.15: EWMA chart of hybrid residuals in period 3 (between two maintenance actions) for bearing 1.

Figure 5.16 shows the EWMA chart of the hybrid method for the fourth bearing in period 2. The fourth bearing fails at the end of this period. Note that the transit model is used to construct this chart. This is because it takes a very long time to obtain enough observations for the combined model during the second period. The chart does not give any signals, therefore we can conclude that the hybrid method does not detect the failure at the end of period 2 in advance. The methods we developed and implemented so far, are designed to detect a persistent upward trend in temperature. Looking at the predictive residuals from Figure 5.8, we see a downward trend towards the end. So, it should not come as a surprise that this failure is not picked up by the hybrid method.

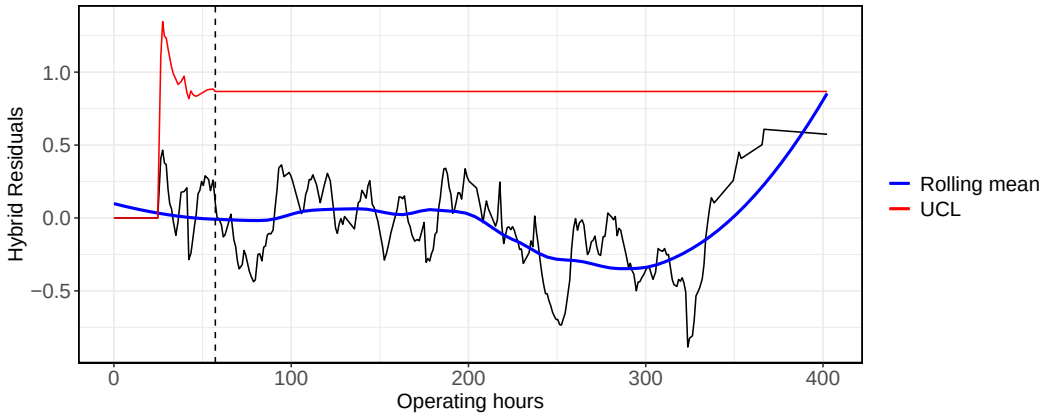


Figure 5.16: EWMA chart of hybrid residuals in period 2 for the failing bearing 4 using the transit model.

In Figure 5.17 the EWMA chart of the hybrid method for the third bearing in period 5 is displayed. The third bearing fails at the end of this period. As can be seen, from 270 hours onward, multiple signals are given by the chart. One should compare this result with the chart in Figure 5.14. Here, the recursive method is used which results in only one signal. This specific case shows that switching to the predictive method can improve the ability to detect a weak but persistent trend. Note that the chart in Figure 5.17 already signals around 180 operating hours. However, the upward trend that causes the signal is not persistent which is why the chart only signals a few times.

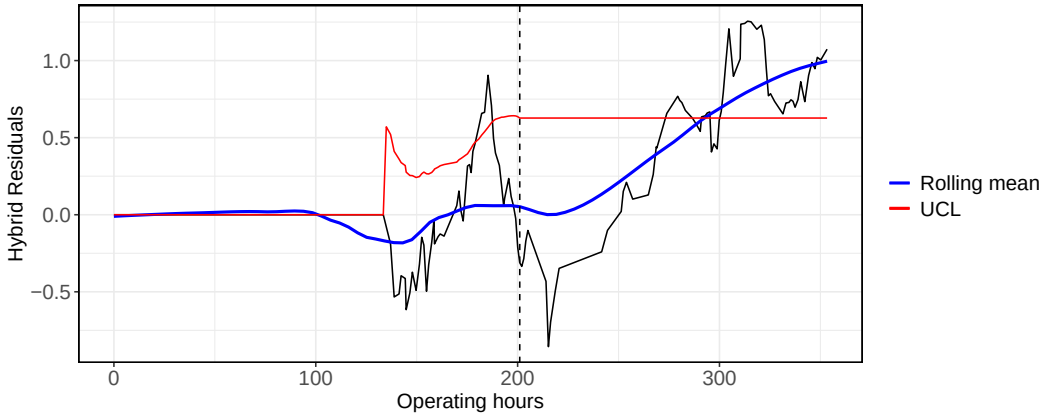


Figure 5.17: EWMA chart of hybrid residuals in period 5 for the failing bearing 3.

As might be noticed, it takes a while before the chart in Figure 5.17 can be used. This is because this chart is based on the combined model and it takes 133 hours before enough observations are collected in both propulsion modes. In this case, the transit model could be used to monitor the process in an earlier stage. The chart that results from the transit model can be viewed in Figure 5.18. Since the initialization period is shorter for the transit model, the switch to the predictive method can be made earlier. This results in multiple signals from 175 operating hours onward. This figure shows that it might be wise to keep monitoring the transit model, even though the combined model can be applied too.

Just like with the recursive method, a critical note should be made. In Figure A.5 in the appendix, all EWMA charts based on the hybrid method in the fifth period are presented. It can be seen that all charts start signalling at the same time as the chart of the third bearing. This means that failures in the bearings can be detected in advance, but the hybrid method cannot detect which bearing will fail. This implies that all bearings need to be investigated first, before maintenance or replacement of the problematic bearing(s) can take place. This investigation could take place manually by the engineers, or via a root cause analysis.

Until this point, we presented multiple EWMA charts, but the choice of the parameter λ has not yet been discussed. All EWMA charts that are presented in this chapter are generated using the scaling parameter $\lambda = 0.1$. According to (Qiu, 2013, Chapter 5) relatively small values of λ should be chosen to detect small changes in mean. The values of λ that Qiu suggest are $\lambda = 0.05$ for small changes in mean, $\lambda = 0.1$ for medium

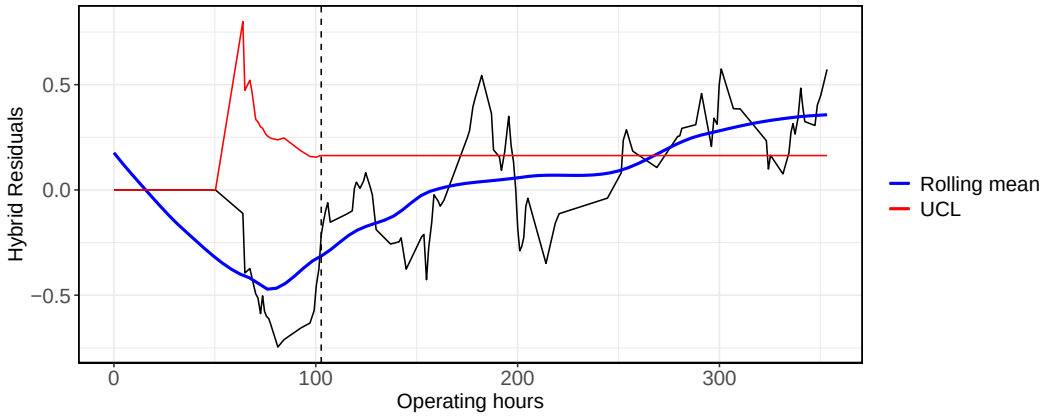


Figure 5.18: EWMA chart of hybrid residuals in period 5 for the failing bearing 3 using the transit model.

changes in mean and $\lambda = 0.2$ for large changes in mean. Since we do not know how large the change is that we want to detect, these three values for λ were tested on the historical data. The value of $\lambda = 0.1$ gave the fewest false alarms in stable periods. On top of that, this value of λ leaded to the earliest signal during the fifth period. All values of λ were unable to detect the failure at the end of the second period. For these reasons we chose to use $\lambda = 0.1$.

Summary and conclusions

In this section it was discussed how the changepoint of the hybrid method can be determined. The various characteristics that were examined, proved to be inappropriate to determine when the recursive model is "good enough". However, the RMSE can be used to determine when the quality of the model is (close to) constant. The changepoint is chosen as the point where the RMSE shows stable behaviour during a period of 10 observations for the first time.

The hybrid method was applied to the historical data. EWMA charts using a scaling parameter of $\lambda = 0.1$ were used to monitor the hybrid residuals. During the third period the method gives no false alarms, which is perfect in this situation. The method was unable to detect the failure at the end of the second period. This is because the method is designed to detect a persistent upward trend, which is not present towards the end of the fourth period. The hybrid method does detect the failure at the end of the fifth period. Using the combined model, the first of a series of multiple signals is generated 75 operating hours before the failure. For the transit model this is 190 hours before the failure. This corresponds to 28 calendar days for the combined model and 47 calendar days for the transit model. Finally, it should be mentioned that the hybrid method cannot be used to find out which bearing is going to fail, since the charts of all bearings start signalling at the same time.

It can be concluded that the hybrid method really is the best of both worlds. The predictive method gives too many false alarms, because of the bias after maintenance. The recursive method has a short initialization period, but it learns the upward trend and therefore only produces a few signals. In this section we showed a specific case in which it can be seen that the hybrid method can detect an upward trend using only a short initialization period. So, for this case, the hybrid method is an improvement of both the predictive and the recursive methods.

5.3 Multivariate SPC

In this section, the multivariate statistical process control methods will be applied to the historical data of the Royal Netherlands Navy. First, the regression adjusted variables method will be discussed extensively. Next, the recursive coefficients method will be discussed. We were unable to implement this method successfully. The issues we encountered will be discussed.

5.3.1 Regression adjusted variables

The Regression Adjusted Variables (RAV) method looks at the monitoring problem from another perspective. The methods based on the predictive, recursive and hybrid residuals are designed to detect an upward trend in temperature. However, we already saw that the failure at the end of the second period cannot be detected using these methods. The RAV method is designed to detect if and when one of the bearings start to deviate from its expected behaviour given the other bearings. So, this is a very different approach to the monitoring problem.

First, we will discuss how we implemented the RAV method in practice. It is argued why the raw data is not appropriate to use in our case. Next, it will be discussed how the parameters k and h for the CUSUM charts are selected. We also discuss whether to use one-sided or two-sided control charts. Finally, the RAV method is applied to the historical data and the results will be discussed.

Parameter selection and determination of control limits

In Section 4.4.3 the assumptions for the characteristics that are monitored via the RAV method are given. It can easily be seen that the raw temperature data violates all of the assumptions. Nevertheless, we applied the RAV method on the raw temperature data of the bearings. As might be expected, the resulting control charts were useless. As an alternative, the recursive residuals of the bearing temperatures are used as input for this method. The models described in Section 5.1 are used to obtain these residuals. The recursive residuals meet the first, third and fourth assumption of the RAV method. The second assumption (the in control mean and covariance are known) is violated. However, Hawkins (1991) already mentions that this is usually not a problem if the sample size is large enough. Since recursive estimation is used, this might be a problem at the start of the monitoring process, but this issue becomes less relevant when more observations are collected over the time.

Since, CUSUM charts are used to monitor the Z-statistics, which are the output of the RAV method, the parameters k and h should be chosen. It is common practice to pre-define the desired ARL_0 and choose a value for k. Then the value of h is chosen such that the desired ARL_0 is achieved. The ARL_0 is chosen as 370 just as in the previous methods. According to the theory, the chart performs optimal when k is chosen as $\delta/2$, where δ is the mean shift, we would like to detect. It hard to determine this δ, since we are looking at regression adjusted variables of the recursive residuals, which is very abstract. Different values of k are applied and $k = 0.25$ showed the best results. By "best results" we mean few false alarms during periods between maintenance and multiple alarms for bearings that failed at the end of a given period. Choosing this value of $k = 0.25$ means there is looked for mean shifts of size $\delta = 0.5$. Since, the regression adjusted variables of the recursive residuals follow a standard normal distribution, this choice of k seems reasonable.

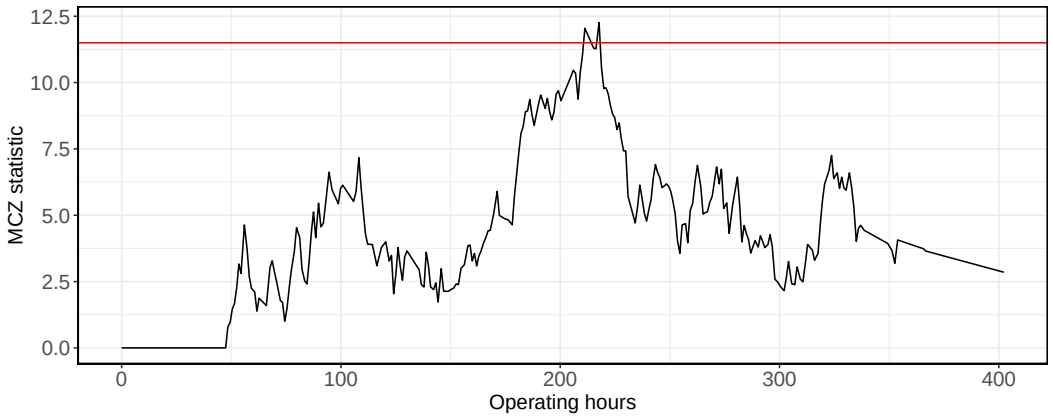
As described in Section 4.3.2, the corresponding value of h can be determined via standard simulations and tables. In our case, a value of $h = 8.008$ should be chosen to obtain a two-sided CUSUM chart with $k = 0.25$ and an average in control run length of 370. As explained in Section 4.4.3 the control limits for the grouped charts cannot be summarized in tables, since they depend on the correlational structure of the Z-statistics. This causes a problem, since the correlational structure can only be computed after the period is finished. However, to achieve online monitoring the control limits need to be available from the start. Fortunately, the correlational structure does not change much over time. We performed the simulations as described in the pseudo-code in Section 4.4.3 and it turned out that the values of h returned by the simulation were equal up to three digits for the five different periods we defined earlier. Therefore, these values of h can be used for all periods and also in the future. However, it might be wise to check if this assumption still holds regularly, for example once a year. For now, $h = 11.6$ is chosen for the MCZ chart and $h = 228$ is chosen for the ZNO chart.

Finally, we want to discuss whether a one-sided or two-sided monitoring approach should be used. For the predictive, recursive and hybrid method we argued that an upper control limit is sufficient to monitor the process. This is because we want to detect an upward shift in the bearing temperatures. The RAV method detects if a bearing starts deviating from the behaviour that is expected given the other bearings. Of course, we want to detect when one bearing has a relatively high temperature, since this is the pattern that is expected in case of wear and cavitation. However, if the temperature of one bearing is relatively low compared to the other bearings, this might also be interesting to know. Therefore, we decide to use a two-sided monitoring approach for the regression adjusted variables method.

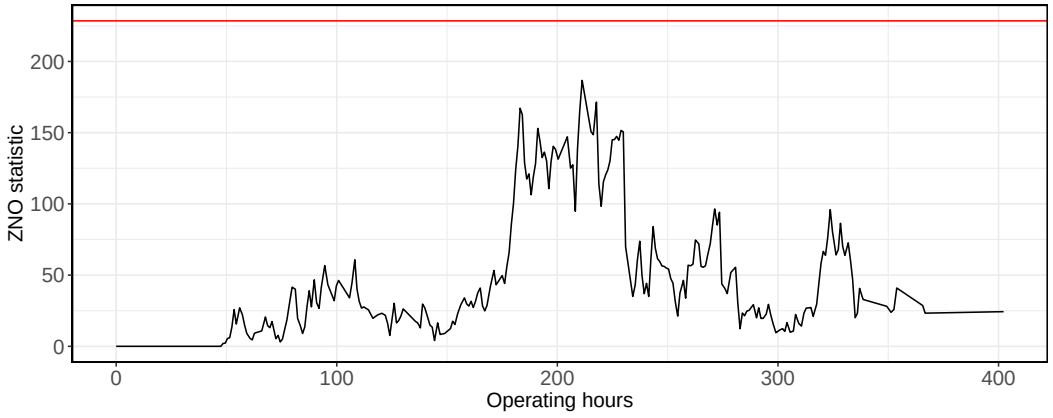
Application to historical data

The results for the second period will be discussed first, because the methods that are discussed earlier fail to detect the failure at the end of this period. Figure 5.19 shows the grouped RAV charts during the second period. In these charts, the black lines represent the grouped statistic (MCZ or ZNO) and the red lines represent the lower and upper control limits. For these charts the recursive residuals obtained by the transit model are used as input. This is because it takes many observations before the combined model can be used during the second period. Figure 5.19a shows that the MCZ chart signals around 215 operating hours. What does this mean?

To gain more information about the cause of a signal one should look at the individual charts during the second period. The individual charts of all seven main bearings can be found in Figure A.7 in the appendix. In the individual charts it can be seen that the fourth bearing exceeds the upper control limit for a period of approximately 50 operating hours. This means that the behaviour of the fourth bearing deviates from the expected behaviour, given the other bearings. The CUSUM statistics for the other bearings stay within the control limits (except for one signal in the first chart). From this is can be concluded that the fourth bearing causes the MCZ chart to signal. The fourth bearing is also the bearing that fails at the end of the second period. The signals in the individual chart of the fourth bearing might be an indication that something is wrong with this bearing. The engineers should take action upon observing such a signal.



(a) CUSUM chart of MCZ statistic during period 2.

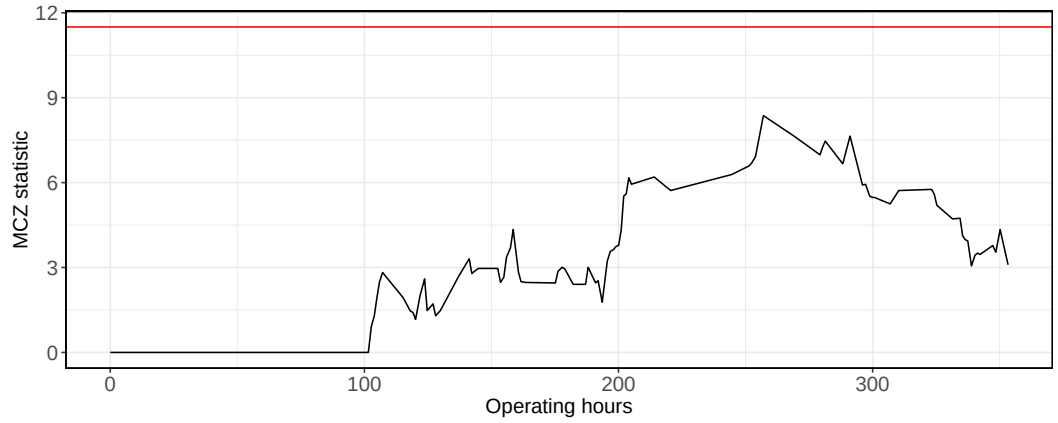


(b) CUSUM chart of the ZNO statistic during period 2.

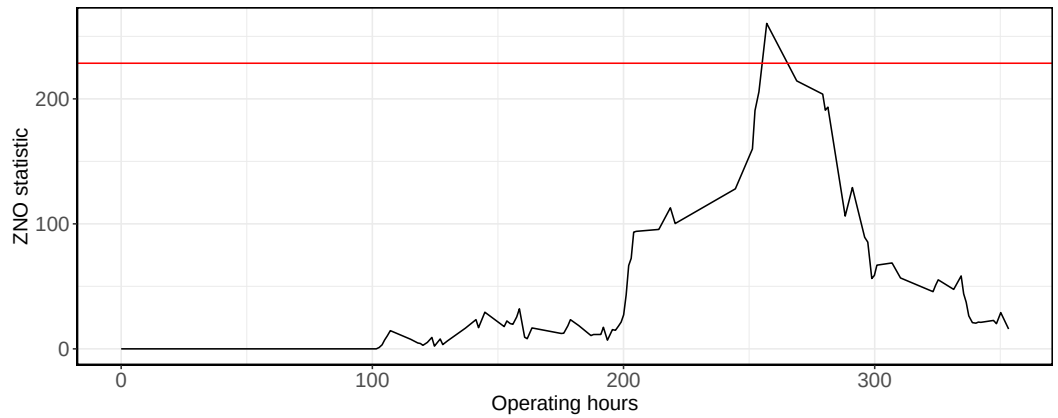
Figure 5.19: CUSUM charts of the grouped RAV statistics during period 2.

The RAV method is able to detect deviant behaviour for the fourth bearing before it fails at the end of the second period. This is an advantage of the RAV method compared to the hybrid residuals method. While the hybrid method is able to detect an upward trend in the temperature of one of the bearings, the RAV method is able to detect deviant behaviour for one of the bearings. One could say that the focus of the RAV method lies more on the cause of the problem and the hybrid method focuses on the consequence. The deviant behaviour that the fourth bearing shows around 200 operating hours could be what causes the bearing to fail eventually.

Figure 5.20 shows the grouped MCZ and ZNO charts during the fifth period. The ZNO chart signals around 260 operating hours. To understand this signal, one must look at the individual charts during this period. These are displayed in Figure A.8 in the appendix. As the figure shows, bearings 3 until 7 approach the upper or lower control limit around 260 operating hours. Bearing 3 however, is the only bearing that crosses one of the control limits. The combination of many bearings being very close to the control limits cause the ZNO chart to signal. Bearing 3 is the bearing that fails at the end of the fifth period. However, it could easily be a coincidence that the chart of this specific bearing crosses the lower control limit very briefly. In this case, from these charts, it cannot be concluded or predicted that bearing 3 is the bearing that will fail before the other bearings do. It should be mentioned that the charts in Figure 5.20 and Figure A.8 are based on the transit model. When the combined model is used none of the grouped charts signal and some of the individual charts give a few false alarms.



(a) CUSUM chart of MCZ statistic during period 5.



(b) CUSUM chart of the ZNO statistic during period 5.

Figure 5.20: CUSUM charts of the grouped RAV statistics during period 5.

Until now, the periods that end with a bearing failure are discussed. It should also be discussed how the RAV method performs in a period between two maintenance actions. Since period 4 contains measurement errors, we will use period 3 to validate the performance of the method in a "stable" period. None of the grouped as well as the individual charts signal when the recursive residuals based on the manoeuvre model are used as input for the RAV method. Note that the manoeuvre model is used here instead of the transit model, since during period 3 the ship sailed in the manoeuvre mode the majority of the time. So, in the manoeuvre mode, the RAV method performs as desired.

Unfortunately, both the MCZ and the ZNO charts signal when the recursive residuals based on the combined model are used as input for the RAV method. To get an idea of the influence that the choice of the model has on the results, the MCZ and ZNO charts for both the combined and manoeuvre models are displayed in Figure A.9 in the appendix. The first two charts of this figure should be compared to the last two charts. The

MCZ chart based on the combined model shows several signals towards the end of the third period. One might think this is an indication of an upcoming failure in the bearings. However, no failures were reported during the maintenance that followed this period. It should be noticed that for all periods, the charts based on the transit or manoeuvre model outperform the charts based on the combined model. This could be a coincidence, but it is definitely something the engineers should keep in mind.

Finally, we want to address one last aspect of this method. Just as Hawkins did in the original article, we used CUSUM charts to monitor the regression adjusted variables. However, CUSUM charts are a little hard to interpret. While the EWMA statistic is a weighted average of the Z-statistics, such an interpretation cannot be given to the CUSUM statistic. We also implemented the EWMA chart in combination with the regression adjusted variables, but this implementation did not lead to useful charts. Therefore, we would recommend using the CUSUM charts, even though they are a little harder to interpret.

Summary and conclusions

In this section, the RAV method was applied to the historical data. During the second period, deviant behaviour was detected for the fourth bearing, which is the bearing that fails at the end of the period. This deviation was large enough to cause signals in grouped MCZ chart. The individual charts of the other bearings did not signal during this period. This means that the RAV method was able to give the root cause of the signal in the grouped chart. During the fifth period, the ZNO chart signalled. This signal originated from multiple bearings being very close to the lower or upper control limit. Deviations were detected, but the root cause of the signal in the ZNO chart was not as clear as during period 2. Both the grouped and individual charts did not signal during the third period when the manoeuvre model was used to obtain the recursive residuals. This is great, since no failures were reported at the end of this period. However, using the combined model, multiple false alarms are generated in the grouped and individual charts.

So, the RAV method is able to detect deviant behaviour for the fourth bearing during the second period. The methods we discussed earlier could not detect any abnormal behaviour during this period. The RAV method also detected some deviant behaviour during the fifth period. However, during this period the RAV signals were less convincing than the signals generated by the hybrid method. Therefore, we conclude that the RAV method is a useful addition to our collection of SPC methodologies, but it cannot replace the hybrid method entirely.

5.3.2 Recursive coefficients method

In the recursive coefficients method, the coefficients of the regression model are re-estimated in every iteration. The coefficient estimates of two successive iterations are subtracted. The result of this subtraction is what we call the recursive coefficient. In case there is stable behaviour, it is expected that the recursive coefficients are approximately equal to zero. When the process starts to show out-of-control behaviour two things can happen: The residuals of the regression model can get larger or the regression coefficients will become different than before, which causes the recursive coefficients to be unequal to zero.

We were unable to successfully implement this recursive coefficient method. In Appendix E the R function that we used to compute the recursive coefficients can be found. This function runs successfully and returns a data frame that contains the recursive coefficients for each of the bearings. However, when we want to monitor the recursive coefficients via T^2 charts, we run in to problems.

In Section 4.1.4, it is shown that the covariance of the recursive coefficients is singular. We cannot explain why this would be the case. We find it hard to comprehend that the covariance of a rather natural expression is singular by definition. However, when we try to compute the T^2 statistic an error messages tells us that the covariance matrix is indeed singular. While we found a theoretical argument for the fact that the T^2 statistic cannot be computed for the recursive coefficients, we should be able to compute it for the recursive residuals. Unfortunately, we ran into the same problems and in this situation a theoretical argument cannot be given. We suspect that the computational singularity has something to do with the high correlations between the variables in the regression model. While we removed very highly correlated variables as described in Section 5.1.1, there

are still variables with high correlations included in the final regression model. The correlation matrix of the variables that are included in the final model can be found in Table 5.7. In this table we find variables that have a pairwise correlation of 0.87 and 0.90. These variables might cause the covariance matrix of the recursive residuals to be computationally singular.

Table 5.7: Correlations between the 10 variables that are included in the final regression model.

Variable	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1	1.00	-0.06	-0.06	-0.11	0.50	0.55	0.87	0.69	0.31	0.87
V2	-0.06	1.00	0.90	0.03	-0.10	-0.11	-0.08	-0.18	-0.18	0.11
V3	-0.06	0.90	1.00	0.03	-0.11	-0.12	-0.07	-0.19	-0.19	-0.01
V4	-0.11	0.03	0.03	1.00	-0.17	0.01	-0.25	0.04	-0.15	-0.11
V5	0.50	-0.10	-0.11	-0.17	1.00	0.22	0.55	0.23	-0.00	0.48
V6	0.55	-0.11	-0.12	0.01	0.22	1.00	0.42	0.52	0.25	0.48
V7	0.87	-0.08	-0.07	-0.25	0.55	0.42	1.00	0.56	0.24	0.72
V8	0.69	-0.18	-0.19	0.04	0.23	0.52	0.56	1.00	0.44	0.49
V9	0.31	-0.18	-0.19	-0.15	-0.00	0.25	0.24	0.44	1.00	0.22
V10	0.87	0.11	-0.01	-0.11	0.48	0.48	0.72	0.49	0.22	1.00

So, the T^2 chart cannot be used to monitor the recursive coefficients. In Section 4.4.2 we also discussed the MEWMA chart, which is a multivariate extension of the EWMA chart. However, the control limits of this chart depend on the correlational structure of the recursive coefficients. The same holds for the MCZ and ZNO charts that are used to monitor the regression adjusted variables. The control limits were simulated in all five periods and it appeared that the correlational structure did not change very much. That is why we were able to fix the control limits so that these charts can be used in practice. However, we did not succeed in performing such an investigation for the recursive coefficients within the time scope of this research. Therefore, for now, we would not recommend using the MEWMA chart on the recursive coefficients.

Monitoring the recursive coefficients using the multivariate T^2 and MEWMA charts is not going to work. One could try to monitor the recursive coefficient vectors using univariate statistical process control methods. We would not recommend this for two main reasons. Firstly, the correlational structure between the coefficient estimates is ignored. Using univariate SPC methods, it is only monitored if the individual components of the coefficient estimates are in-control. If two or more components change together, for example, from a positive to a negative correlation, this might not be picked up by the univariate chart. Secondly, the use of univariate SPC methods will lead to an overload of control charts. The regression model contains 10 variables and each of the seven bearings is monitored individually. This leads to 70 control charts which is way too much for the engineers to keep track of.

Summary and conclusions

The recursive coefficients method could not be implemented successfully because we did not find a multivariate control chart to properly monitor these coefficient vectors. The T^2 chart could not be used because of the singularity of the covariance matrix. The MEWMA chart could not be used because we do not know if the correlational structure of the recursive coefficients changes over time. The control limits of this chart need to be simulated and if the correlational structure changes, the control limits change too. This makes this chart very impractical. Monitoring the recursive coefficients via univariate charts is not an option, since this leads to an overload of charts and the correlational structure between the variables is ignored in this way.

Chapter 6

Conclusions

Based on the results from the previous chapters, we draw several conclusions. In this chapter, these conclusions will be summarized and the research question will be answered. In the first chapter, five sub-questions were defined. Looking back, the most attention was paid to answering the second and third sub-questions. These questions led to the most important conclusions. Therefore, the focus will lie on answering these questions in the first and second section of this chapter.

The main question that is answered in this thesis is: How can statistical process control be applied to predict failures in the main bearings of the diesel engines from the OPVs? The answer is that a combination of the hybrid residuals method and the regression adjusted variables method can be used to predict failures in the main bearings. By recursively estimating the residuals of the multiple linear regression model, we were able to shorten the initialization period significantly for both methods. As a consequence, the methods can be restarted after maintenance actions are performed on the diesel engines. In this way real-time monitoring of the bearing temperatures can be realized.

The methods that we developed in this thesis build on the previous work of Heek (Heek (2021)). We were able to shorten the initialization period of the methods from 800 operation hours to only 20 operating hours, which makes that the methods can really be used in practice at the Royal Netherlands Navy. On top of that, the RAV method that we implemented can detect deviant behaviour prior to the failure at the end of the second period. Using the methods Heek developed, this failure could not be detected. By implementing and combining multiple methods, more insights can be given to the engineers on the failures. In the sections below, more details on this conclusion are given by answering the sub-questions.

6.1 The MLR model

The first sub-question that was defined in Chapter 1 is: Which (output) variables should be monitored to timely detect failures in the main bearings of the diesel engines? This question was actually already answered in Heek (2021). He concluded that abrasive wear and cavitation are the most important failure modes for the diesel engines of the OPVs. Both wear and cavitation lead to increased bearing temperatures. That is why the bearing temperatures contain information on the quality of the bearings. The temperature of each main bearing is measured individually by a temperature sensor below the bearing. Therefore, the seven bearing temperatures are the output variables that should be monitored in order to timely detect failures in the main bearings of the diesel engine.

The second sub-question is: How can the relations between output variables and input variables of the main bearings of the diesel engines be modelled? In his research, Heek used a Multiple Linear Regression (MLR) model to describe the relations between the input and output variables of the diesel engines. An MLR model seemed the most appropriate, since it is an easy to understand white box model. Using a white box model, the engineers can interpret the residuals themselves. Since Heek's research showed very promising results, we decided to keep the MLR model. However, we reconsidered the variable selection process.

To reduce the number of highly correlated variables in the data, clusters of pairwise highly correlated variables were determined. For each cluster one variable was selected to represent the cluster in the data set. Next, LASSO variable selection was performed on multiple sub-samples to overcome the large sample problem.

The variable selection process was performed separately for the transit and manoeuvre mode, but it turned out that the selected variables were the same for both propulsion modes. Therefore, the models for both modes could be combined. Adding the propulsion mode and all interactions with the propulsion modes and the other predictors proofed to be a significant improvement of the model. The regression model was validated in a stable period using 10-fold cross validation. The RMSE and MAE were very small (< 0.06 for all bearings) and the adjusted R^2 was very large (> 0.994 for all bearings). From this we concluded that the model is able to predict the bearing temperatures in a stable period accurately. Regarding the signal-to-noise ratio, the model outperformed the model as proposed in Heek (2021). Therefore, we conclude that the more sophisticated variable selection process we performed really leads to an improvement of the model. This answers the second sub-question.

6.2 Statistical Process Control

The third sub-question we defined in Chapter 1 is: What is an appropriate (multivariate) control chart to monitor the main bearings of the diesel engines making use of the relation between output and input variables? In the previous chapters, we discussed multiple statistical process control methods. Methods based on predictive residuals did not prove to be appropriate, since these residuals become biased after maintenance actions. This means that after maintenance actions a new initialization period of several hundreds of hours is needed to refit and validate a new regression model. During this period the quality of the bearings cannot be monitored. This makes methods based on solely predictive residuals useless.

As an alternative to the predictive residuals, the recursive residuals are proposed. In the recursive method, the parameters of the regression model are re-estimated in every iteration. Then the model is used to predict the upcoming observation. An advantage of the recursive method is that the initialization period is very short (20 operating hours). A disadvantage is that the model learns a weak but persistent trend. The historical data shows exactly such a trend before the failure at the end of period 5. The recursive method is not able to pick up this trend. Therefore, we conclude that methods based on solely the recursive residuals are not able to predict bearing failures in advance.

The shortcoming of the recursive method appeared to be solvable by switching to the predictive method. This leads to a hybrid method in which the strengths of the recursive and predictive methods are combined. In the hybrid method, the model is reset after maintenance actions. When 20 observations are collected, the recursive method is used to "start up" the model. When the model is "stable enough" a switch is made to the predictive method. The model that is estimated in the last iteration of the recursive method, is used as the input model for the predictive method. In this way we can take advantage of the short initialization period of the recursive method. The adaptive property is however no longer present, since the control limits of the predictive method are constant.

The hybrid method was implemented using Shewhart and EWMA charts. Since the Shewhart charts use the individual observations as charting statistic, the resulting chart has a low signal-to-noise ratio. Therefore, you cannot see a clear upward trend being developed. In the EWMA charts you do see a clear upward trend develop during the fifth period. The EWMA chart based on the hybrid method starts signalling 190 operating hours before the failure occurs. This corresponds to 47 calendar days. This gives the engineers enough time to take action upon these signals. The failure at the end of the second period is not picked up by the EWMA chart based on the hybrid residuals. This is because the behaviour before this failure is not what is expected (a weak but persistent upward trend).

A multivariate statistical process control method we discussed is the Regression Adjusted Variables (RAV) method. This method is designed to detect when the temperatures of one of the bearings start to deviate from what is expected, given the temperatures of the other bearings. The raw temperature data does not meet the requirements for this method. As a consequence, the resulting charts of the RAV method applied to the raw data are not usable. Alternatively, the RAV method is applied to the recursive residuals, since these residuals do meet the assumptions. The RAV method results in grouped CUSUM charts for all bearings and individual CUSUM charts per bearing.

When applied to the second period, the grouped MCZ chart signals around 220 operating hours. Looking at the individual charts, it can be seen that the chart of the fourth bearing shows deviant behaviour for a period of approximately 50 operating hours. The fourth bearing is also the bearing that fails at the end of this period.

The engineers should take action upon these signals. So, the RAV method is not only able to detect a failure, it can also give the root cause of this failure. Such results were not achieved by the methods implemented by Heek. Since the signals of the MCZ charts take place approximately 200 operating hours before the failure, which corresponds to 12 calendar days, there is plenty time to research this deviant behaviour and maintain the fourth bearing.

Another method that was investigated in this research is the method of recursive coefficients. Since the recursive coefficients are multivariate quantities, they should be monitored using multivariate SPC approaches. However, the T^2 chart cannot be applied because of the singularity of the covariance matrix of the recursive coefficients. The MEWMA chart is impractical to use for online monitoring, since it depends on the correlational structure of the recursive coefficients. Univariate monitoring is not an option, since this leads to an overload of control charts. Therefore, we conclude that the recursive coefficient method cannot be used to predict failures in the main bearings of the diesel engines at the Royal Netherlands Navy.

To answer the third sub-question, EWMA charts based on the hybrid residuals and grouped and individual CUSUM charts based on the regression adjusted variables should be used together to monitor the quality of the main bearings. Signals for multiple kinds of failures can be picked up in advance using a combination of these methods.

The fourth sub-question was: Can a root cause analysis be developed to determine from the control chart the cause of the change when a change in the process is detected? Using a combination of the grouped and individual charts in the RAV method, it can be seen which bearing (or potentially which combination of bearings) is causing the signal in the grouped chart. The RAV method tells us exactly which bearing is showing deviant behaviour. The hybrid method does not have this property. All charts start signalling around the same time, so from these charts it cannot be determined which bearing eventually is going to fail. PCA based methods might give more insight in the cause of the signals, unfortunately this could not be realized within the time scope of this research.

The final sub-question was: How can the Remaining Useful Life (RUL) be predicted when a change in the process is detected? Because of the limited amount of failure data, we concluded in a very early stage of the research that performing a RUL analysis is infeasible for this research. As mentioned before, the hybrid chart started signalling 47 calendar days before the failure at the end of period 5. The RAV charts started signalling 12 calendar days before the failure at the end of period 2. However, we have no reason to believe that failures in the future will be detected within the same time range as these two failures were.

We finish this chapter with a summary of the key insights that were gained during this research. They can be found in the box below.

Key insights

- An improvement on the work of Heek is that the hybrid and RAV method can actually be implemented to achieve online monitoring of the bearing temperatures. After an initialization period of 20 operating hours the monitoring can start. Heek uses an initialization period of 800 operating hours. Since the initialization period of the new methods is very short, the model can be reset after maintenance. This makes that the methods that are presented in this thesis can be used in practice at the Royal Netherlands Navy.
- It can take a while before enough observations are collected in both propulsion modes so that the combined regression model (including the propulsion mode and interactions) can be applied. Therefore, the charts based on the residuals of the transit or manoeuvre model should also be available for the engineers. It turned out that these charts might pick up signals of failures even faster than charts based on the combined model.
- The RAV method not only gives signals in case of deviant behaviour. The method can also be used to tell which bearing is causing this deviant behaviour. This is an important advantage for the RAV method compared to the hybrid method. However, historical data showed that the hybrid method is able to pick up signals that are not picked up by the RAV method. Therefore, the control charts resulting from both methods should be monitored.

Chapter 7

Discussion

In this chapter the conclusions from the previous chapter will be discussed. First, we will name some limitations of the research we did and we give suggestions for future research in this area. Next, we will give recommendations regarding the implementation of statistical process control methods at the Royal Netherlands Navy.

7.1 Limitations and future research

There are several limitations regarding this research. One aspect that made the validation of the methods we developed particularly hard is the limited amount of failure data. The data set that was available only contained two failures. On top of that, the behaviour right before these failures was very different for these two cases. We found a collection of methods that is able to detect both failures in advance. However, we cannot know for sure that all failures in the future will also be picked up by these methods. Also, the choice of $\lambda = 0.1$ for the EWMA charts cannot be validated, because of the limited amount of failure data. The optimal value for the historical data set is chosen, but we have no guarantees that this value also will be optimal in the future. Since it is hard to validate the model properly, the conclusions we drew in the previous chapter should be interpreted with care.

The limited amount of failure data also makes it hard to perform a Remaining Useful Life (RUL) analysis. However, being able to predict the RUL of the bearings would be really useful for the Royal Netherlands Navy. The statistical process control methods that we implemented show when the bearing temperatures are behaving out-of-control. However, it is not clear when the engineers should take action and start maintenance. A RUL prediction could be helpful to set up a maintenance policy. Literature on the RUL prediction for bearings is available, but for example in Tayade et al. (2019) the data set contains 17 complete runs to failure. It is infeasible to obtain such a data set at the Royal Netherlands Navy. But by collecting failure data from the other OPVs and maybe even from vessels in other classes, the number of runs to failure can be increased. If more runs to failure are collected, the RUL prediction techniques described earlier in this report can be applied. In Si et al. (2011) it is stated that it is desirable to develop a new RUL estimation technique in case limited or no failure data is available. Statistical methods cannot be used in these two cases and physical models cannot be validated properly. Therefore, this is also a really interesting topic for future research.

In this research Shewhart and EWMA charts were used to monitor the hybrid residuals. The EWMA chart is preferred because it shows a clear upward trend, while the signal-to-noise ratio of the Shewhart chart is much lower. However, for the RAV method the CUSUM chart turned out to be the most appropriate. Since the CUSUM chart is a great improvement compared to the EWMA chart for the RAV method, this could also be the case for the hybrid method. Due to the limited time, we were unable to explore the possibility of monitoring the hybrid residuals via CUSUM charts. This possibility could be investigated by the Royal Netherlands Navy in the future, since it might improve the hybrid method. Note that CUSUM charts should only be used if there is a great improvement. A trade-off should be made between the quality and the interpretability of the chart.

Another limitation of the research we did, is that we were unable to successfully implement the recursive coefficients method. Initially, we wanted to include the recursive regression coefficients in our analysis. However, due to the singularity of the covariance matrix of the recursive coefficients, the T^2 chart could not be used to

monitor the recursive coefficient vectors. The MEWMA chart might be an alternative. Before the MEWMA chart can be applied, it must be investigated if the correlational structure of the recursive coefficients changes over time. This is because the control limits of the MEWMA chart depend on this correlational structure. The correlational structure must be available at the beginning of a monitoring period to achieve online monitoring. This is only true if the correlational structure does not change much over time. Then the correlations from a previous period can be used to determine the control limits of the MEWMA chart in advance. Within the time scope of this research, we were unable to investigate this ourselves. Therefore, this is left open for future research.

There are also some mathematically interesting opportunities for future research. For example, it is proven that the recursive residuals are independent assuming that the mean and standard deviation are known. We suspect that this is still true if the mean and standard deviation are estimated. However, we were not able to give a formal proof. The same holds for the recursive coefficients. The recursive coefficients are independent over time, given that the mean and covariance are known. However, this is not proven in the case that the mean and covariance are estimated. These proves are therefore left open for future research.

Another topic for mathematical future research is the use of multivariate multiple linear regression models. In this research a univariate multiple linear regression model is used to obtain the residuals for all bearings individually. However, the system is a multivariate process and the bearing temperatures are highly correlated as we saw in Chapter 2. Therefore, it would be more appropriate to use a multivariate regression model. To be able to properly monitor the residuals that result from this model, the distributions of the predictive and recursive residuals must be derived. In Van Dalen (2020) the distribution of the predictive residuals of the multivariate multiple linear regression model is given. These residuals are dependent, just as in the univariate case. However, a distribution of the recursive residuals is not derived. In this research we have seen how important the recursive method is. Without the recursive method, the control charts cannot be used in practice. The distribution of the recursive residuals of the multivariate multiple linear regression model is thus an important topic for future research.

7.2 Recommendations

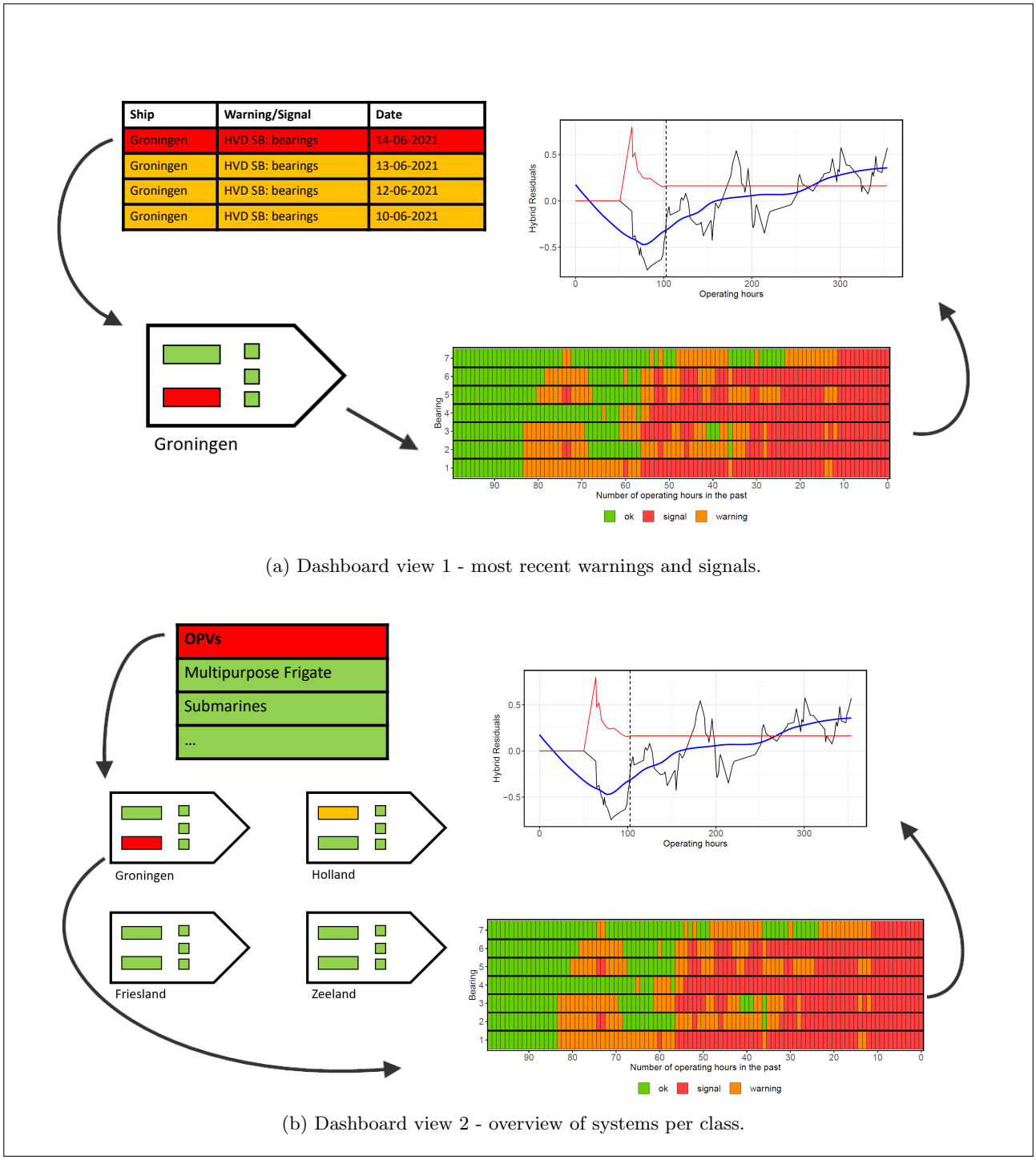
So far, we did not discuss how exactly the Royal Netherlands Navy can implement the methods we designed. Therefore, we give some recommendations regarding the implementations of the methods in this section.

While in this research we only looked at the main bearings of the port side diesel engine of one of the OPVs, the same methods can be applied to the other OPVs and even to the diesel engines of vessels from different classes. Implementing the same methods for the other OPVs would be rather easy, since the engines in these ships are identical. Therefore, the same variables should be used in the regression model. However, the parameters have to be re-estimated for each engine individually. When the methods are applied to the main bearings of the diesel engines of ships from other classes, the variable selection process should be performed again. Different engines might contain different sensors and the dynamics inside the engines might be different too. Therefore, the analysis in Section 5.1 should be executed again.

Apart from applying these methods to other ships, they can also be applied to other components. However, the components should meet several requirements in order to make the monitoring process useful. Firstly, failure data of these components must be available. If there is a lack of failure data, the pattern prior to a failure cannot be studied. It is then very hard to determine which variables should be monitored to be able to assess the quality of the component. On top of that, the method that is developed cannot be validated if there is no failure data available. Secondly, special attention should be paid to the question: Does monitoring the quality of this component really contributes to the maintenance process? For example, if the failure of a certain component does not have very big consequences regarding the availability of the ship, it might not be worth the trouble of constructing models for the quality of these components. One should try to avoid an information overload by only monitoring components that really matter.

The monitoring techniques that are presented in this research can be used to their full potential if a real-time implementation is realized. When the data is sent from the ship directly to the engineers in Den Helder, signals from the control charts can be reacted upon immediately. If the data is, for example, sent once a week, there might not be enough time to prevent bearing failures (during the second period a signal was given 12 calendar days before the failure). Therefore, we would recommend facilitating a real-time implementation of the models.

As written in the conclusions, it is advised to implement both the hybrid method and the RAV method, since these methods are able to detect different patterns leading to failures. It is also advised to monitor the bearings using the combined and the transit or manoeuvre model (depending on which mode is sailed in most). The hybrid method results in seven control charts per model and the RAV method results in nine control charts per model. This leads to a total of 32 control charts. Note that these charts are only used to monitor the main bearings of the port side engine. When in a later stadium more components and more ships are being monitored it becomes impossible to keep track of all these charts. To overcome this issue, an easy to understand and structured dashboard could be developed. To give an idea of what such a dashboard could look like, some sketches are presented in Figure 7.1.



(a) Dashboard view 1 - most recent warnings and signals.

(b) Dashboard view 2 - overview of systems per class.

Figure 7.1: Sketch of a possible future implementation of the control charts in a dashboard.

Figure 7.1 shows two different suggestions for a possible future dashboard. In Figure 7.1a, the first suggestion is visualized. The arrows represent clicks in the dashboard. The dashboard opens with a table that contains the latest signals and warnings. The table summarizes the warnings in terms of the ship, component and date of the warning. The engineers can click on these warnings and signals to open a top view of the ship. In this top view, colour codes represent the state of the components. If components are coloured orange (warning) or red (signal) this indicates that there might be a problem with these components. The engineer can click on such a component to open a warnings table that shows the state of the component over time. The figure shows an example of the warning table for the main bearings of the diesel engine. In this example one model is used (hybrid method in the transit mode), however in real life this table should be extended with the RAV method and the combined model. If the engineer is triggered by one of the bearings, he or she can click on this bearing to open the actual control chart.

Figure 7.1b shows a different set-up for the dashboard. Here, the dashboard opens with a table that contains an overview of all classes of ships. The state of the ships in each class is represented by a colour. When an engineer is interested in the ships of one specific class, he or she can click on an entry of this table to open the top view of all the sips in this class. Again, the components are color-coded to indicate the state of that specific component. Just as in the previous set-up, the engineer can open a table that contains the warnings over time or even some specific control charts.

The warnings that are present in the warnings table in this figure are created using a 2σ warning limit. The chart signals when the statistic crosses the 3σ limit. By combining multiple charts for multiple models in one table, the false alarm rate exceeds the rate of $1/370$ of the individual charts. The engineers should report it if too many false alarms are generated in practice. If this is the case, the criteria for warnings could be changed. For example, one could implement that a warning is generated only if the chart exceeds the 2σ limit for three successive observations.

We would like to make one final recommendation. We think it is important that the engineers that eventually will use these methods read parts of this report. Seeing the examples on the historical data will help them to learn how the signals should be interpreted.

Bibliography

A Alin. Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):370–374, 2010.

S. Bersimis, J. Panaretos, and S. Psarakis. Multivariate statistical process control charts and the problem of interpretation: A short overview and some applications in industry. 02 2005.

N.H. Bingham and J.M Fry. *Regression - Linear Models in Statistics*. Springer, 2010.

R.L. Brown, J. Durbin, and J.M. Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 149–192, 1975.

G. Capizzi. Recent advances in process monitoring: Nonparametric and variable-selection methods for phase i and phase ii. *Quality Engineering*, 27:44–67, 2015.

G. Capizzi and G. Masarotto. Comparison of phase ii control charts based on variable selection methods. *Frontiers in Statistical Quality Control*, 11:151–162, 2015.

J.M. Dufour. Recursive stability analysis of linear regression relationships. *Journal of Econometrics*, 19(1): 31–76, 1982.

M. Frisén. Properties and use of the shewhart method and its followers. *Sequential Analysis*, 26:171–193, 2007.

F.E. Harrell. *Regression Modeling Strategies*. Springer, Nashville, 2015.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2016.

D.M. Hawkins. Multivariate quality control based on regression-adjusted variables. *Technometrics*, 33(1):61–75, 1991.

D.A. Heek. A data-driven condition monitoring approach for the main bearings of a marine diesel engine. Master's thesis, Eindhoven University of Technology, Department of Industrial Engineering & Innovation Sciences, Eindhoven, The Netherlands, 2021.

H. Hotelling. Multivariate quality control. *Techniques of Statistical Analysis (C. Eisenhart, M. Hastay, and W.A. Wallis, eds.)*, pages 111–184, 1947.

G. James, D. Witten, T. Hastie, and R Tibshirani. *Introduction to Statistical Learning*. Springer, 2014.

J.M. Juran. Early sqc: A historical supplement. *Quality Progress*, 1997.

Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin. Machinery health prognostics: A systematic review from data acquisition to rul prediction. *Mechanical Systems and Signal Processing*, 104:799–834, 2018.

M. Lin, H.C. Lucas, and G. Shmueli. Too big to fail: Large samples and the p-value problem. *Information Systems Research*, pages 1–12, 2013.

K. Linderman, K.E. McKone-Sweet, and J.C. Anderson. An integrated systems approach to process controland maintenance. *European Journal of Operational Research*, 164:324–340, 2005.

H. Liu, W. Jiang, A. Tangirala, and S. Shah. An adaptive regression adjusted monitoring and fault isolation scheme. *Journal of Chemometrics*, 20:280–293, 2006.

C.A. Lowry, W.H. Woodall, C.W. Champ, and S.E. Rigdon. A multivariate exponentially weighted moving average control chart. *Technometrics*, 34:46–53, 1992.

R. L. Mason and J.C. Tracy, N. D.and Young. Decomposition of T^2 for multivariate control chart interpretation. *Journal of Quality Technology*, 27:99–1108, 1995.

R.L. Mason and J.C. Young. *Multivariate statistical process control with industrial applications*, volume 9 of *ASA-SIAM Series on Statistics and Applied Probability*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

B.M. McClurg. A self-starting statistical control chart methodology for data exhibiting linear trend. Master's thesis, University of Iowa, Iowa City, The United States of America, 2016.

F.M. Megahed and L.A. Jones-Farmer. Statistical perspectives on "big data". *Frontiers in Statistical Quality Control*, 11:29–47, 2015.

Z. Mehrafrooz and R. Noorossana. An integrated model based on statistical process control and maintenance. *Computers & Industrial Engineering*, 61:1245–1255, 2011.

P.J. Miranti. Corporate learning and quality control at the bell system, 1877-192. *Business History Review*, 79: 39–72, 2005.

E.S. Page. Control charts with warning lines. *Biometrika*, 42(1-2):243–257, 1955.

S. Panagiotidou and G. Tagaras. Statistical process control and condition-basedmaintenance: A meaningful relationship throughdata sharing. *Production and Operations Management*, 19(2):156—-171, 2010.

D.S.G. Pollock. Recursive estimation in econometrics. *Computational Statistics & Data Analysis*, 44:37–75, 2003.

P. Qiu. *Introduction to Statistical Process Control*. Chapman and Hall/CRC, 2013.

S.W. Roberts. Control charts based on geometric moving averages. *Technometrics*, 1:239–250, 1959.

D. Salomé. Estimating remaining useful lifetime bearings using raw vibration measurements. Master's thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, Eindhoven, The Netherlands, 2019.

W. A. Shewhart. *Economic Control of Quality of Manufactured Product*. Macmillan, London, 1931.

A.N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, 8:22–46, 1963.

X. Si, Z. Zhang, and C. Hu. *Data-Driven Remaining Useful Life Prognosis Techniques*. Springer, Beijing, 2017.

X.S. Si, W. Wang, C.H. Hu, and D.H. Zhou. Remaining useful life estimation – a review on the statistical data driven approaches. *European Journal of Operational Research*, 213:1–14, 2011.

A. Tayade, S. Patil, V. Phalle, F. Kazi, and S. Powar. Remaining useful life (RUL) prediction of bearing by using regression model and principal component analysis (PCA) technique. *Vibroengineering PROCEDIA*, 23:30–36, 2019.

M. Teunisse. Anomalie detectie voor voorspelbaar onderhoud (Dutch). Bachelor thesis, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands, 2021.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58: 267–288, 1996.

N.D. Tracy, J.C. Young, and R.L. Mason. Multivariate control charts for individual observations. *Journal of Quality Technology*, 24:88–95, 1992.

O. Van Dalen. Statistical monitoring of wind turbines. Bachelor thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, Eindhoven, The Netherlands, 2018.

O. Van Dalen. Contextual anomaly detection through multivariate monitoring of regression profiles. Master's thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, Eindhoven, The Netherlands, 2020.

A. Vencl and A. Rac. Diesel engine crankshaft journal bearings failures: Case study. *Engineering Failure Analysis*, 44:217–228, 2014.

S. Vidal-Puig and A. Ferrer. A comparative study of different methodologies for fault diagnosis in multivariate quality control. *Communications in Statistics - Simulation and Computation*, 43(5):986–1005, 2013.

S.J. Wierda. Multivariate statistical process control—recent results and directions for future research. *Statistica Neerlandica*, 48:147–168, 1994.

W.H. Woodall and D.C. Montgomery. Research issues and ideas in statistical process control. *Journal of Quality Technology*, 31(4):376–386, 1999.

Appendix A

Additional figures

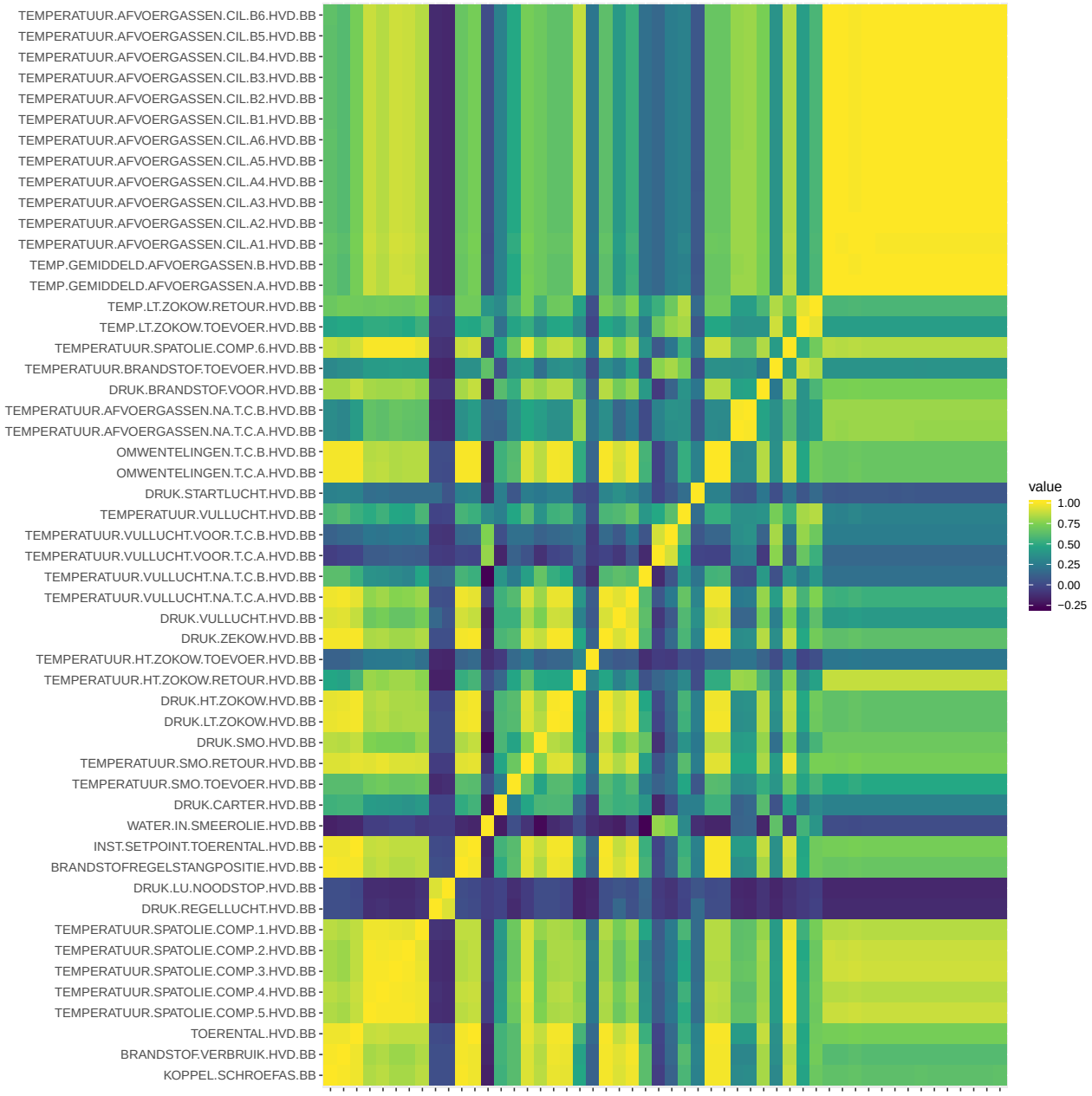


Figure A.1: Correlations between the predictor variables. Adding labels to the x-axis, makes this figure unreadable. Therefore, keep in min that the order of the variables on the x-axis is the same as the order of the variables on the y-axis.

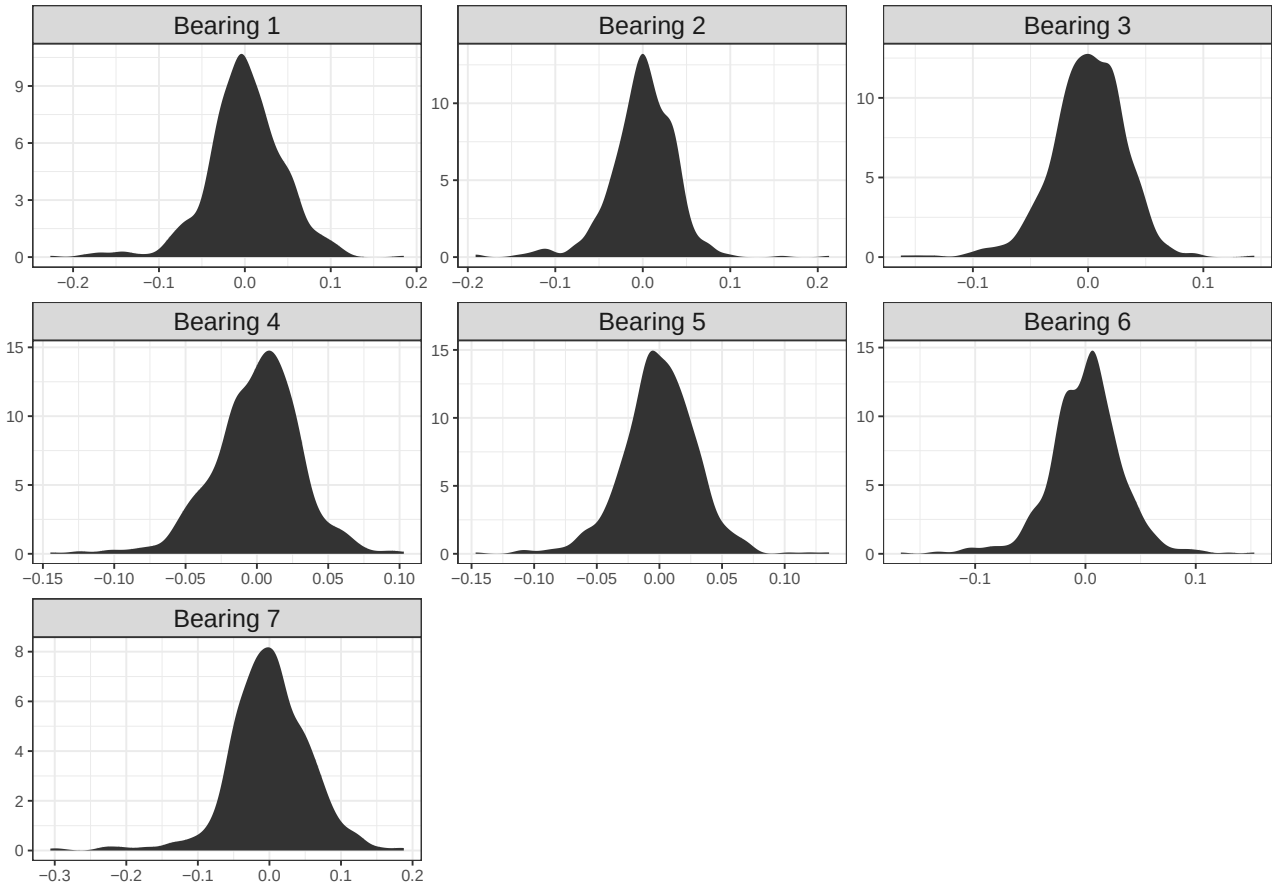


Figure A.2: Residual densities for the stable period per bearing.

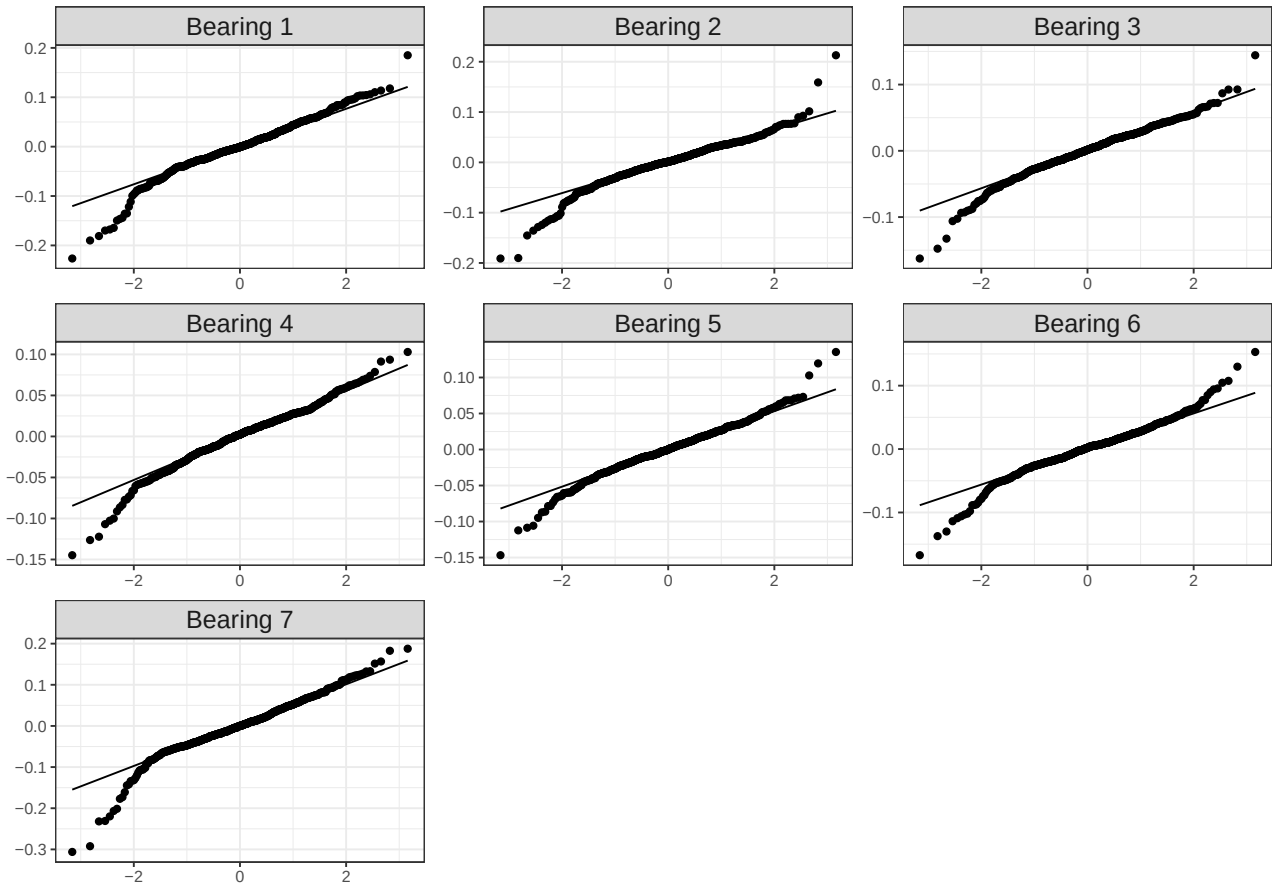


Figure A.3: QQ-plots of the residuals for the stable period per bearing.

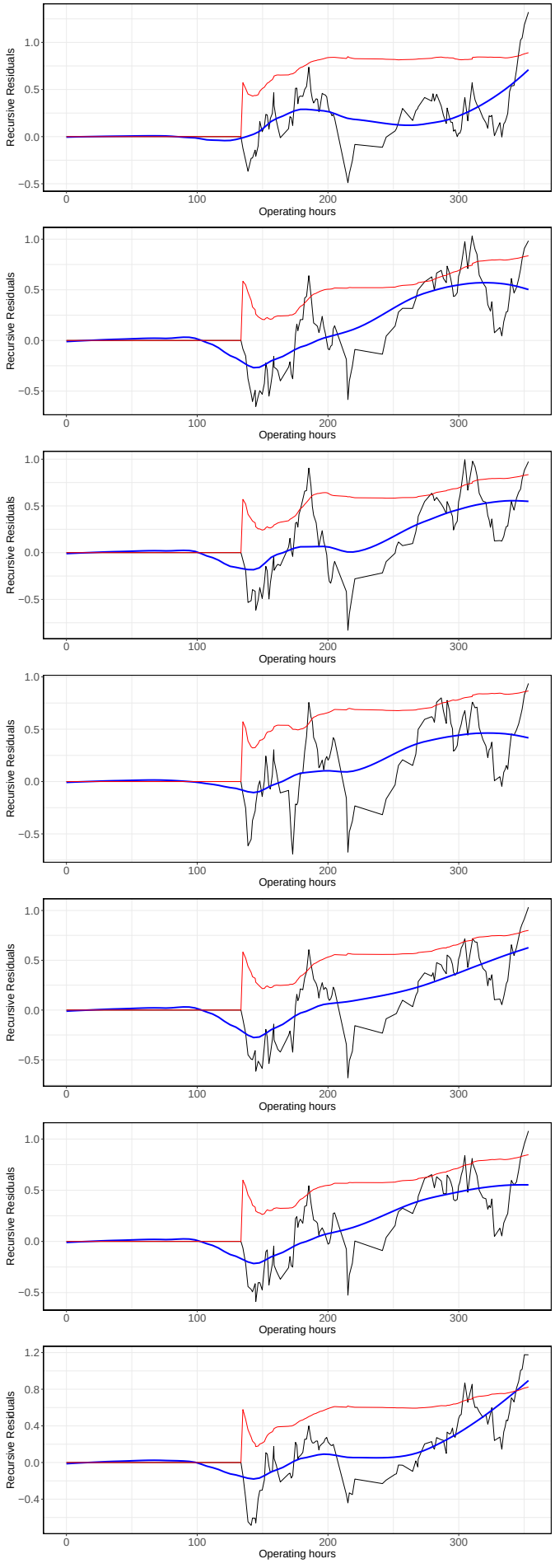


Figure A.4: Recursive EWMA charts during period 5 of the seven main bearings in ascending order.

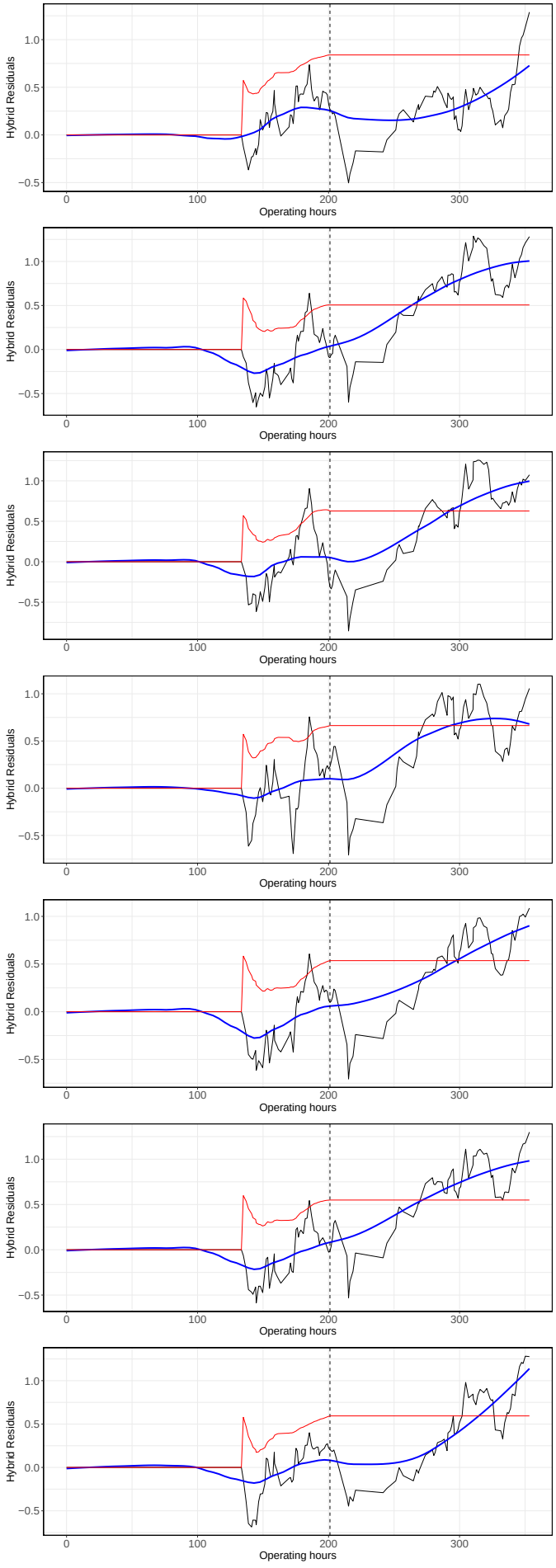


Figure A.5: Hybrid EWMA charts during period 5 of the seven main bearings in ascending order.

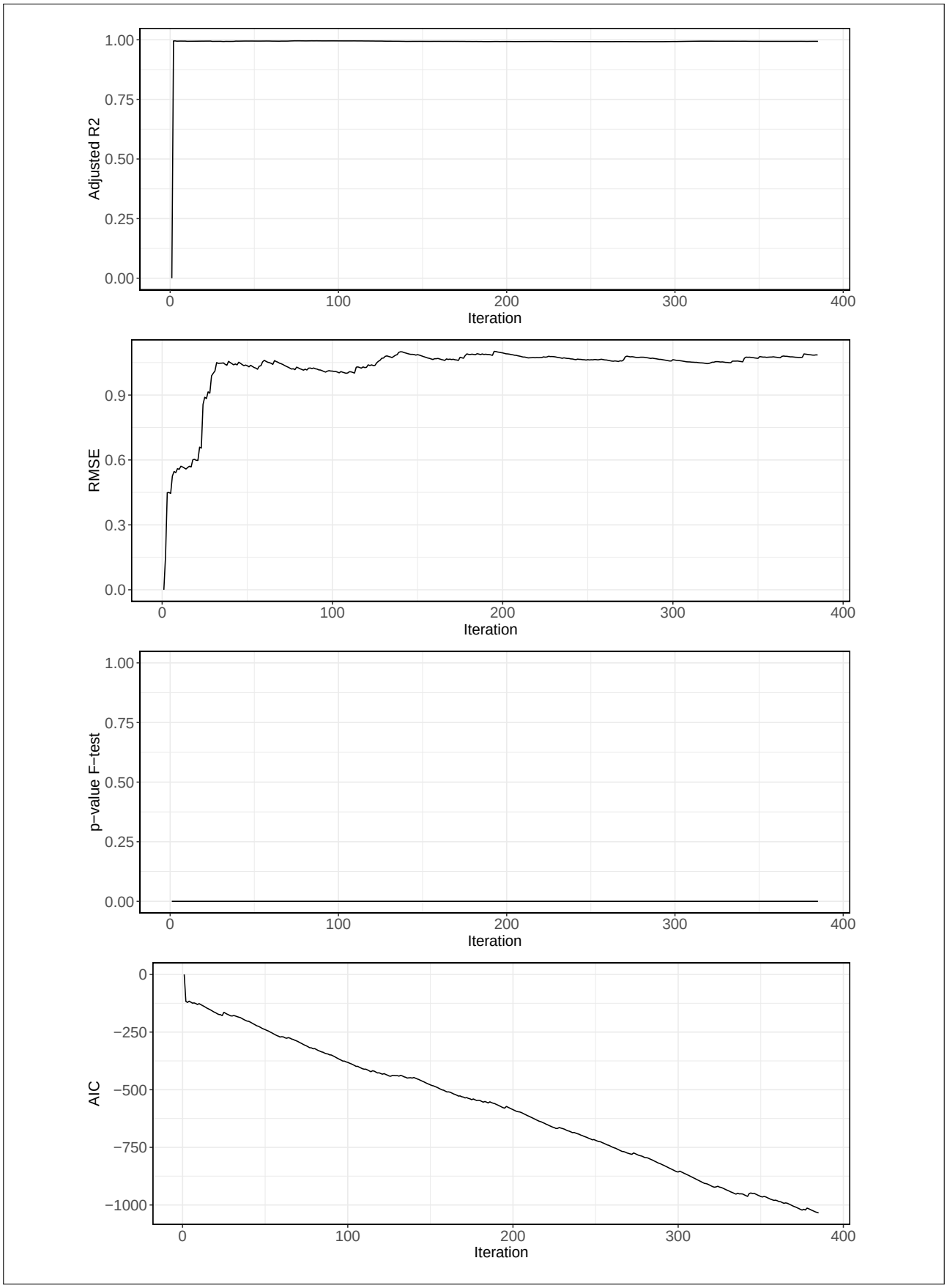


Figure A.6: Characteristics of the recursive regression model for period 3 using the combined model.

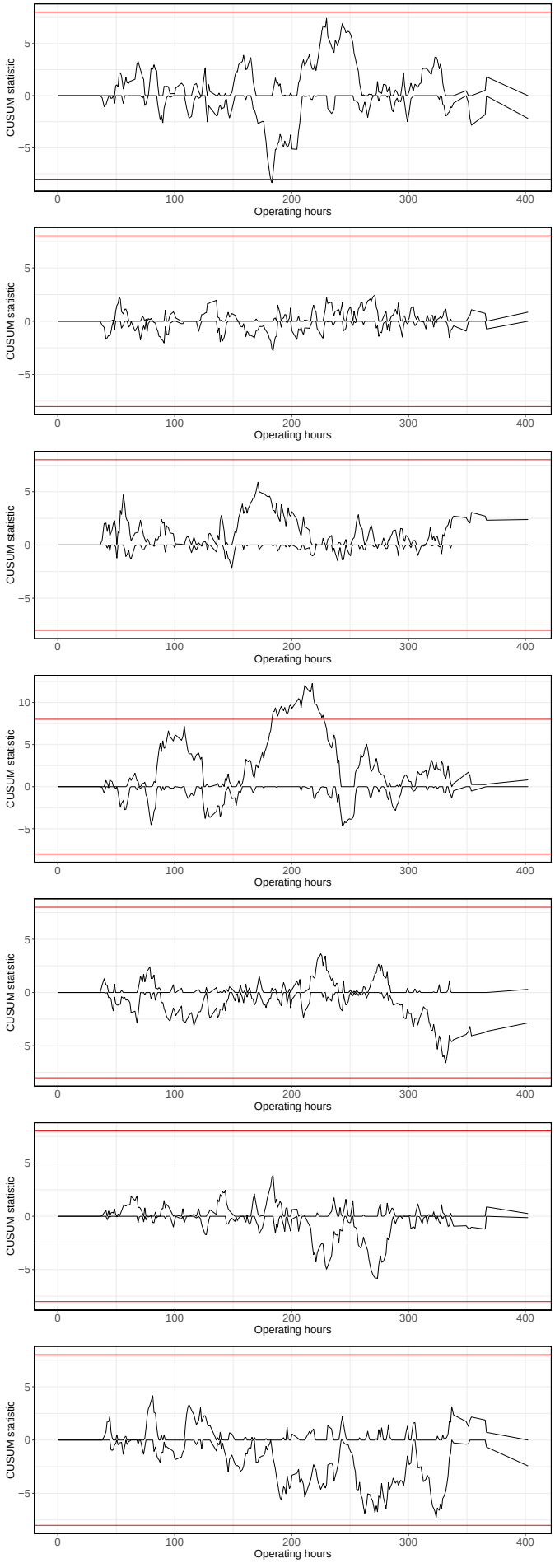


Figure A.7: Individual RAV CUSUM charts during period 2 of the seven main bearings in ascending order.

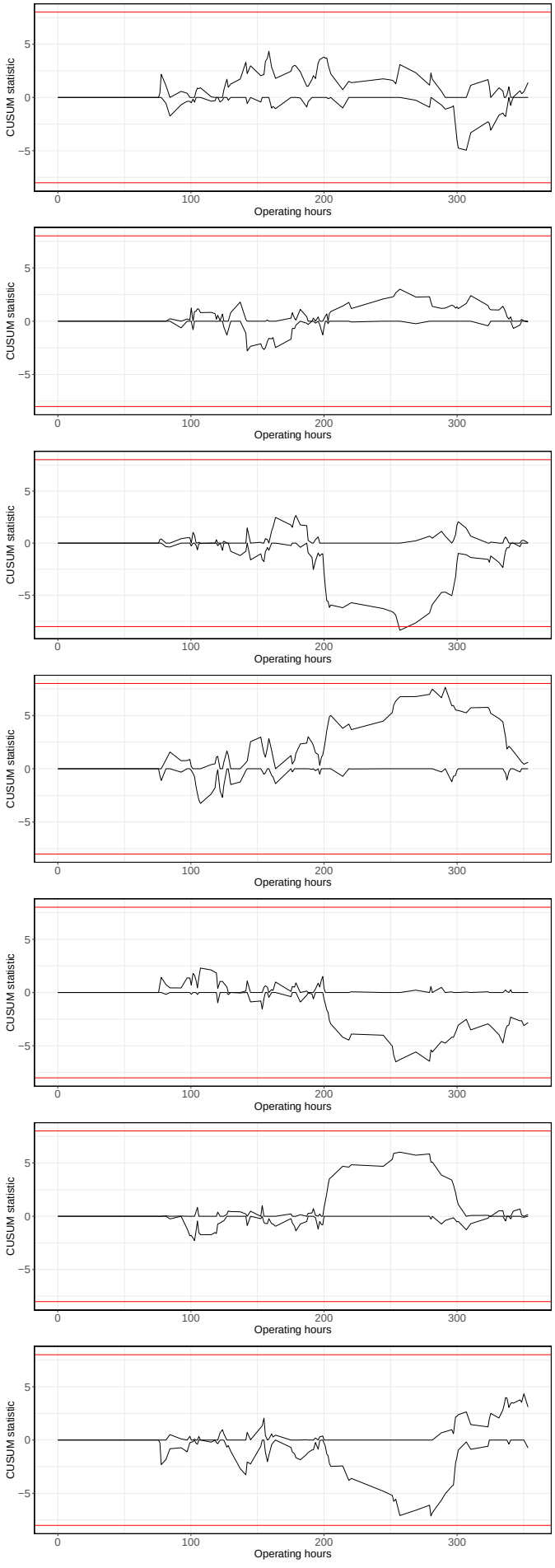
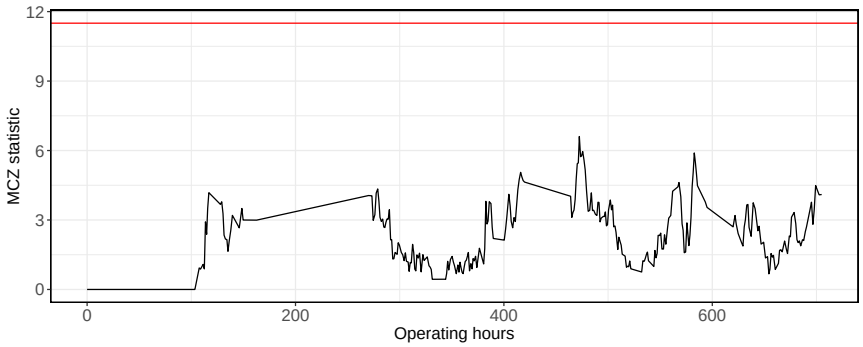
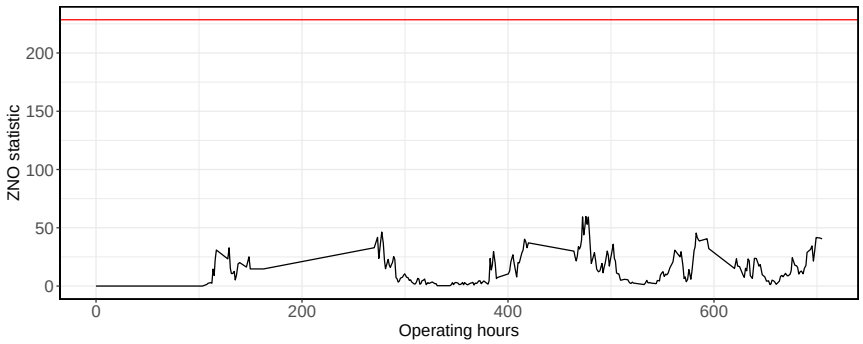


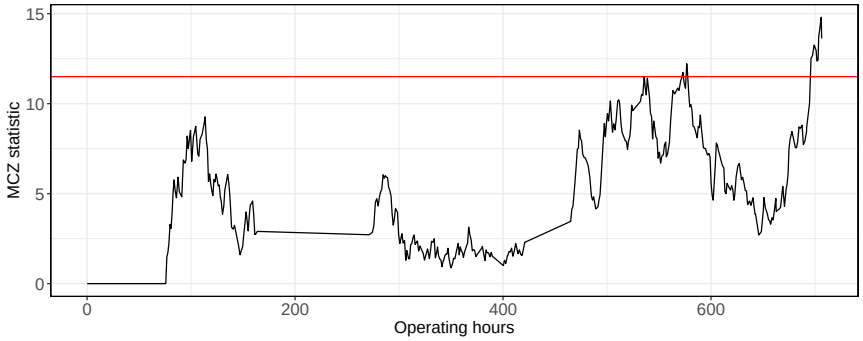
Figure A.8: Individual RAV CUSUM charts during period 5 of the seven main bearings in ascending order.



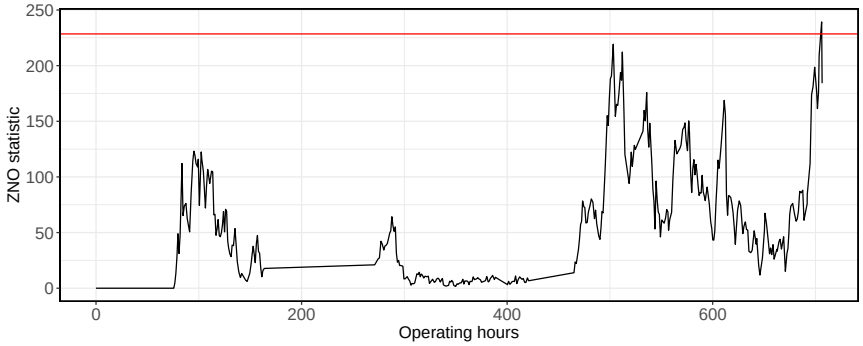
(a) MCZ chart during period 3 based on the manoeuvre model.



(b) ZNO chart during period 3 based on the manoeuvre model.



(c) MCZ chart during period 3 based on the combined model.



(d) ZNO chart during period 3 based on the combined model.

Figure A.9: Grouped RAV CUSUM charts during period 3 based on the manoeuvre and combined models.

Appendix B

Additional tables

Table B.1: Computed values of h for different values of k and ARL_0 for the two-sided CUSUM chart.

	k						
ARL_0	0.10	0.25	0.50	0.75	1.00	1.25	1.50
50	6.362	4.418	2.849	2.037	1.532	1.164	1
100	8.520	5.597	3.502	2.481	1.874	1.458	1.131
200	11.019	6.852	4.171	2.933	2.214	1.741	1.387
300	12.622	7.610	4.568	3.200	2.413	1.903	1.531
370	13.486	8.008	4.774	3.339	2.516	1.986	1.604
400	13.813	8.157	4.851	3.391	2.555	2.017	1.631
500	14.764	8.585	5.071	3.538	2.665	2.105	1.708
1000	17.846	9.931	5.757	3.999	3.009	2.379	1.942

Table B.2: Computed values of ρ for different values of λ and ARL_0 for the two-sided EWMA chart.

	λ							
ARL_0	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.75
50	0.845	1.520	1.811	2.054	2.166	2.229	2.268	2.315
100	1.152	1.879	2.148	2.360	2.453	2.504	2.534	2.568
200	1.500	2.216	2.454	2.635	2.713	2.754	2.777	2.802
300	1.710	2.399	2.619	2.785	2.854	2.890	2.911	2.931
370	1.819	2.490	2.701	2.859	2.925	2.959	2.978	2.996
400	1.859	2.523	2.731	2.886	2.950	2.984	3.002	3.020
500	1.973	2.615	2.814	2.962	3.023	3.054	3.071	3.087
1000	2.308	2.884	3.059	3.187	3.238	3.263	3.277	3.289

This table is copied from (Qiu, 2013, Section 5.2.1).

Table B.3: Correlation of the bearing temperature with the predictor variables.

Variable	**Correlation**
KOPPEL.SCHROEFAS.BB	0.96
BRANDSTOF.VERBRUIK.HVD.BB	0.95
TOERENTAL.HVD.BB	0.98
TEMPERATUUR.SPATOLIE.COMP.5.HVD.BB	0.89
TEMPERATUUR.SPATOLIE.COMP.4.HVD.BB	0.91
TEMPERATUUR.SPATOLIE.COMP.3.HVD.BB	0.88
TEMPERATUUR.SPATOLIE.COMP.2.HVD.BB	0.89
TEMPERATUUR.SPATOLIE.COMP.1.HVD.BB	0.87
DRUK.REGELLUCHT.HVD.BB	-0.01
DRUK.LU.NOODSTOP.HVD.BB	-0.01
BRANDSTOFREGELSTANGPOSITIE.HVD.BB	0.97
INST.SETPOINT.TOERENTAL.HVD.BB	0.98
WATER.IN.SMEEROLIE.HVD.BB	-0.14
DRUK.CARTER.HVD.BB	0.50
TEMPERATUUR.SMO.TOEVOER.HVD.BB	0.65
TEMPERATUUR.SMO.RETOUR.HVD.BB	0.97
DRUK.SMO.HVD.BB	0.80
DRUK.LT.ZOKOW.HVD.BB	0.97
DRUK.HT.ZOKOW.HVD.BB	0.97
TEMPERATUUR.HT.ZOKOW.RETOUR.HVD.BB	0.57
TEMPERATUUR.HT.ZOKOW.TOEVOER.HVD.BB	0.20
DRUK.ZEKOW.HVD.BB	0.97
DRUK.VULLUCHT.HVD.BB	0.88
TEMPERATUUR.VULLUCHT.NA.T.C.A.HVD.BB	0.94
TEMPERATUUR.VULLUCHT.NA.T.C.B.HVD.BB	0.45
TEMPERATUUR.VULLUCHT.VOOR.T.C.A.HVD.BB	-0.01
TEMPERATUUR.VULLUCHT.VOOR.T.C.B.HVD.BB	0.12
TEMPERATUUR.VULLUCHT.HVD.BB	0.53
DRUK.STARTLUCHT.HVD.BB	0.26
OMWENTELINGEN.T.C.A.HVD.BB	0.97
OMWENTELINGEN.T.C.B.HVD.BB	0.97
TEMPERATUUR.AFVOERGASSEN.NA.T.C.A.HVD.BB	0.40
TEMPERATUUR.AFVOERGASSEN.NA.T.C.B.HVD.BB	0.41
DRUK.BRANDSTOF.VOOR.HVD.BB	0.85
TEMPERATUUR.BRANDSTOF.TOEVOER.HVD.BB	0.36
TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB	0.93
TEMP.LT.ZOKOW.TOEVOER.HVD.BB	0.45
TEMP.LT.ZOKOW.RETOUR.HVD.BB	0.67
TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB	0.69
TEMP.GEMIDDELD.AFVOERGASSEN.B.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.A1.HVD.BB	0.71
TEMPERATUUR.AFVOERGASSEN.CIL.A2.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.A3.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.A4.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.A5.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.A6.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.B1.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.B2.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.B3.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.B4.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.B5.HVD.BB	0.69
TEMPERATUUR.AFVOERGASSEN.CIL.B6.HVD.BB	0.69

Table B.4: Clusters with a minimal pairwise correlation of 0.95 for the transit mode.

Cluster	Variables
1	TEMP.LT.ZOKOW.RETOUR.HVD.BB
	TEMP.LT.ZOKOW.TOEVOER.HVD.BB
2	TEMPERATUUR.AFVOERGASSEN.NA.T.C.A.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.NA.T.C.B.HVD.BB
3	TEMPERATUUR.SMO.RETOUR.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.4.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
4	TEMPERATUUR.SPATOLIE.COMP.2.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.3.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.4.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.5.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
5	TEMPERATUUR.SPATOLIE.COMP.1.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.3.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.4.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.5.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
6	DRUK.ZEKOW.HVD.BB
	KOPPEL.SCHROEFAS.BB
	OMWENTELINGEN.T.C.A.HVD.BB
	OMWENTELINGEN.T.C.B.HVD.BB
	TEMPERATUUR.VULLUCHT.NA.T.C.A.HVD.BB
	DRUK.HT.ZOKOW.HVD.BB
	DRUK.LT.ZOKOW.HVD.BB
	INST.SETPOINT.TOERENTAL.HVD.BB
	BRANDSTOFREGELSTANGPOSITIE.HVD.BB
	TOERENTAL.HVD.BB
	BRANDSTOF.VERBRUIK.HVD.BB
7	TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB
	TEMP.GEMIDDELD.AFVOERGASSEN.B.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A1.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A2.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A4.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A5.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A6.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B1.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B2.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B4.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B5.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B6.HVD.BB

Table B.5: Clusters with a minimal pairwise correlation of 0.95 (and 0.947) for the manoeuvre mode.

Cluster	**Variables**
1	TEMPERATUUR.AFVOERGASSEN.NA.T.C.A.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.NA.T.C.B.HVD.BB
2	TEMPERATUUR.SPATOLIE.COMP.1.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.2.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.3.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.4.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.5.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
	TEMPERATUUR.SMO.RETOUR.HVD.BB
3	DRUK.HT.ZOKOW.HVD.BB
	KOPPEL.SCHROEFAS.BB
	OMWENTELINGEN.T.C.A.HVD.BB
	OMWENTELINGEN.T.C.B.HVD.BB
	DRUK.ZEKOW.HVD.BB
	DRUK.LT.ZOKOW.HVD.BB
	INST.SETPOINT.TOERENTAL.HVD.BB
	BRANDSTOFREGELSTANGPOSITIE.HVD.BB
	TOERENTAL.HVD.BB
	BRANDSTOF.VERBRUIK.HVD.BB
4	TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB
	TEMP.GEMIDDELD.AFVOERGASSEN.B.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A1.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A2.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A4.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A5.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A6.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B1.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B2.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B4.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B5.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B6.HVD.BB

Table B.6: Clusters with a minimal pairwise correlation of 0.947 for the transit mode.

Cluster	Variables
1	TEMP.LT.ZOKOW.RETOUR.HVD.BB
	TEMP.LT.ZOKOW.TOEVOER.HVD.BB
2	TEMPERATUUR.AFVOERGASSEN.NA.T.C.A.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.NA.T.C.B.HVD.BB
3	TEMPERATUUR.SMO.RETOUR.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.4.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
4	TEMPERATUUR.SPATOLIE.COMP.1.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.2.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.3.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.4.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.5.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
5	DRUK.ZEKOW.HVD.BB
	KOPPEL.SCHROEFAS.BB
	OMWENTELINGEN.T.C.A.HVD.BB
	OMWENTELINGEN.T.C.B.HVD.BB
	TEMPERATUUR.VULLUCHT.NA.T.C.A.HVD.BB
	DRUK.HT.ZOKOW.HVD.BB
	DRUK.LT.ZOKOW.HVD.BB
	INST.SETPOINT.TOERENTAL.HVD.BB
	BRANDSTOFREGELSTANGPOSITIE.HVD.BB
	TOERENTAL.HVD.BB
	BRANDSTOF.VERBRUIK.HVD.BB
6	TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB
	TEMP.GEMIDDELD.AFVOERGASSEN.B.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A1.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A2.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A4.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A5.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A6.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B1.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B2.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B4.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B5.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B6.HVD.BB

Table B.7: Variables that are left after pairwise highly correlated variables are deleted.

Variables
TOERENTAL.HVD.BB
DRUK.REGELLUCHT.HVD.BB
DRUK.LU.NOODSTOP.HVD.BB
WATER.IN.SMEEROLIE.HVD.BB
DRUK.CARTER.HVD.BB
TEMPERATUUR.SMO.TOEVOER.HVD.BB
DRUK.SMO.HVD.BB
TEMPERATUUR.HT.ZOKOW.RETOUR.HVD.BB
TEMPERATUUR.HT.ZOKOW.TOEVOER.HVD.BB
DRUK.VULLUCHT.HVD.BB
TEMPERATUUR.VULLUCHT.NA.T.C.A.HVD.BB*
TEMPERATUUR.VULLUCHT.NA.T.C.B.HVD.BB
TEMPERATUUR.VULLUCHT.VOOR.T.C.A.HVD.BB
TEMPERATUUR.VULLUCHT.VOOR.T.C.B.HVD.BB
TEMPERATUUR.VULLUCHT.HVD.BB
DRUK.STARTLUCHT.HVD.BB
TEMPERATUUR.AFVOERGASSEN.NA.T.C.A.HVD.BB
DRUK.BRANDSTOF.VOOR.HVD.BB
TEMPERATUUR.BRANDSTOF.TOEVOER.HVD.BB
TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
TEMP.LT.ZOKOW.TOEVOER.HVD.BB*
TEMP.LT.ZOKOW.RETOUR.HVD.BB
TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB

Variables indicated with a * are still present in the manoeuvre mode, but are deleted for the transit mode.

Table B.8: Clusters with pairwise correlation of 0.90 for the transit mode.

Cluster	Variables
1	DRUK.LU.NOODSTOP.HVD.BB
	DRUK.REGELLUCHT.HVD.BB
2	TEMPERATUUR.VULLUCHT.VOOR.T.C.A.HVD.BB
	TEMPERATUUR.VULLUCHT.VOOR.T.C.B.HVD.BB
3	TEMPERATUUR.AFVOERGASSEN.NA.T.C.A.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.NA.T.C.B.HVD.BB
4	TEMP.LT.ZOKOW.RETOUR.HVD.BB
	TEMP.LT.ZOKOW.TOEVOER.HVD.BB
5	TEMPERATUUR.BRANDSTOF.TOEVOER.HVD.BB
	TEMP.LT.ZOKOW.TOEVOER.HVD.BB
6	TEMPERATUUR.SPATOLIE.COMP.1.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.2.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.3.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.4.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.5.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
	TEMPERATUUR.SMO.RETOUR.HVD.BB
7	TEMPERATUUR.SPATOLIE.COMP.2.HVD.BB
	TEMPERATUUR.SPATOLIE.COMP.3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A1.HVD.BB
8	TEMPERATUUR.SPATOLIE.COMP.3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B6.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A1.HVD.BB
	TEMP.GEMIDDELD.AFVOERGASSEN.B.HVD.BB
	TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB
9	TEMPERATUUR.SPATOLIE.COMP.6.HVD.BB
	TOERENTAL.HVD.BB
	TEMPERATUUR.SMO.RETOUR.HVD.BB
	INST.SETPOINT.TOERENTAL.HVD.BB
10	DRUK.VULLUCHT.HVD.BB
	KOPPEL.SCHROEFAS.BB
	OMWENTELINGEN.T.C.A.HVD.BB
	OMWENTELINGEN.T.C.B.HVD.BB
	TEMPERATUUR.VULLUCHT.NA.T.C.A.HVD.BB
	DRUK.ZEKOW.HVD.BB
	BRANDSTOFREGELSTANGPOSITIE.HVD.BB
	BRANDSTOF.VERBRUIK.HVD.BB
11	DRUK.ZEKOW.HVD.BB
	KOPPEL.SCHROEFAS.BB
	OMWENTELINGEN.T.C.A.HVD.BB
	OMWENTELINGEN.T.C.B.HVD.BB
	TEMPERATUUR.VULLUCHT.NA.T.C.A.HVD.BB
	DRUK.HT.ZOKOW.HVD.BB
	DRUK.LT.ZOKOW.HVD.BB
	TEMPERATUUR.SMO.RETOUR.HVD.BB
	INST.SETPOINT.TOERENTAL.HVD.BB
	BRANDSTOFREGELSTANGPOSITIE.HVD.BB
	TOERENTAL.HVD.BB
	BRANDSTOF.VERBRUIK.HVD.BB
12	TEMP.GEMIDDELD.AFVOERGASSEN.A.HVD.BB
	TEMP.GEMIDDELD.AFVOERGASSEN.B.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A1.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A2.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A4.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A5.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.A6.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B1.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B2.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B3.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B4.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B5.HVD.BB
	TEMPERATUUR.AFVOERGASSEN.CIL.B6.HVD.BB

Appendix C

MLR parameter estimation

C.1 Estimation of β

We assume that during the in-control period we obtain n observations $Y = (y_1, \ldots, y_n)^T$ that come from a normal distribution. During the same period we also observe n realizations of our p predictors. These observations can be used to obtain an estimate for the coefficient vector β. First, we will look at least squares estimation to obtain an estimate $\hat{\beta}$ for the coefficient vector β.

In least squares estimation, the sum of squares needs to be minimized, which is defined as

$$SS = (Y_{[1:n]} - X\beta)^T (Y_{[1:n]} - X\beta) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2. \tag{C.1}$$

In order to minimize this expression, we take the derivatives with respect to β_r $(r = 0, 1, \ldots, p)$ and equate these derivatives to 0. Hence, we obtain

$$\sum_{i=1}^{n} x_{ir} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right) = 0 \Leftrightarrow \sum_{i=1}^{n} x_{ir} y_i = \sum_{j=0}^{p} \left(\sum_{i=1}^{n} x_{ir} x_{ij} \right) \beta_j \tag{C.2}$$

A solution of the equations (C.2) satisfies

$$X^T X \hat{\beta} = X^T Y_{[1:n]} \tag{C.3}$$

which we call the normal equations. By Lemma 3.3 from Bingham and Fry (2010) we know that a solution of the normal equations exists if X has full rank. Usually the number of regressors is way smaller than the number of observations. Therefore, X has full rank if all the regressors are linearly independent. We assume that this is the case. Therefore, the solution to the normal equations is

$$\hat{\beta} = (X^T X)^{-1} X^T Y_{[1:n]}. \tag{C.4}$$

So, we now have obtained an estimator for β using the least squares estimation. Since we assumed normally distributed errors, we can also use maximum likelihood estimation to estimate β. The likelihood function is

$$L(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y_i - X_{i\bullet}\beta}{\sigma} \right)^2 \right), \tag{C.5}$$

and thus the log-likelihood function is

$$\ell(\beta, \sigma) = \log(L(\beta, \sigma)) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2. \tag{C.6}$$

From this it can be seen that maximizing the likelihood function with respect to β is equivalent to minimizing the sum of squares. So, the estimator for β we obtain via maximum likelihood estimation is the same as the estimator defined in equation (C.4). So both maximum likelihood estimation and least squares estimation lead to the same estimate $\hat{\beta}$ for β. Using Lemma D.0.1 the distribution of $\hat{\beta}$ can be derived to be

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X^T X)^{-1})\right). \tag{C.7}$$

The elaborate computations can be found in Van Dalen (2018).

In statistics, we like our estimators to be unbiased and of minimal variance. From the result in equation (C.7) it can be seen that $\hat{\beta}$ is an unbiased estimator of β. On top of that Theorem D.0.2 tells us that $\hat{\beta}$ is also of minimum variance, since $\hat{\beta}$ is the BLUE.

C.2 Estimation of σ

What rests us is to find an estimate of σ. Since we assumed normally distributed errors, we can do this via maximum likelihood estimation. Recall that

$$\ell(\beta, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2. \tag{C.8}$$

Taking the derivative with respect to σ of the log-likelihood and equating this to 0 gives

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2 = 0 \Leftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2. \tag{C.9}$$

At the maximum $\beta = \hat{\beta}$ we obtain

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j\right)^2. \tag{C.10}$$

Note that the minimized sum of squares, the sum of squares for error (SSE), is defined as

$$SSE = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j\right)^2 = (Y_{[1:n]} - X\hat{\beta})^T(Y_{[1:n]} - X\hat{\beta}). \tag{C.11}$$

Thus, it follows that $\hat{\sigma}^2 = \frac{SSE}{n}$ is an estimator of σ^2. Theorem D.0.8 tells us that $\hat{\sigma}^2 \sim n\sigma^2\chi^2_{(n-p-1)}$.

So, now we know the distribution of both $\hat{\beta}$ and $\hat{\sigma}^2$. At last, it is good to notice that $\hat{\sigma}^2$ is not an unbiased estimator of σ^2 as the expected value of $\hat{\sigma}^2$ is $\frac{n-p-1}{n}\sigma^2$.

Appendix D

Additional theorems and lemmas

Lemma D.0.1. *Let $Y \in \mathbb{R}^n$ be a random vector and let $X \in \mathbb{R}^{m \times n}$ be a linear transformation, then $\mathbb{E}[XY] = X\mathbb{E}[Y]$ and $Var(XY) = X\,Var(Y)X^T$.*

Proof. The expression for the expectation follows directly by the linearity of the expectation operator. Then, for the (co)variance we have:

$$\begin{aligned}
\mathrm{Cov}(Xy_i, Xy_j) &= \mathbb{E}[(Xy_i - \mathbb{E}[Xy_i])(Xy_j - \mathbb{E}[Xy_j])] \\
&= \mathbb{E}[\textstyle\sum_k x_{ik}(y_k - \mathbb{E}[y_k]) \sum_s x_{js}(y_s - \mathbb{E}[y_s])] \\
&= \textstyle\sum_{ks} x_{ik}x_{js}\mathbb{E}([y_k - \mathbb{E}[y_k]](y_s - \mathbb{E}[y_s])] \\
&= \textstyle\sum_{ks} x_{ik}x_{js}\mathrm{Cov}(y_i, y_j) \\
&= X\mathrm{Cov}(y_i, y_j)X^T.
\end{aligned}$$

This leads to $\mathrm{Var}(XY) = X\mathrm{Var}(Y)X^T$. □

Theorem D.0.2 (Gauss-Markov Theorem)**.** *Among all linear unbiased estimators $\tilde{\beta} = Cy$ of β, the least-squares estimator $\hat{\beta} = (X^TX)^{-1}X^Ty$ has the minimum variance in each component. That is $\hat{\beta}$ is the BLUE.*

Proof. Let $\tilde{\beta} = Cy$ be another linear estimator of β. Then we can write

$C = D + (X^TX)^{-1}X^T$.

Using this result it follows that

$\tilde{\beta} = Cy = Dy + (X^TX)X^Ty = Dy + \hat{\beta}$.

The expected value of $\tilde{\beta}$ is

$$\begin{aligned}
\mathbb{E}(\tilde{\beta}) &= \mathbb{E}(Dy) + \mathbb{E}(\hat{\beta}) \\
&= D\mathbb{E}(y) + \beta \\
&= DX\beta + \beta \\
&= (DX + I)\beta.
\end{aligned}$$

Thus $\tilde{\beta}$ is an unbiased estimator of β if and only if $DX = 0$. If we look at the variance of $\tilde{\beta}$ we see that

$$\begin{aligned}
\text{Var}(\tilde{\beta}) &= \text{Var}(Cy) \\
&= \text{Var}((D + (X^TX)^{-1}X^T)y) \\
&= (D + (X^TX)^{-1}X^T)\text{Var}(y)(D + (X^TX)^{-1}X^T) \\
&= \sigma^2(D + (X^TX)^{-1}X^T)(D + (X^TX)^{-1}X^T)^T \\
&= \sigma^2(DD^T + DX(X^TX)^{-1} + (X^TX)^{-1}X^TDT + (X^TX)^{-1}X^TX(X^TX)^{-1}) \\
&= \sigma^2(DD^T + (X^TX)^{-1}) \\
&= \text{Var}(\hat{\beta}) + \sigma^2DD^T.
\end{aligned}$$

Because of Lemma D.0.3 we know that DD^T is positive semi-definite. Therefore, we have that $\hat{\beta}$ has minimal variance among all linear unbiased estimators and thus that $\hat{\beta}$ is the BLUE. □

Lemma D.0.3. *Let $X \in \mathbb{R}^{n\times m}$, then XX^T is positive semi-definite (nonnegative definite).*

Proof. A matrix A is positive semi-definite iff

$$z^TAz \geq 0, \quad \text{for all } z \neq 0. \tag{D.1}$$

Let $A = XX^T$, then

$$z^TAz = z^TXX^Tz = (X^Tz)^TXTz = ||X^Tz||_2^2 \geq 0. \tag{D.2}$$

This holds for all $z \neq 0$, therefore XX^T is positive semi-definite. □

Lemma D.0.4. *Let $P = X(X^TX)^{-1}X^T$, then P and $(I - P)$ are idempotent.*

Proof.

$$P^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = P$$
$$(I - P)^2 = I^2 - 2IP + P^2 = I - 2P + P = I - P$$

So indeed, P and$(I - P)$ are idempotent and thus projections. □

Lemma D.0.5. *Let P be a projection matrix, then its eigenvalues λ are 0 and 1 and the trace of P is equal to its rank.*

Proof. Let λ be an eigenvalue of P with corresponding eigenvector $x \neq 0$. Then $Px = \lambda x$. Furthermore,

$$P^2x = P(Px) = P(\lambda x) = \lambda(Bx) = \lambda^2x. \tag{D.3}$$

So if λ is an eigenvalue of P, then λ^2 is an eigenvalue of P^2. But, since P is a projection matrix, we have $P^2 = P$ and thus $\lambda^2 = \lambda$. Therefore, λ must be either 0 or 1. Since the trace of a matrix is equal to the sum of its non-zero eigenvalues we obtain that $\text{trace}(P) = \text{rank}(P)$. □

Lemma D.0.6. *Let P be a symmetric projection matrix of rank r and let x_i $\mathcal{N}(0, \sigma^2)$ independent, then*

$$x^TPx \sim \sigma^2\chi_r^2. \tag{D.4}$$

Proof. Since P is idempotent we have by Lemma D.0.5 that its eigenvalues are 0 and 1. Let r be the rank of P. Then, P can be diagonalized to a matrix with first r ones one the diagonal followed by $n - r$ zeros. This means that x^TPx can be written as the sum of r squares of independent normally distributed random variables. The χ_{df}^2-distribution is defined as the sum of the squares of *df* independent standard normal random variables. And thus, $x^TPx \sim \sigma^2\chi_r^2$. □

Lemma D.0.7. *Let P be defined as $P = X(X^TX)^{-1}X^T$, then rank$(P) = p + 1$ and rank$(I - P) = n - p - 1$.*

Proof. Since P is idempotent, by Lemma D.0.5 that $\text{rank}(P) = \text{trace}(P)$. Therefore,

$$\begin{aligned}\text{rank}(P) &= \text{trace}(P)\\ &= \text{trace}(X(X^TX)^{-1}X^T)\\ &= \text{trace}((X^TX)^{-1}X^TX)\\ &= \text{trace}(I),\end{aligned}$$

where I is the $((p+1)\times(p+1))$-identity matrix. So it follows that $\text{rank}(P) = p+1$. If we use this result we obtain

$$\begin{aligned}\text{rank}(I-P) &= \text{trace}(I-P)\\ &= \text{trace}(I) - \text{trace}(P)\\ &= n-p-1,\end{aligned}$$

where I is the $(n\times n)$-identity matrix. □

Theorem D.0.8. $SSE \sim \sigma^2\chi^2_{(n-p-1)}$.

Proof. Define $X\hat{\beta} = X(X^TX)^{-1}X^Ty = Py$. Then by Lemma D.0.4 P and $(I-P)$ are idempotent. We can rewrite the SSE as

$$\begin{aligned}SSE &= (y - X\hat{\beta})^T(y - X\hat{\beta})\\ &= (y-Py)^T(y-Py)\\ &= y^T(I-P)(I-P)y\\ &= y^T(I-P)^{1/2}(I-P)(I-P)^{1/2}\\ &= \tilde{y}^T(I-P)\tilde{y}.\end{aligned}$$

Making use of Lemma D.0.1 we can derive the expectation and variance of $\tilde{y}$:

$$\begin{aligned}\mathbb{E}(\tilde{y}) &= \mathbb{E}((I-P)^{1/2}y)\\ &= (I-P)^{1/2}X\beta\\ &= (I-P)^{-1/2}(I-P)X\beta\\ &= (I-P)^{-1/2}(X\beta - X\beta)\\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Var}(\tilde{y}) &= \text{Var}((I-P)^{1/2}y)\\ &= \text{Var}((I-P)^{-1/2}(I-P)y)\\ &= (I-P)^{-1/2}(I-P)\text{Var}(y)((I-P)^{-1/2}(I-P))^T\\ &= \sigma^2(I-P)^{-1/2}(I-P)(I-P)(I-P)^{-1/2}\\ &= \sigma^2 I\end{aligned}$$

So, $\tilde{y} \sim \mathcal{N}(0, \sigma^2 I)$. Therefore, by Lemma D.0.6 $SSE \sim \sigma^2\chi^2_r$, where r is the rank of $(I-P)$. The rank of $(I-P) = n-p-1$ by Lemma D.0.7. And thus, $SSE \sim \sigma^2\chi^2_{(n-p-1)}$. □

Appendix E

R functions

In this appendix some of the most important R functions that are used in this research are discussed. The input and output arguments are explained, but we do not explain how these functions exactly work.

E.1 Predictive residuals

The R function `Predictive.Residuals()` computes the residuals via the predictive method. The first input argument `data` is a data frame that contains the bearing temperatures, the predictor variables and the operating hours. The second and third input arguments `stab_per` and `mon_per` are arrays of length 2 that contain the beginning and the end of the stable period and the period during which the bearing temperatures are monitored. The last input argument `formula` is a string that indicates which formula is used in the regression model. These formulas should be specified beforehand.

The output of this function is a list of two data frames. The first data frame contains the residuals for all the bearings during the stable period. The second data frame contains the residuals for all bearings during the monitoring period. Both data frames are provided with the operating hours.

```
Predictive.Residuals <- function(data, stab_per, mon_per, formula){
    stab_data = subset(data, DRAAI.UREN >= stab_per[[1]] & DRAAI.UREN <= stab_per[[2]])
    mon_data = subset(data, DRAAI.UREN >= mon_per[[1]] & DRAAI.UREN <= mon_per[[2]])
    y = grep("^TEMPERATUUR.HOOFDLAGER", colnames(data))
    Y = mon_data[,y]
    X = mon_data[,-y]
    stab_X = stab_data[,-y]
    X = X[,-ncol(X)]
    stab_X = stab_X[,-ncol(stab_X)]
    ONES = rep(1, nrow(X))
    X = as.matrix(cbind(ONES, X))
    ONES = rep(1, nrow(stab_X))
    stab_X = as.matrix(cbind(ONES, stab_X))

    if(formula == 'transit'){
        formula = formula_transit
    }
    else if(formula == 'manoeuvre'){
        formula = formula_manoeuvre
    }
    else if(formula == 'combined'){
        formula = formula_combined
    }

    RES_mon = matrix(rep(0,7*nrow(Y)), nrow(Y), 7)
    RES_stab = matrix(rep(0,7*nrow(stab_data)), nrow(stab_data),7)
    for(i in 1:7){
        model = lm(paste(paste(colnames(Y)[[i]],"~"),formula), data=stab_data)
        RES_stab[,i] = rstandard(model)
        sigma = summary(model)$sigma
        pred = predict(model, mon_data)
```

```
        e = Y[,i] - pred
        XTX_inv = solve(t(stab_X)%*%stab_X)
        var = apply(X, 1, function(X, XTX_inv) X%*%XTX_inv%*%as.matrix(X)+1, XTX_inv=XTX_inv)
        res = (e/(sigma*sqrt(var)))
        RES_mon[,i] = res
    }
    RES_stab = data.frame(RES_stab)
    RES_mon = data.frame(RES_mon)
    RES_mon$DRAAI.UREN = mon_data$DRAAI.UREN - min(mon_data$DRAAI.UREN)
    RES_stab$DRAAI.UREN = stab_data$DRAAI.UREN - min(stab_data$DRAAI.UREN)
    colnames(RES_mon) = c("TEMPERATUUR.HOOFDLAGER.1.HVD.BB", "TEMPERATUUR.HOOFDLAGER.2.HVD.BB
    ", "TEMPERATUUR.HOOFDLAGER.3.HVD.BB", "TEMPERATUUR.HOOFDLAGER.4.HVD.BB", "TEMPERATUUR.
    HOOFDLAGER.5.HVD.BB", "TEMPERATUUR.HOOFDLAGER.6.HVD.BB",
    "TEMPERATUUR.HOOFDLAGER.7.HVD.BB", "DRAAI.UREN")
    colnames(RES_stab) = colnames(RES_mon)
    return(list(RES_stab, RES_mon))
}
```

The R function `DF.Shewhart.Pred()` creates data frames that contain all the input that is necessary to construct a Shewhart chart of the predictive residuals. The first input argument `stab_res` is the data frame that contains the predictive residuals for all bearings during the stable period. The second input argument `mon_res` is the data frame that contains the predictive residuals of all bearings during the monitoring period. The last input argument `sides` is a scalar that indicates if a two-sided or a one-sided chart will be created.

The output of this function is a list of seven data frames. Each data frame contains the inputs for the Shewhart chart of a specific bearing. In these data frames the residuals, mean, UCL and operating hours are stored and in case of a two-sided chart also the LCL is included.

```
DF.Shewhart.Pred <- function(stab_res, mon_res, sides){
    DF_Shewhart = c()
    for(i in 1:7){
        e = mon_res[,i]
        sigma = sd(stab_res[,i])
        n = nrow(stab_res)
        p = 11
        pred_res_shew = data.frame(e)
        pred_res_shew$MEAN = mean(e)
        if(sides == 2){
            pred_res_shew$UCL = rep(mean(stab_res[,i]) + qt(1-(0.0027/2), n-p)*sigma, length(e
    ))
            pred_res_shew$LCL = rep(mean(stab_res[,i]) + qt(0.0027/2, n-p)*sigma, length(e))
        }
        if(sides == 1){
            pred_res_shew$UCL = rep(mean(stab_res[,i]) + qt(1-(0.0027), n-p)*sigma, length(e))
        }
        pred_res_shew$DRAAI.UREN = mon_res$DRAAI.UREN - min(mon_res$DRAAI.UREN)
        DF_Shewhart[[i]] = pred_res_shew
    }
    return(DF_Shewhart)
}
```

The R function `DF.EWMA.Pred()` creates data frames that contain all the input that is necessary to construct an EWMA chart of the predictive residuals. The input arguments are the same as the input arguments of `DF.Shewhart.Pred()`, except for the argument `lambda`. This argument represents the constant λ that is used in the EWMA chart.

The output of this function is a list of seven data frames. Each data frame contains all the inputs that are necessary to construct an EWMA chart for a specific bearing. In this data frame the EWMA-statistic, mean, UCL and operating hours are stored and in case of a two-sided chart also the LCL is included in the data frame.

```
DF.EWMA.Pred <- function(stab_res, mon_res, lambda, sides){
    DF_EWMA = c()
    for(i in 1:7){
        e = mon_res[,i]
        E = ewmaSmooth((1:length(e)),e,lambda)$y
        E_stab = ewmaSmooth((1:length(stab_res[,i])),stab_res[,i],lambda)$y
        sigma_control = sqrt(lambda/(2-lambda))*sd(scale(stab_res[,i]))
```

```
        if(sides == 2){
            rho = xewma.crit(lambda, 370, sided="two")
        }
        if(sides == 1){
            rho = xewma.crit(lambda, 370, sided="one")
        }
        ucl_EWMA = rho*sigma_control

        pred_res_ewma = data.frame(E)
        pred_res_ewma$MEAN = rep(mean(E), length(E))
        pred_res_ewma$UCL = mean(E_stab) + ucl_EWMA
        pred_res_ewma$LCL = mean(E_stab) - ucl_EWMA
        pred_res_ewma$DRAAI.UREN = mon_res$DRAAI.UREN - min(mon_res$DRAAI.UREN)
        colnames(pred_res_ewma) = c("E", "MEAN", "UCL", "LCL", "DRAAI.UREN")
        DF_EWMA[[i]] = pred_res_ewma
        }
    return(DF_EWMA)
}
```

E.2 Recursive residuals

The R function `Recusive.Residuals()` computes the residuals via the recursive method. The `data`, `mon_per` and `formula` input arguments of `Recursive.Residuals()` are the same as in `Predictive.Residuals()`. The third input argument `m` indicates the length of the initialization period.

The output of the `Recursive.Residuals()` function is a list of two data frames and a scalar. The first data frame `RES` contains the recursive residuals of the seven main bearings. This data frame is also provided with the operating hours. The second data frame `SIGMA` contains the estimates of the standard deviation of the recursive residuals for every iteration. The scalar `jstart` indicates the length of the initialization period. If the transit or manoeuvre model is used, then `jstart` is equal to `m+1`. If the combined model is used, then `jstart` is equal to the iteration in which enough observations in both the transit and manoeuvre mode are obtained.

```
Recursive.Residuals <- function(data, mon_per, m, formula){
    data_mon = subset(data, DRAAI.UREN >= mon_per[[1]] & DRAAI.UREN <= mon_per[[2]])
    y = grep("^TEMPERATUUR.HOOFDLAGER", colnames(data))
    Y = data_mon[,y]
    X = data_mon[,-y]
    X = X[,-ncol(X)]
    ONES = rep(1, nrow(X))
    X = as.matrix(cbind(ONES, X))

    if(formula == 'transit'){
        formula = formula_transit
    }
    else if(formula == 'manoeuvre'){
        formula = formula_manoeuvre
    }
    else if(formula == 'combined'){
        formula = formula_combined
    }

    RES = matrix(rep(0,7*nrow(Y)), nrow(Y), 7)
    SIGMA = matrix(rep(0,7*nrow(Y)), nrow(Y), 7)
    n = nrow(data_mon)
    p = ncol(X)
    count_manoeuvre = 0
    count_transit = 0
    jstart = m

    while((0!=1)){
        j = j+1
        if(j > (n-1)){
            break
        }
```

```
        if(formula == formula_combined){
            if(count_manoeuvre <= p-1+10 | count_transit <= p-1+10){
                count_manoeuvre = sum(data_mon[(1:(j+1)),]$INDICATOR)
                count_transit = j + 1 - count_manoeuvre
                jstart = j
             }
            else{
               for(i in 1:7){
                model <- lm(paste(paste(colnames(Y)[[i]],"~"),formula), data_mon[(1:j),])
                suppressWarnings(pred <- predict(model, newdata=data_mon[(j+1),]))
                res <- Y[(j+1),i] - pred
                var <- 1 + X[(j+1),]%*%solve(t(X[1:(j),])%*%X[1:(j),])%*%as.matrix(X[(j+1),])
                sigma <- summary(model)$sigma
                w = (res/(sigma*sqrt(var)))*(sqrt(n-p)/sqrt(n))

                RES[j+1,i] = w
                SIGMA[j+1,i] = sd(rstudent(model))
                }
            }
        }
        else if(formula == formula_transit | formula == formula_manoeuvreer){
            jstart = m+1

            for(i in 1:7){
                model <- lm(paste(paste(colnames(Y)[[i]],"~"),formula), data_mon[(1:j),])
                suppressWarnings(pred <- predict(model, newdata=data_mon[(j+1),]))
                res <- Y[(j+1),i] - pred
                var <- 1 + X[(j+1),]%*%solve(t(X[1:(j),])%*%X[1:(j),])%*%as.matrix(X[(j+1),])
                sigma <- summary(model)$sigma
                w = (res/(sigma*sqrt(var)))*(sqrt(n-p)/sqrt(n))

                RES[j+1,i] = w
                SIGMA[j+1,i] = sd(rstudent(model))
            }
        }
    }

    for(i in 1:7){
        SIGMA[(1:jstart),i] = rep(SIGMA[(jstart+1),i], jstart)
    }

    RES = data.frame(RES)
    RES$DRAAI.UREN = data_mon$DRAAI.UREN - min(data_mon$DRAAI.UREN)
    colnames(RES) = c("TEMPERATUUR.HOOFDLAGER.1.HVD.BB", "TEMPERATUUR.HOOFDLAGER.2.HVD.BB",
                  "TEMPERATUUR.HOOFDLAGER.3.HVD.BB", "TEMPERATUUR.HOOFDLAGER.4.HVD.BB",
      "TEMPERATUUR.HOOFDLAGER.5.HVD.BB", "TEMPERATUUR.HOOFDLAGER.6.HVD.BB", "TEMPERATUUR.
    HOOFDLAGER.7.HVD.BB", "DRAAI.UREN")
    colnames(SIGMA) = colnames(RES)[1:7]
    return(list(RES,SIGMA,jstart))
}
```

The R function `DF.Shewhart.Rec()` creates data frames that contain all the input that is necessary to construct a Shewhart chart of the recursive residuals. The first input argument `res` is the data frame that contains the recursive residuals for all bearings. The second input argument `sigma` is a data frame with the standard deviation of the recursive residuals in each iteration as outputted by `Recursive.Residuals()`. The third input argument `m` is a scalar that represents the length of the initialization period. The last input argument `sides` is a scalar that indicates if a two-sided or a one-sided chart will be created.

The output of this function is a list of seven data frames. Each data frame contains the inputs for the Shewhart chart of a specific bearing. In these data frames the residuals, mean, UCL and operating hours are stored and in case of a two-sided chart also the LCL is included.

```
DF.Shewhart.Rec <- function(res, sigma, m, sides){
    DF_Shewhart = c()
    for(i in 1:7){
        e = res[,i]
```

```
        n = nrow(res)
        p = 11
        ucl_shew = c()
        lcl_shew = c()
        for(k in (m+1):n){
            if(sides == 2){
                ucl_shew[k-m] = qt(1-(0.0027/2), k-p)
                lcl_shew[k-m] = qt(0.0027/2, k-p)
            }
            if(sides == 1){
                ucl_shew[k-m] = qt(1-(0.0027), k-p)
                lcl_shew[k-m] = 0
            }
        }

        cummean = cummean(e[(m+1):n])
        rec_res_shew = data.frame(e)
        rec_res_shew$MEAN = rbind(matrix(rep(0,m),m,1), as.matrix(cummean))
        rec_res_shew$UCL = rec_res_shew$MEAN + rbind(matrix(rep(0,m),m,1), as.matrix(ucl_shew)
    )
        rec_res_shew$LCL = rec_res_shew$MEAN + rbind(matrix(rep(0,m),m,1), as.matrix(lcl_shew)
    )
        rec_res_shew$DRAAI.UREN = res$DRAAI.UREN - min(res$DRAAI.UREN)
        DF_Shewhart[[i]] = rec_res_shew
    }
    return(DF_Shewhart)
}
```

The R function `DF.EWMA.Rec()` creates data frames that contain all the input that is necessary to construct an EWMA chart of the recursive residuals. The input arguments are the same as the input arguments of `DF.Shewhart.Rec()`, except for the argument `lambda`. This argument represents the constant λ that is used in the EWMA chart.

The output of this function is a list of seven data frames. Each data frame contains all the inputs that are necessary to construct an EWMA chart for a specific bearing. In this data frame the EWMA-statistic, mean, UCL and operating hours are stored and in case of a two-sided chart also the LCL is included in the data frame.

```
DF.EWMA.Rec <- function(res, sigma, lambda, sides, m){
    DF_EWMA = c()
    for(i in 1:7){
        e = res[,i]
        E = ewmaSmooth((1:length(e)),e,lambda)$y
        SIGMA = sqrt(lambda/(2-lambda))*sigma[,i]
        if(sides == 2){
            rho = xewma.crit(lambda, 370, sided="two")
        }
        if(sides == 1){
            rho = xewma.crit(lambda, 370, sided="one")
        }
        ucl_EWMA = rho*SIGMA

        n = nrow(res)
        cummean = cummean(E[(m+1):n])

        rec_res_ewma = data.frame(E)
        rec_res_ewma$MEAN = rbind(matrix(rep(0,m),m,1), as.matrix(cummean))
        rec_res_ewma$UCL = rec_res_ewma$MEAN + ucl_EWMA
        rec_res_ewma$LCL = rec_res_ewma$MEAN - ucl_EWMA
        rec_res_ewma$DRAAI.UREN = res$DRAAI.UREN - min(res$DRAAI.UREN)
        colnames(rec_res_ewma) = c("E", "MEAN", "UCL", "LCL", "DRAAI.UREN")
        DF_EWMA[[i]] = rec_res_ewma
    }
    return(DF_EWMA)
}
```

E.3 Hybrid residuals

The R function `Hybrid.Residuals()` computes the residuals via the hybrid method. The input arguments of `Hybrid.Residuals()` are the same as the input arguments of `Recursive.Residuals()`. The output of `Hybrid.Residuals()` is a list of two data frames and two scalars. The first data frame `RES` contains the hybrid residuals for all the bearings. This data frame is also provided with the operating hours. The second data frame `SIGMA` contains the estimates of the standard deviation for the hybrid residuals in each iteration. The first scalar `tau` represents the changepoint. This is the point where the switch from the recursive to the predictive method is made. The second scalar `jstart` represents the length of the initialization period, just as in `Recursive.Residuals()`.

```
Hybrid.Residuals <- function(data, mon_per, m, formula){
    data_mon = subset(data, DRAAI.UREN >= mon_per[[1]] & DRAAI.UREN <= mon_per[[2]])
    y = grep("^TEMPERATUUR.HOOFDLAGER", colnames(data))
    Y = data_mon[,y]
    X = data_mon[,-y]
    X = X[,-ncol(X)]
    ONES = rep(1, nrow(X))
    X = as.matrix(cbind(ONES, X))

    if(formula == 'transit'){
        formula = formula_transit
    }
    else if(formula == 'manoeuvre'){
        formula = formula_manoeuvre
    }
    else if(formula == 'combined'){
        formula = formula_combined
    }

    RES = matrix(rep(0,8*nrow(Y)), nrow(Y), 8)
    SIGMA = matrix(rep(0,7*nrow(Y)), nrow(Y), 7)
    n = nrow(data_mon)
    p = ncol(X)
    count_manoeuvre = 0
    count_transit = 0
    jstart=0
    check = TRUE

    j = m
    while(check == TRUE | count_manoeuvre < 20 | count_transit < 20){
        j = j+1
        if(j > (n-1)){
            break
        }
        if(formula == formula_combined){
            if(count_manoeuvre <= p-1+10 | count_transit <= p-1+10){
                count_manoeuvre = sum(data_mon[(1:(j+1)),]$INDICATOR)
                count_transit = j + 1 - count_manoeuvre
                jstart = j
            }
             else{
                for(i in 1:7){
                model <- lm(paste(paste(colnames(Y)[[i]],"~"),formula), data_mon[(1:j),])
                suppressWarnings(pred <- predict(model, newdata=data_mon[(j+1),]))
                res <- Y[(j+1),i] - pred
                var <- 1 + X[(j+1),]%*%solve(t(X[1:(j),])%*%X[1:(j),])%*%as.matrix(X[(j+1),])
                sigma <- summary(model)$sigma
                w = (res/(sigma*sqrt(var)))*(sqrt(n-p)/sqrt(n))

                RES[j+1,i] = w
                SIGMA[j+1,i] = sd(rstudent(model))
                }

                count_manoeuvre = sum(data_mon[(1:(j+1)),]$INDICATOR)
```

```
                count_transit = j+1-count_manoeuvre

                if(j > (m+20)){
                    rmse2 = mean(RMSE[(j-9):j])
                    rmse1 = mean(RMSE[(j-19):(j-10)])
                    if(rmse2 > (1-k)*rmse1 & rmse2 < (1+k)*rmse1){
                    check = FALSE
                    }
                }
            }
        }

       else if(formula == formula_transit | formula == formula_manoeuvreer){

            jstart = m+1

            for(i in 1:7){
                model <- lm(paste(paste(colnames(Y)[[i]],"~"),formula), data_mon[(1:j),])
                suppressWarnings(pred <- predict(model, newdata=data_mon[(j+1),]))
                res <- Y[(j+1),i] - pred
                var <- 1 + X[(j+1),]%*%solve(t(X[1:(j),])%*%X[1:(j),])%*%as.matrix(X[(j+1),])
                sigma <- summary(model)$sigma
                w = (res/(sigma*sqrt(var)))*(sqrt(n-p)/sqrt(n))

                RES[j+1,i] = w
                SIGMA[j+1,i] = sd(rstudent(model))
                sum = summary(model)
            }

            count_manoeuvre = 30
            count_transit = 30

            if(j > (m+20)){
                rmse2 = mean(RMSE[(j-9):j])
                rmse1 = mean(RMSE[(j-19):(j-10)])
                if(rmse2 > (1-k)*rmse1 & rmse2 < (1+k)*rmse1){
                    check = FALSE
                }
            }
        }
    }


    SIGMA = SIGMA[(1:(j-1)),]
    for(i in 1:7){
        SIGMA[(1:jstart),i] = rep(SIGMA[(jstart+1),i], jstart)
    }

    tau = j-1

    stab_X = X[(1:(tau-1)),]
     for(i in 1:7){
            model = lm(paste(paste(colnames(Y)[[i]],"~"),formula), data=data_mon[(1:(tau-1))
    ,])
            sigma = summary(model)$sigma
            pred = predict(model, data_mon[(tau:n),])
            e = Y[(tau:n),i] - pred
            XTX_inv = solve(t(stab_X)%*%stab_X)
            var = apply(X, 1, function(X, XTX_inv) X%*%XTX_inv%*%as.matrix(X)+1, XTX_inv=
    XTX_inv)
            res = (e/(sigma*sqrt(var[(tau:n)])))
            RES[(tau:n),i] = res
        }
    RES[,8] = data_mon$DRAAI.UREN - min(data_mon$DRAAI.UREN)
    RES = data.frame(RES)
    colnames(RES) = c("1", "2", "3", "4", "5", "6", "7", "DRAAI.UREN")
    return(list(RES,SIGMA,tau,jstart))
```

```
}
```

The R function `DF.Shewhart.Hybrid()` creates data frames that contain all the input that is necessary to construct a Shewhart chart of the hybrid residuals. The first input argument `res` is the data frame that contains the hybrid residuals for all bearings. The second input argument `m` is a scalar that represents the length of the initialization period. The third input argument `tau` is the changepoint. The last input argument `sides` is a scalar that indicates if a two-sided or a one-sided chart will be created.

The output of this function is a list of seven data frames. Each data frame contains the inputs for the Shewhart chart of a specific bearing. In these data frames the residuals, mean UCL and operating hours are stored and in case of a two-sided chart also the LCL is included.

```
DF.Shewhart.Hybrid <- function(res, m, tau, sides){
    DF_Shewhart = c()
    for(i in 1:7){
        e = res[,i]
        n = nrow(res)
        p = 11
        ucl_shew = c()
        lcl_shew = c()
        for(k in (m+1):n){
            if(sides == 2){
                ucl_shew[k-m] = qt(1-(0.0027/2), k-p)
                lcl_shew[k-m] = qt(0.0027/2, k-p)
            }
            if(sides == 1){
                ucl_shew[k-m] = qt(1-(0.0027), k-p)
                lcl_shew[k-m] = 0
            }
        }

        cummean = cummean(e[(m+1):tau])
        mean_rec = rbind(matrix(rep(0,m),m,1), as.matrix(cummean))
        mean_pred = matrix(rep(tail(mean_rec,1), (nrow(res)-tau)), nrow(res)-tau, 1)
        mean = rbind(mean_rec, mean_pred)
        hyb_res_shew = data.frame(e)
        hyb_res_shew$MEAN = mean
        hyb_res_shew$UCL = hyb_res_shew$MEAN + rbind(matrix(rep(0,m),m,1), as.matrix(ucl_shew)
    )
        hyb_res_shew$LCL = hyb_res_shew$MEAN + rbind(matrix(rep(0,m),m,1), as.matrix(lcl_shew)
    )
        #hyb_res_shew$DRAAI.UREN = res$DRAAI.UREN - min(res$DRAAI.UREN)
        hyb_res_shew$DRAAI.UREN = res$DRAAI.UREN
        DF_Shewhart[[i]] = hyb_res_shew
    }
    return(DF_Shewhart)
}
```

The R function `DF.EWMA.Hybrid()` creates data frames that contain all the input that is necessary to construct an EWMA chart of the hybrid residuals. The input arguments are the same as the input arguments of `DF.Shewhart.Hybrid()`, except for the argument `lambda`. This argument represents the constant λ that is used in the EWMA chart.

The output of this function is a list of seven data frames. Each data frame contains all the inputs that are necessary to construct an EWMA chart for a specific bearing. In this data frame the EWMA-statistic, mean, UCL and operating hours are stored and in case of a two-sided chart also the LCL is included in the data frame.

```
DF.EWMA.Hybrid <- function(res, sigma, m, lambda, tau, sides){
    DF_EWMA = c()
    for(i in 1:7){
        e = res[,i]
        E = ewmaSmooth((1:length(e)),e,lambda)$y
        SIGMA = sqrt(lambda/(2-lambda))*sigma[,i]
        if(sides == 2){
            rho = xewma.crit(lambda, 370, sided="two")
        }
        if(sides == 1){
```

```
            rho = xewma.crit(lambda, 370, sided="one")
        }
        ucl_EWMA_rec = matrix(rho*SIGMA, tau, 1)
        ucl_EWMA_rec[(1:m),] = rep(0,m)
        ucl_EWMA_pred = matrix(rep(tail(ucl_EWMA_rec,1), nrow(res)-tau), nrow(res)-tau, 1)
        ucl_EWMA = rbind(ucl_EWMA_rec, ucl_EWMA_pred)
        cummean = cummean(E[(m+1):tau])
        mean_rec = rbind(matrix(rep(0,m),m,1), as.matrix(cummean))
        mean_pred = matrix(rep(tail(mean_rec,1), (nrow(res)-tau)), nrow(res)-tau, 1)
        mean = rbind(mean_rec, mean_pred)

        hyb_res_ewma = data.frame(E)
        hyb_res_ewma$MEAN = mean
        hyb_res_ewma$UCL = hyb_res_ewma$MEAN + ucl_EWMA
        hyb_res_ewma$LCL = hyb_res_ewma$MEAN - ucl_EWMA
        hyb_res_ewma$DRAAI.UREN = res$DRAAI.UREN - min(res$DRAAI.UREN)
        colnames(hyb_res_ewma) = c("E", "MEAN", "UCL", "LCL", "DRAAI.UREN")
        DF_EWMA[[i]] = hyb_res_ewma
    }
    return(DF_EWMA)
}
```

E.4 Recursive coefficients

The R function `Recursive.Coefficients()` computes the difference between two successive recursively estimated coefficient vectors. The input arguments of this function are the same as the input arguments of `Recursive.Residuals()` and `Hybrid.Residuals()`. The output of this function is a list that contains two other lists. The first list `RECCOEF` is a list of length 7 that contains a data frame with the recursive coefficients for each of the seven bearings. `SIGMA_ALL` is a list of lists of data frames. This requires some explanation. For each bearing in each iteration the coefficient estimates are recursively estimated. By re-estimating the coefficient vector, the covariance matrix also needs to be re-estimated. Therefore, we obtain a list of covariance matrices for each bearing. In the end, these lists are combined to one large list `SIGMA_ALL`.

```
Recursive.Coefficients <- function(data, mon_per, m, formula){
    data = subset(data, DRAAI.UREN >= mon_per[[1]] & DRAAI.UREN <= mon_per[[2]])
    y = grep("^TEMPERATUUR.HOOFDLAGER", colnames(data))
    Y = data[,y]
    X = data[,-y]
    X = X[,-ncol(X)]
    ONES = rep(1, nrow(X))
    X = as.matrix(cbind(ONES, X))

    if(formula == 'transit'){
        formula = formula_transit
    }
    else if(formula == 'manoeuvre'){
        formula = formula_manoeuvre
    }
    else if(formula == 'combined'){
        formula = formula_combined
    }

    n = nrow(data)
    p = nrow(X)
    RECCOEF = list()
    sigma = matrix(rep(0, 7*nrow(data)), nrow(data), 7)
    SIGMA = list()
    SIGMA_ALL = list()
    for(i in 1:7){
        delta = matrix(rep(0, 11*nrow(data)), 11, nrow(data))
        intercept = list('INTERCEPT')
        rownames(delta)  = append(intercept, colnames(df_transit_X))
        for(k in(m+1:(n-m))){
```

```
            model1 <- lm(paste(paste(colnames(Y)[[i]],"~"),formula), data[(1:k-1),])
            model2 <- lm(paste(paste(colnames(Y)[[i]],"~"),formula), data[(1:k),])
            delta[,k] = model2$coefficients - model1$coefficients
            sigma = cov(t(delta[,(m:k)]))
            SIGMA[[k]] = sigma
        }
        RECCOEF[[i]] = delta
        SIGMA_ALL[[i]] = SIGMA
    }
    return(list(RECCOEF, SIGMA_ALL))
}
```

E.5 Regression adjusted variables

E.5.1 Individual charts

The R function `Compute.Z.rec()` computes the Z-statistics recursively. The first input argument `bearing` is a string that indicates the bearing for which the Z-statistic should be computed. The second input argument `data` is the data frame that contains the recursive residuals of the bearing temperatures of all the bearings. The final input variable `m` is a scalar that represents the length of the initialization period. The output of this function is a data frame that contains the Z-statistics for the desired bearing.

```
Compute.Z.rec <- function(bearing, data, m){
    y = grep("^TEMPERATUUR.HOOFDLAGER", colnames(data))
    Y = data[,y]
    Ypred = setdiff(grep("^TEMPERATUUR.HOOFDLAGER", colnames(data)), grep(bearing, colnames(
    data)))
    Ypred = data[,Ypred]

    Z = rep(0, nrow(data))
    formula = paste(paste(bearing,"~"),paste(colnames(Ypred), collapse="+"))
    for(i in (m+1):nrow(data)){
        model = lm(as.formula(formula), data=data[(1:i),])
        betas = model$coefficients[-1]
        tau = summary(model)$sigma
        betas = as.matrix(betas)
        summation = 0
        for(j in (1:6)){
            sum = as.matrix(betas[j,]*Ypred[i,j])
            summation = summation + sum
        }
        y = grep(bearing, colnames(data))
        Z[i] = (data[i,y] - summation)/sqrt(tau)
    }
    return(Z)
}
```

The R function `Cusum.Statistics()` computes the CUSUM statistics of the Z-statistics. The first input argument `bearing` is a string that indicates for which bearing the CUSUM statistics will be computed. The second input argument `data` contains the Z-statistics of that bearing. The final input argument `k` is the parameter k that is used to compute the CUSUM statistic. The output of this function is a data frame that contains the CUSUM statistics for the desired bearing.

```
Cusum.Statistics <- function(bearing, data, k){
    Z = data[,bearing]

    Li_plus = matrix(rep(0,nrow(data)))
    Li_min = matrix(rep(0,nrow(data)))

    for(i in 2:nrow(data)){
        Li_plus[i] = max(0, Li_plus[(i-1),] + Z[i] - k)
        Li_min[i] = min(0, Li_min[(i-1),] + Z[i] + k)
    }

```

```
    CUSUM = data.frame(cbind(Li_plus, Li_min))
    CUSUM$DRAAI.UREN =data$DRAAI.UREN
    colnames(CUSUM) = c("Li_plus", "Li_min", "DRAAI.UREN")
    return(CUSUM)
}
```

The R function `CUSUM.H()` is used to determine the value of h that needs to be used to obtain the desired ARL_0 value given the value of k via simulations. The input arguments are the desired `ARL0`, the value `k`, the search interval in which the value for h is searched in `HL` and `HU`, the accuracy parameter `rho` and the maximum number of iterations `maxit`. The output of this function is the value of h that needs to be used for the control limits of the CUSUM charts.

```
CUSUM.H <- function(ARL0, k, HL, HU, rho, maxit){
    for(i in (1:maxit)){
        h = (HL + HU)/2
        approx_ARL0 = xcusum.arl(k, h, mu=0, sided='two')
        if(approx_ARL0 >= ARL0 - rho & approx_ARL0 <= ARL0 + rho){
            break
        }
        if(approx_ARL0 > ARL0){
            HU = (HU+HL)/2
        }
        if(approx_ARL0 < ARL0){
            HL = (HL+HU)/2
        }
    }
    return(h)
}
```

The R function `plot.CUSUM()` returns the CUSUM chart that is used to monitor the regression adjusted variables. The first input argument is the data frame `data`. This data frame contains the CUSUM statistics for one of the bearings and also the operating hours. The second input argument `h` is a scalar that represents the upper control limit of the CUSUM chart.

```
plot.CUSUM <- function(data, h){
    return(
    ggplot() +
    geom_line(data=data, aes(x=DRAAI.UREN, y=Li_plus)) +
    geom_line(data=data, aes(x=DRAAI.UREN, y=Li_min)) +
    geom_hline(yintercept=h, color="red") + geom_hline(yintercept=-h, color="red") +
    theme_bw() +
    theme(legend.position = 'none', axis.text.x = element_text(size=14), axis.text.y =
    element_text(size=14),
        axis.title.x = element_text(size=14), axis.title.y = element_text(size=14),
        panel.border = element_rect(colour = 'black', size=1, fill=NA)) +
    xlab('Operating hours') + ylab('CUSUM statistic'))
}
```

In the R function `RAV.individual.CUSUM()` all the above functions are combined. The input arguments speak for themselves. The idea is that a data frame that contains the bearing temperatures, the predictors and the operating hours is used as input. The output are then seven CUSUM control charts of the regression adjusted variables.

```
RAV.individual.CUSUM <- function(data, prop_mode, mon_period, m, k, ARL0){
res = Recursive.Residuals(data, mon_period, m, prop_mode)

Z = matrix(rep(0, 7*nrow(res[[1]])), nrow(res[[1]]), 7)
j = 1
for(i in c("TEMPERATUUR.HOOFDLAGER.1.HVD.BB", "TEMPERATUUR.HOOFDLAGER.2.HVD.BB", "TEMPERATUUR.
    HOOFDLAGER.3.HVD.BB", "TEMPERATUUR.HOOFDLAGER.4.HVD.BB", "TEMPERATUUR.HOOFDLAGER.5.HVD.BB
    ", "TEMPERATUUR.HOOFDLAGER.6.HVD.BB",
"TEMPERATUUR.HOOFDLAGER.7.HVD.BB")){
     Z[,j] = Compute.Z.rec(i, res[[1]], res[[3]]+10)[[1]]
    j = j + 1
}
Z = data.frame(Z)
```

```
colnames(Z) = c("Z1", "Z2", "Z3", "Z4", "Z5", "Z6", "Z7")

h = CUSUM.H(ARL0, k, 1, 20, 0.001, 100000)
for(i in 1:7){
    CUSUM = Cusum.Statistics(i, Z, res[[1]], k)
    print(plot.CUSUM(CUSUM, i, prop_mode, h))
    }
}
```

E.5.2 Grouped charts

There are two different grouped RAV charts: the MCZ chart and the ZNO chart. While we give the R functions for both charts, only the functions for the MCZ chart will be explained.

The R function `Compute.MCZ()` computes the MCZ statistic. The first input argument `Z` is the data frame that contains the Z-statistics for all bearings. The second input argument `k` is the parameter k as used in the CUSUM charts. The output of this function is a data frame that contains the grouped MCZ statistic.

```
Compute.MCZ <- function(Z, k){
    Li_max = matrix(rep(0,8*nrow(Z)), nrow(Z),8)
    for(i in 1:7){
        CUSUM = Cusum.Statistics(i, Z, k)
        Li = cbind(CUSUM$Li_plus, - CUSUM$Li_min)
        Li_max[,i] = apply(Li, 1, max)
    }
    Li_max[,8] = apply(Li_max[,1:7], 1, max)
    MCZ = data.frame(Li_max[,8])
    MCZ$DRAAI.UREN = Z$DRAAI.UREN
    colnames(MCZ) = c('Statistic', 'DRAAI.UREN')
    return(MCZ)
}
```

```
Compute.ZNO <- function(Z, k){
    Li_sum = matrix(rep(0,8*nrow(Z)), nrow(Z),8)
    for(i in 1:7){
        CUSUM = Cusum.Statistics(i, Z, k)
        Li_sum[,i] = (CUSUM$Li_plus + CUSUM$Li_min)^2
    }
    Li_sum[,8] = apply(Li_sum[,1:7], 1, sum)
    ZNO = data.frame(Li_sum[,8])
    ZNO$DRAAI.UREN = Z$DRAAI.UREN
    colnames(ZNO) = c('Statistic', 'DRAAI.UREN')
    return(ZNO)
}
```

The R function `Compute.ARL.MCZ()` performs simulations to determine the ARL_0 of the CUSUM chart of the MCZ statistic. This ARL_0 value depends on the correlations in `Z` and the parameters h and k of the CUSUM chart.

```
Compute.ARL.MCZ <- function(Z, k ,h, maxARL, maxIT){
    vector_RL = matrix(rep(0,maxIT))
    for(n in 1:maxIT){
        X = as.matrix(mvrnorm(maxARL, rep(0,7), cor(Z)))
        Li_plus = matrix(rep(0,maxARL),maxARL, 1)
        Li_min = matrix(rep(0,maxARL), maxARL, 1)

        Li_max = matrix(rep(0,8*maxARL), maxARL,8)
        for(j in 1:7){
            for(i in 2:maxARL){
                Li_plus[i] = max(0, Li_plus[(i-1),] + X[i,j] - k)
                Li_min[i] = min(0, Li_min[(i-1),] + X[i,j] + k)
            }
            Li = cbind(Li_plus, - Li_min)
            Li_max[,j] = apply(Li, 1, max)
        }
        Li_max[,8] = apply(Li_max[,1:7], 1, max)
```

```
18
19          RL = min(which(Li_max[,8] > h))
20          if(RL == Inf){
21              RL = maxARL
22          }
23
24          vector_RL[n] = RL
25      }
26      return(mean(vector_RL))
27  }
```

```
1   Compute.ARL.ZNO <- function(Z, k ,h, maxARL, maxIT){
2       vector_RL = matrix(rep(0,maxIT))
3       for(n in 1:maxIT){
4           X = as.matrix(mvrnorm(maxARL, rep(0,7), cor(Z)))
5           Li_plus = matrix(rep(0,maxARL),maxARL, 1)
6           Li_min = matrix(rep(0,maxARL), maxARL, 1)
7
8           Li_sum = matrix(rep(0,8*maxARL), maxARL,8)
9           for(j in 1:7){
10              for(i in 2:maxARL){
11                  Li_plus[i] = max(0, Li_plus[(i-1),] + X[i,j] - k)
12                  Li_min[i] = min(0, Li_min[(i-1),] + X[i,j] + k)
13              }
14              Li_sum[,j] = (Li_plus + Li_min)^2
15          }
16          Li_sum[,8] = apply(Li_sum[,1:7], 1, sum)
17
18          RL = min(which(Li_sum[,8] > h))
19          if(RL == Inf){
20              RL = maxARL
21          }
22
23          vector_RL[n] = RL
24      }
25      return(mean(vector_RL))
26  }
```

The R function `Compute.h.MCZ()` determines the value of h via simulations given the ARL_0 and the parameter k of the CUSUM chart. The value of h is searched in the interval $[\texttt{HL}, \texttt{HU}]$.

```
1   Compute.h.MCZ <- function(ARL0, k, HL, HU, rho, maxIT){
2           for(i in (1:maxIT)){
3           h = (HL + HU)/2
4           approx_ARL0 = suppressWarnings(Compute.ARL.MCZ(Z, k, h, 5*ARL0, 1000))
5           if(approx_ARL0 >= ARL0 - rho & approx_ARL0 <= ARL0 + rho){
6               break
7           }
8           if(approx_ARL0 > ARL0){
9               HU = (HU+HL)/2
10          }
11          if(approx_ARL0 < ARL0){
12              HL = (HL+HU)/2
13          }
14      }
15      return(h)
16  }
```

```
1   Compute.h.ZNO <- function(ARL0, k, HL, HU, rho, maxIT){
2           for(i in (1:maxIT)){
3           h = (HL + HU)/2
4           approx_ARL0 = suppressWarnings(Compute.ARL.ZNO(Z, k, h, 5*ARL0, 1000))
5           if(approx_ARL0 >= ARL0 - rho & approx_ARL0 <= ARL0 + rho){
6               break
7           }
8           if(approx_ARL0 > ARL0){
9               HU = (HU+HL)/2
```

```
        }
        if(approx_ARL0 < ARL0){
            HL = (HL+HU)/2
        }
    }
    return(h)
}
```

In the R function `RAV.grouped()` all the above grouped functions are combined. The input arguments speak for themselves. The idea is that a data frame that contains the bearing temperatures, the predictors and the operating hours is used as input. The output are then two grouped CUSUM control charts of the regression adjusted variables.

```
RAV.grouped <- function(data, prop_mode, mon_period, m, k, ARL0){
    if(type == 'recursive'){
        res = Recursive.Residuals(data, mon_period, m, prop_mode)
    }

    Z = matrix(rep(0, 7*nrow(res[[1]])), nrow(res[[1]]), 7)
    j = 1
    for(i in c("TEMPERATUUR.HOOFDLAGER.1.HVD.BB", "TEMPERATUUR.HOOFDLAGER.2.HVD.BB", "
    TEMPERATUUR.HOOFDLAGER.3.HVD.BB",
    "TEMPERATUUR.HOOFDLAGER.4.HVD.BB", "TEMPERATUUR.HOOFDLAGER.5.HVD.BB", "TEMPERATUUR.
    HOOFDLAGER.6.HVD.BB",
    "TEMPERATUUR.HOOFDLAGER.7.HVD.BB")){
        Z[,j] = Compute.Z.rec(i, res[[1]], 2*m)[[1]]
        j = j + 1
    }
    Z = data.frame(Z)
    colnames(Z) = c("Z1", "Z2", "Z3", "Z4", "Z5", "Z6", "Z7")

    h.MCZ = as.numeric(subset(H.MCZ, ARL=ARL0, select = toString(k)))
    h.ZNO = as.numeric(subset(H.ZNO, ARL=ARL0, select = toString(k)))
    MCZ = Compute.MCZ(Z, res[[1]], k)
    ZNO = Compute.ZNO(Z, res[[1]], k)
    print(plot.CUSUM.grouped(MCZ, h.MCZ))
    print(plot.CUSUM.grouped(ZNO, h.ZNO))
}
```