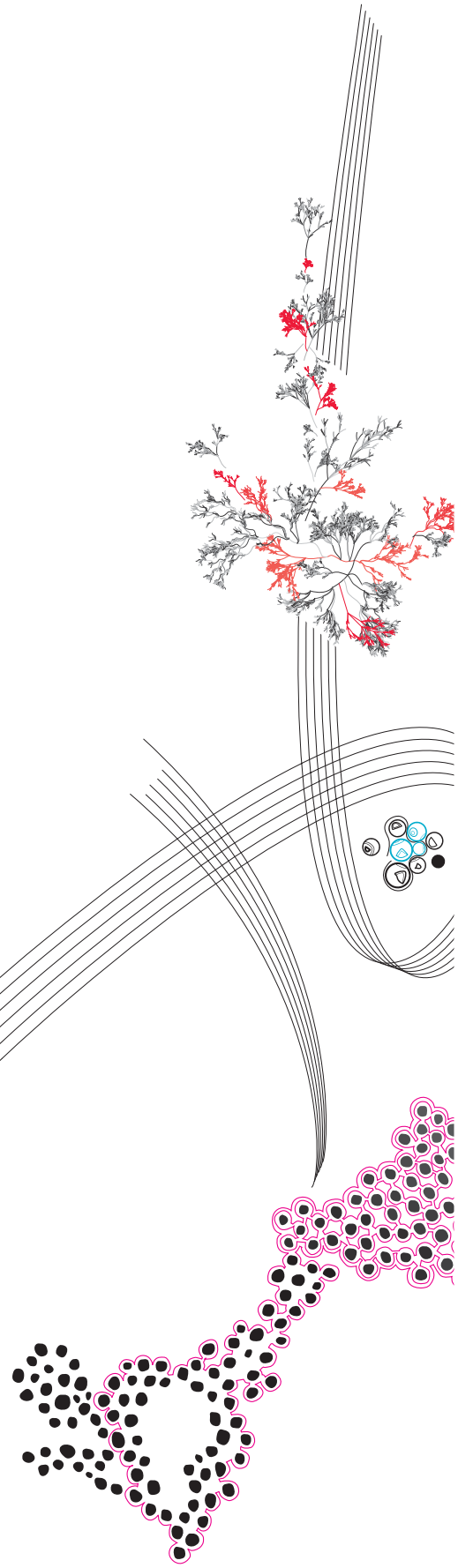


MSc Maintenance engineering and operations
Thesis report

Using Granger Causality to generate early warnings from Patrol Vessel platform sensor data

Jasper Wieringa

Supervisors:
prof. dr. ir. T. Tinga
ir. E. L. Wictor

August, 2024

Commando Materieel en IT
Ministerie van Defensie
Herculeslaan 1
3584 AB, Utrecht
The Netherlands

Faculty of Engineering Technology
University of Twente
Drienerlolaan 5
7500 AE, Enschede
The Netherlands

UNIVERSITY OF TWENTE.

Contents

1 Introduction **1**
- 1.1 The Royal Netherlands navy . . . 1
- 1.2 Motivation . . . 1
- 1.3 Involved bodies . . . 2
- 1.4 Goal of the Research . . . 3
- 1.5 Report organization . . . 3

2 Literature Study **4**
- 2.1 Maintenance strategies and smart maintenance . . . 4
  - 2.1.1 Preventive maintenance (PM) . . . 4
  - 2.1.2 Condition-Based maintenance (CBM) . . . 4
  - 2.1.3 Predictive maintenance (PdM) . . . 5
  - 2.1.4 Smart maintenance (SM) . . . 5
  - 2.1.5 Applications in the maritime field . . . 6
- 2.2 Granger causality and other data analysis techniques . . . 7
  - 2.2.1 Causality, Granger causality and correlation . . . 7
  - 2.2.2 Granger causality . . . 7
  - 2.2.3 Testing Granger causality: F-tests . . . 9
  - 2.2.4 Additions to the Granger causality method . . . 12
  - 2.2.5 Applications of Granger causality . . . 14
  - 2.2.6 SCADA alarm analysis . . . 16
  - 2.2.7 Outlier detection using a variable window . . . 16
  - 2.2.8 Data correlation and dimensionality reduction . . . 18

3 Selection of testing data **20**
- 3.1 Data acquisition on the OPV . . . 20
  - 3.1.1 IPMS . . . 20
  - 3.1.2 SAP system . . . 20
  - 3.1.3 DINO . . . 21
- 3.2 Selection Criteria . . . 21
  - 3.2.1 Corrective maintenance actions . . . 21
  - 3.2.2 Sensor count (S) . . . 22
  - 3.2.3 Sensor types (P) . . . 22
- 3.3 Method of selection . . . 22
- 3.4 Selection results and potential datasets . . . 23
  - 3.4.1 BSMI Rankings . . . 24
  - 3.4.2 Dataset selection . . . 25

4 Methodology **26**
4.1 Exploratory analysis . . . 26
4.1.1 Initial data-processing . . . 26
4.1.2 First in-dept application to find GC . . . 27
4.2 Further application of the method . . . 28
4.2.1 Application on larger datasets . . . 28
4.2.2 Fault detection . . . 28
4.3 Software and other data analysis techniques . . . 30
4.3.1 Software . . . 30
4.3.2 Other data analysis techniques . . . 30

5 Results **31**
5.1 Exploratory Analysis: BSMI 1214 . . . 31
5.1.1 BSMI 1214: Electric engine propulsion system . . . 31
5.1.2 Initial data processing . . . 31
5.1.3 Selection of in- and output parameter . . . 35
5.1.4 Application of Granger causality test of in- and output data . . . 35
5.1.5 Preliminary conclusions . . . 41
5.2 Fault detection using Granger causality . . . 43
5.2.1 Visual inspection of the dataset . . . 44
5.2.2 Application 1: BSMI 1214, set-point RPM t RPM . . . 45
5.2.3 Application 2: BSMI 1214, set-point RPM to POWER . . . 51
5.2.4 Similarities between undetected faults and 0 RPM errors in application 1 and 2 . . . 56
5.2.5 Application 3: BSMI 1214, RPM from/to POWER . . . 57
5.2.6 Linking SAP events to the results of the fault detection method . . . 59

6 Conclusions, Discussion & Recommendations **60**
6.1 Conclusions . . . 60
6.2 Discussion . . . 64
6.3 Recommendations . . . 65
6.3.1 Further research . . . 65
6.3.2 Recommendations for the RNLN . . . 66

A SAP events recorded at the RNLN **71**

B List of sensors in BSMI codes **73**

C Timelines of events in BSMI codes **77**

D All visually found potential faults **78**

List of Abbreviations

PC	Pearson's correlation
OPV	Oceangoing Patrol Vessel
RNLN	Royal Netherlands Navy
SM	Smart maintenance
CBM	Condition based maintenance
COMMIT	Commando Materieel en IT
DVO	Data voor onderhoud
GC	Granger causality
PdM	Predictive maintenance
PM	Preventive maintenance
DATs	Data analysis techniques
VAR	vector auto-regressive model
SAP	System Analysis Program Development
IPMS	Integrated Platform Management System
DMI	Directie Materiele Instandhouding
NLDA	Nederlandse Defensie Academie
BSMI	Basis Standaard Materieel Indeling
RPM	Rotations Per Minute

Abstract

The Royal Netherlands Navy (RNLN), will undergo various replacement programs in the coming 20 years. To improve on availability of the systems on-board the need for more sophisticated maintenance systems has arisen. With the goal of creating smart maintenance (SM) on the ships, a method called Granger causality (GC) was proposed. The method, originating from econometrics, aims to evaluate the causality between two parameters. Datasets selected from archived sensor readings of the Oceangoing Patrol Vessels in the RNLN are used for the analysis. In selected use-cases, in- and output parameters were identified for the application of the method. Causality between in- and output parameters could then confirmed in a normal working condition of a propulsion system. With the assumption that GC is lost during malfunction a detection system was proposed and applied on a larger dataset. Applications showed that with a performance expressed in a $F1$-score ranging between 0.88 and 0.96, visually identified malfunctions could be detected. The thesis therefore proved the feasibility of the application of GC in a naval setting.

Keywords: Predictive maintenance, Granger causality (GC), bivariate regression model, F-test, data manipulation, fault detection, F1-scores

Chapter 1

Introduction

This chapter will provide the necessary background information on the research topic. First, the status of the Royal Netherlands Navy (RNLN) and info on the ships will be given. Then, the relevance of this research is pointed and linked to the greater renewal project of the RNLN. Next, the goal of the research is stated together with sub-questions. Finally, the organization of the report is elaborated upon.

1.1 The Royal Netherlands navy

The RNLN is expected to renew most of its current fleet in the coming 20 years [28]. A part of this transition is the renewal of the so-called Oceangoing Patrol Vessel (OPV). By the late 2030's these OPV's will be phased out in favour of a more modern model. In total the Netherlands operates 4 of these vessels. In figure 1.1 the first of its class, HNLMS Holland is shown. The ship was commissioned in 2012, by the end of 2013 the last of its class, HNLMS Groningen was put to service [4].
The OPV has a wide range of tasks: Next to coastal guard duties it is also built for anti-piracy operations, counter-narcotics operations and is also capable of granting humanitarian aid where needed. Most notably the OPV have joined European anti-piracy missions in 2015 and have aided Haiti and Sint Maarten on multiple occasions during the aftermath of hurricanes in the area [4].

The Holland class is operated by roughly 50 crew members, has high-tech sensor equipment and has received midlife upgrades in 2020. These upgrades include both soft- and hardware changes [3]. As stated before, the upgrades will ensure the class's availability until the late 2030's until it is phased out.

1.2 Motivation

The modernisation efforts of the RNLN create the opportunity to learn from the current fleet and implement these lessons in the new ships. A part of these lessons is the aspect of implementing Smart maintenance (SM) technologies into the next generation OPV. SM is defined as a collection of techniques used to predict maintenance actions on board, preferably with real-time monitoring. The ultimate goal of any of these improvements will be the increased availability of the ships. Moreover, SM will also be able to improve safety onboard, lighten logistical burdens and thus save costs [5]. Below in figure 1.2 the method for achieving the SM goals is shown: In this example, data gathered from ships is analysed and based on trend-detection warnings are given leading to maintenance actions. This

FIGURE 1.1: HNLMS Holland at sea [4]

method can broadly be applied for all sensor data. Although, it should be noted that this is an example and in reality different techniques may be used instead of trend-detection.

1.3 Involved bodies

The research has been carried out at the Commando Materieel en IT (COMMIT) in Utrecht, as well as in collaboration with the group Data voor onderhoud (DVO) in Den Helder.
The DVO group is run by the over-arching Directie Materiele Instandhouding (DMI). This body, with the use of the DVO group, has set out this goal for SM.

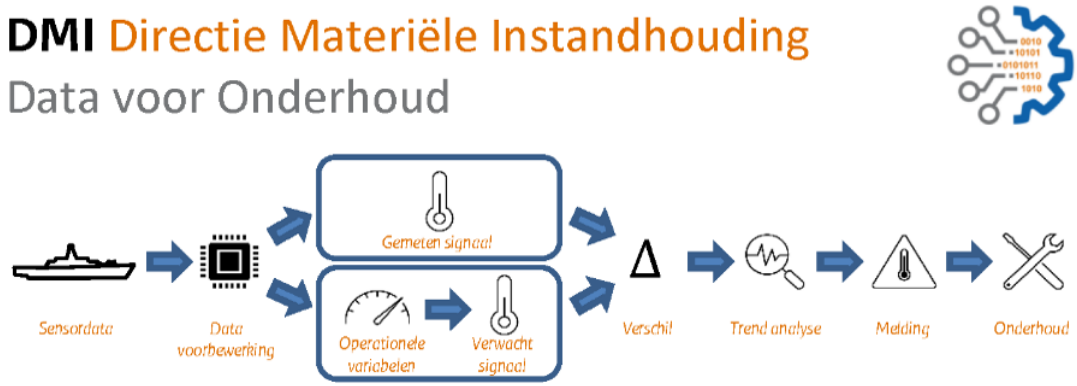


FIGURE 1.2: A scheme showing the perceived goal of using data for maintenance starting from sensor readings on the ship to eventual maintenance actions following from Data analysis techniques (DATs) such as trend detection [5].

1.4 Goal of the Research

As part of these Innovations at the RNLN this research will explore a possible use of available data currently stored at databases. With this data, new SM techniques may be introduced on the next generation of ships.

A recent study conducted in Canada showed the possibility of achieving accurate Condition based maintenance (CBM) using a method called Granger causality, a method created to prove causality between parameters [24]. As a consequence, a feasibility study will be conducted to explore the use of this technique in the RNLN

This goal thus aims to answer the following main research question:

To what extent can Granger causality be applied on available data from current Oceangoing patrol vessels at the Royal Netherlands Navy with the goal of implementing smart maintenance

Five sub-questions are also defined to answer the main research question:

1. *What data is currently gathered and stored on the Oceangoing patrol vessels?*
2. *How can possible use-cases to test these techniques be selected on the Oceangoing patrol vessels based on their maintenance history and sensor data availability?*
3. *How is the Granger causality test applied in this data environment and what variables are involved in its application?*
4. *What requirements have to be met in a dataset in order to successfully conduct a Granger causality test?*
5. *With what performance can faults in system be detected using Granger causality?*

1.5 Report organization

In chapter 2 the findings of a literature study are presented regarding relevant studies and elaborate on the theory behind Granger causality (GC).

Then, in chapter 3 more background information will be given on the OPV which will result in answering the first and second sub-questions with a proposal of a suitable method for the use-case selection.

DATs found in 2 are used to formulate a method of testing the selected data in chapter 4. Here sub-question three is addressed.

Using the proposed methods the application is made in chapter 5 showcasing the results. Here sub-question four and five are answered.

The report is concluded by listing its findings, shortcomings and recommendations for further research in chapter 6.

Chapter 2

Literature Study

The aim of the literature study is to acquire background knowledge to successfully research the topic at hand. Firstly, a deeper dive into the different maintenance strategies will point out which types of maintenance styles there are and which ones are addressed in this paper. In addition to this, research implementing similar maintenance techniques in the same field are discussed. Secondly, GC is discussed with its potentials and limitations. Lastly, other techniques are discussed that could be relevant to the research.

2.1 Maintenance strategies and smart maintenance

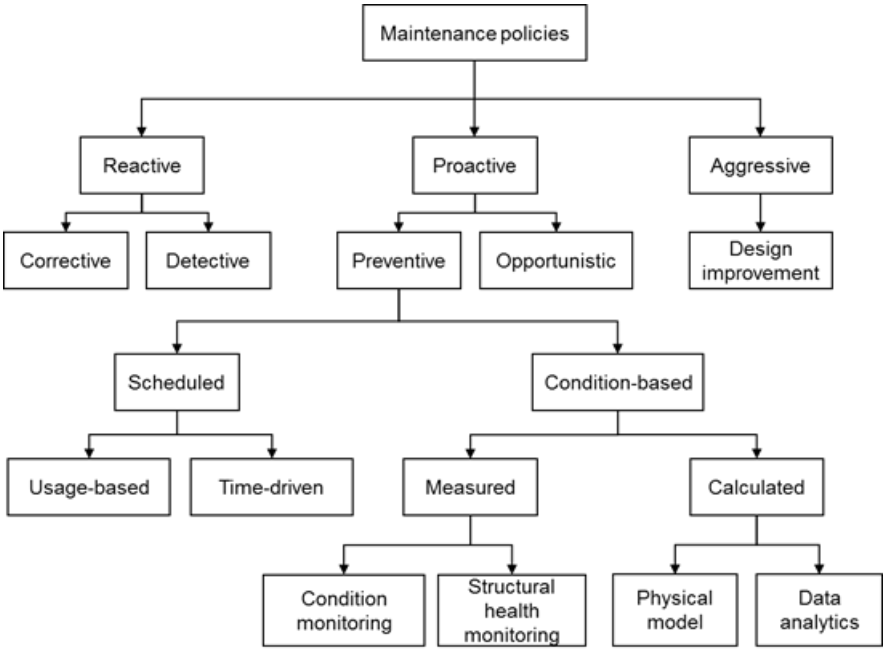
Generally speaking there are two main philosophies when talking about maintenance [21]. There are preventive and corrective styles of maintenance. Here, corrective maintenance means that maintenance actions are only undertaken after a failure in the system has already occurred. Basically, a run-to-failure strategy. On the other hand, there is the more active approach, which is the preventive or Preventive maintenance (PM). Here, there are a number of different implementations. Currently, the RNLN employs a combination of multiple strategies depending numerous factors such as systems' costs, availability target, etc. As a side note, during the design phase of a system parts can also be designed out in any way. However, this aggressive style of maintenance policy is not in scope of the research.

2.1.1 Preventive maintenance (PM)

Figure 2.1 shows the main groups of maintenance strategies. On the preventive side of maintenance there is the Scheduled and condition-based styles. The more advanced of the two is CBM. The styles of maintenance can be specified further. This research will focus on a Predictive maintenance (PdM) system. Here, condition monitoring is applied with the use of either sensors or visual inspections.

2.1.2 Condition-Based maintenance (CBM)

As the name implies, CBM has the aim of predicting maintenance actions as the target system reaches a pre-determined condition. The overall advantage of this type of maintenance is that availability is kept as high as possible while repair costs are kept low; repairs are carried out before costly failures occur and parts are not prematurely replaced [21]. Figure 2.2 shows the potential ways of detection of failure. The definition of what is preventive, predictive or reactive is not set in stone such that definitions may vary across organisations of what they deem to be preventive or predictive [8]. Generally, increasing the complexity



Rea

FIGURE 2.1: The main classification of maintenance strategies [27]

of DATs results in better capabilities.

Figure [27] further defines two groups within the CBM strategies: **Measured CBM** strategies are based on measured sensor data. As stated, the conditions are set and maintenance actions are based directly on the data [27]. **Calculated CBM** are more advanced using physical models based on a failure mechanism or data models based on historical data [27].

2.1.3 Predictive maintenance (PdM)

When an failure is detected using a determined condition, the next level in maintenance is the application of PdM. Here, an prognosis is made on the remaining time before failure [12]. In turn, maintenance actions could be postponed or executed earlier based on this remaining time. The prognosis can be based on analysis of earlier failure data or maintenance history of similar products [34].

2.1.4 Smart maintenance (SM)

As stated, this research will focus on a condition-based form of maintenance with the use of sensor data onboard similar to the early warning system highlighted in figure 2.2. Implementing this technique will, over time, lead to SM in the RNLN. In the future, a prognosis on systems may be added to the maintenance system as well. Finally prescriptive maintenance techniques give the user indications of where in the system problems are detected, and what actions are needed. This, in turn, will reduce downtime even further [17].

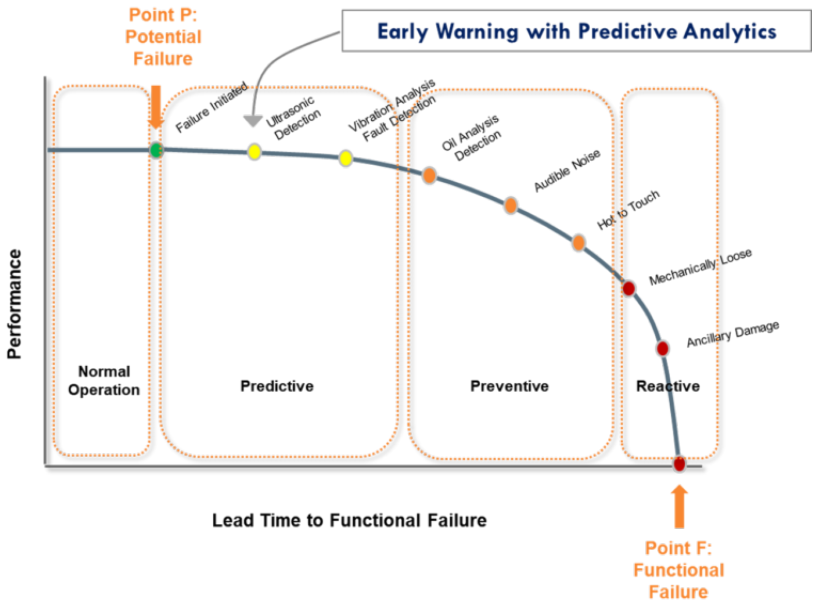


FIGURE 2.2: A P-F curve showing potential detection techniques over time from normal operation to failure [8]

2.1.5 Applications in the maritime field

In the Netherlands Defence Academy

In a naval application exploratory research was carried out by the Nederlandse Defensie Academie (NLDA) showing the potential benefits of CBM [19]. Sensor data is used in combination with pre-determined thresholds to give alarms leading to maintenance actions. The example also shows how, in further research, trends could be identified enabling even earlier detection of failures [19]. Figure 2.3 shows a clear trend leading to a sharp increase in, in this case, registered pressure difference is a pump setup. While both the implementation of an alarm and the detection of trends data could be considered SM, the trend detection is much stronger as problems are detected earlier [19].

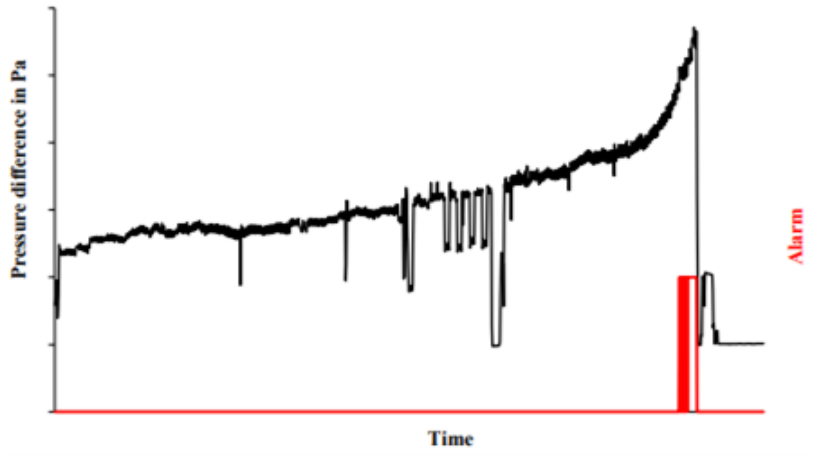


FIGURE 2.3: An example of CBM using an alarm [19].

In commercial maritime applications

A study put forward CBM techniques applied on a seismic survey ship. The ship, used to locate sub-sea pockets of oil and gas, was outfitted with sensors for the research [12]. With the use of DATs such as a correlation analysis and outlier detection a CBM model was created [12]. In the paper, failure modes were stated. In one case, contamination of oil for engine components was found to be an indicator of failure. Increased contamination levels led to warnings. The result is very similar to the one put forward by the NLDA.

Fatigue damage monitoring is applied using a digital twin achieving PdM. A freight ship was modeled with the aim of monitoring hull damage. Positional and environmental data such as GPS, temperature and wave height were acquired from the ship and used as input in the model. The model was then able to calculate fatigue damage to the ships hull. Using visual inspections, the model is updated with the eventual goal of PdM where the model is able to predict the time until critical damage based on planned use of the ship [31].

2.2 Granger causality and other data analysis techniques

In this section a range of DATs will be discussed. Here, their possible use in the research will be put forward. As discussed the GC method is to be applied. The results may be improved if elements of other techniques are applied as well.

2.2.1 Causality, Granger causality and correlation

There is a distinct difference between a correlative and causal relationship of data. Where a causal relationship between two factors explains that one of them causes the other, a correlative relationship simply explains that the two factors move generally in the same direction [20].

Often, data is mistaken to be causal while often correlation is the case [36]. Explanations for this can vary: For instance, coincidence may result in an causal relationship where there is none [36]. Another reason is the wrong interpretation of highly-correlated data where an relation is assumed to be causal in one direction when the opposite is actually the case [36]. Most commonly though, two independent measures are caused by a single factor not in scope [36].

At the end, causality is a somewhat deceptive term as it implies that one thing causes another while in reality countless other reasons could explain the relation [20]. The concept of causality is not defined and can not fully be quantified. There are however, methods that aim to do so to an extent. Granger causality itself is one scientific way to look at, and quantify causal relationships [1].

2.2.2 Granger causality

The main focus of this thesis will be to apply the GC method on the data available. This in itself is not particularly what GC was meant for: In 1969 the method was initially proposed by Prof. Clive Granger within the field of econometrics [1]. Here, the goal was to prove causality between two or more terms. It is important to note that true casualty is not proven here, 'Granger causality' is merely the proof that term X has a predictive

capability towards a term Y [1]. In mathematical terms the relationships are described in equation 2.1a and 2.1b. Equation 2.1a shows that for term X_1 an approximation is made by looking at the previous value of X_1 and X_2 which is $t-j$. Here $j = 1$, or one step earlier in a time-series. The value E_1 is the error, or the number that should be added to the first two terms in order to reach the value X_1. If the addition of new terms such as X_2 or any term X_k results in a smaller error it means that the addition of the term has a predictive capability towards the term X_1 [1]. The equation 2.1b shows the relation in terms of X_2 [1]. These two equations depict what is called a bivariate auto-regressive model meaning that a model is constructed to predect a value in an vector based on earlier values of its own vector in addition to another vectors earlier values.

$$X_1 = \sum_{j=1}^{p} A_{11,j} X_1(t-j) + \sum_{j=1}^{p} A_{12,j} X_2(t-j) + E_1(t) \tag{2.1a}$$

$$X_2 = \sum_{j=1}^{p} A_{21,j} X_1(t-j) + \sum_{j=1}^{p} A_{22,j} X_2(t-j) + E_2(t) \tag{2.1b}$$

The equations now only include the first lag, or $j = 1$. While in reality the relation may be stronger with addition of further lags, or delays in the time-series. Within the time-series the choice can be made to include values that are two, three, or further points back. The results of the test will determine what added value these lags have. In practice it could mean that if one term increases the other term follows almost instantly. It could also be found that an incline seen in term X results in an incline in term Y only after a longer time.

Figure 2.4 shows an example where term X 'granger' causes term Y. Here the relation between the two seems to be that there is a 5 second delay from X to Y . The data is recorded at a $1Hz$ frequency. The results of the test will potentially show that $X \rightarrow Y$ with $j = 5$. This will the be case while (most likely) $X \not\rightarrow Y$ for $j = [1-4]$. The j value thus depends on the data but also the sampling frequency of the data. If the data in figure 2.4 was recorded at a frequency of $0.2Hz$ GC would have been found at the first lag, $j = 1$ [7]. It should be noted here that if causality is proven or rejected in one way it does not say anything about the reverse yet; when looking at figure 2.4 term Y it does not 'Granger' cause term X.

Regression modeling

Equations 2.1a and 2.1b already showed how the determination of GC is set up broadly. In practice this means that in order to test GC two regression models have to be made. One of them is the before mentioned, bivariate auto-regressive model and one of them is an univariative one where only one time-series is included. Equations 2.2a and 2.2b show the equations.

$$X_t = c_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_j X_{t-j} + ... + e_t \tag{2.2a}$$

$$X_t = c_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_j X_{t-j} + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_j Y_{t-j} + ... + \epsilon_t \tag{2.2b}$$

Essentially, regression modeling has the aim of rewriting data into a function which best describes it [7], this is why an constant (c) is added to the model. In the univariate model (equation 2.2a) only past values of time-series X are used to describe the time-series, while the bivariate model (equation 2.2b) has the addition of the second parameter Y. The two

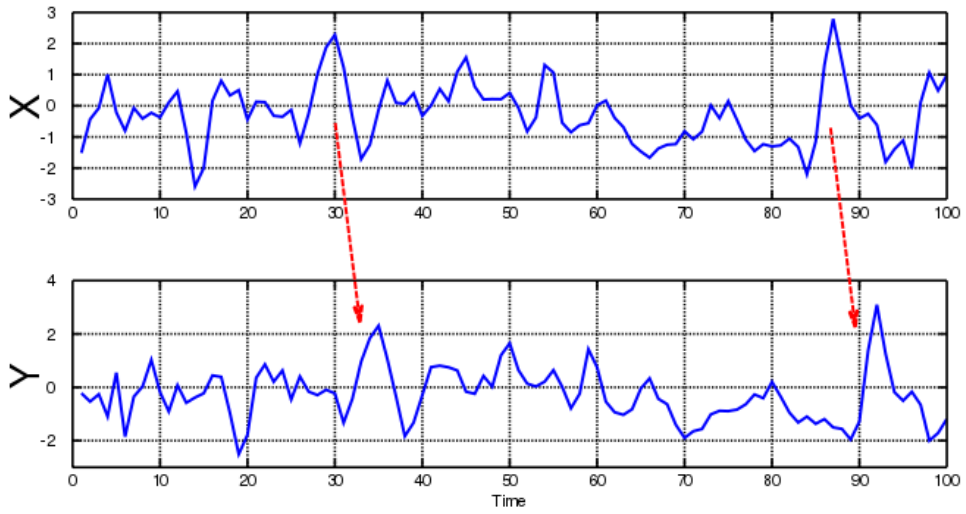


FIGURE 2.4: 2 graphs showing terms X and Y where X granger causes Y with an delay of 5 seconds [7]

equations result in two residual errors. To test GC the idea is that the addition of the Y terms will reduce the error that results from the model. Therefore, to prove GC, $e > \epsilon$ [13].

2.2.3 Testing Granger causality: F-tests

Acceptance and rejection of Granger causality

When determining causality such as in the example shown in figure 2.4 so-called F-tests have to be executed for every lag chosen [13]. The F-tests are based on a null-hypothesis where the hypothesis is as follows [13]:

The exclusion of the tested lag j in X strengthens the approximation of the next value in time-series Y.

The hypothesis can be accepted (H_0) or rejected (H_1) looking at equation 2.2b this would mean the following:

1. H_0= $\beta_1 = \beta_2 = \beta_j = 0$ will accept the hypothesis meaning that a no degree of a delay in Y decreases the error, there is no GC present.
2. H_1= $\beta_1 = \beta_2 = \beta_j \neq 0$ will lead to the rejection of the hypothesis meaning that a degree of a delay in Y decreases the error. GC is present.

As is standard with a null-hypothesis, the result of the test is a probability measure which measures the chance that the hypothesis is correct. A probability lower than 5% is considered a threshold where the hypothesis is rejected [10]. This means that if the result of the test is lower than 5% or 0.05 this hypothesis can not be accepted and thus this lag could contribute to GC of the time-series. If the test results in values higher than than the threshold, higher than 0.05, the null-hypothesis is accepted and the exclusion of the lag does indeed strengthens the approximation. In other words, no causal relationship is present between the parameters. The threshold for acceptance of the test can be altered

based on the application [10]. If, for instance, greater certainty is needed it may be lowered. The downside to this is that a larger sample size is needed [10]. A commonly used alternative to the 5% level is a 0.01 or 1% one [10].

When applied to lags $j = [1 - 5]$ in the time-series in figure 2.4, results could look like the following: $P > 0.05$ in lag $j = [1 - 4]$ thus accepting the null-hypothesis and implying no causality. At lag $j = 5$, $P < 0.05$ will reject the null-hypothesis implying a possible causal relationship.

In order to accept the P value another result is needed from the F-test. This is the F-test statistic, or F-statistic [13]. marked by F and is the actual measure to describe the strength of GC. Moreover, it can also be used to validate the resulting P values. Generally speaking, F should be higher than P in order to accept the F-test results [13]. F is determined using equation 2.3 where:

1. T marks the sample size
2. e and ϵ are the sum of residual squares $E(t)$ before and after inclusion of the lag [13]. They are therefore the resulting errors of the univariate vs the bivariate model. $E(t)$ is the same error as previously described in equations 2.2a and 2.2b.
3. p marks the degrees of freedom in the regression model, this is based on the amount of lags taken into account. $p = j + 1$

The F-test statistic gives a a measure of improvement of the bivariate model against the univariate model. If the error is reduced more greatly by adding a lag of another time-series the F value should become larger. A very low F-statistic could indicate that although there might be an Granger causal relationship this could also be the result of noise in data, occasional spikes in data, high variance, et cetera [13].

$$F = \frac{(e - \epsilon)/p}{\epsilon/(T - 2p - 1)} \tag{2.3}$$

To sum up, two values result from F-testing:

1. P-value: The P-value is the probability value resulting from the F-test. It is always valued between 0 and 1 as it describes the probability that the hypothesis is true. The default threshold for acceptance of a null-hypothesis is 0.05 however, it can be altered.
2. F-statistic: To verify that the P-value is legitimate the F-statistic is calculated and should always result in an higher value than the P-value. It is based on the size and content of the data and gives a measure of improvement of the bivariate over the univariate regression model.

Interpretation of F-test results

It is important to interpret the results of the F-test correctly, Therefore, before conducting the test, a determination of what relationship is assumed between the to-be analysed parameters must be established [16]. If tests validate the assumed relationship the analysis will be assumed to be correct. If, on the other hand, test results differ from the assumption, the following two main problems may be the cause of this:

1. Problems with the frequency rate of data. This can lead to instant causality where a Granger causal relationship is concluded in both directions. The frequency of data may be incorrect for the application. It could be too high for a dataset where changes are only happening over very long periods of time. Also, frequency could be too low resulting in overall data being too scattered [16].

2. Both terms could also be influenced by a third term [16]. This is called common cause fallacy or a confounding factor and may cause test results to be distorted.

The result of these problems could for instance be Bidirectional GC which means that in both directions GC is concluded and validated with the F-statistic. The before mentioned problems could be the cause of this [16].

A good method of validating the test is by conducting it the other way around. If the assumption is that parameter A Granger causes B it can be validated by testing if B Granger causes A [16]. The result of this test should then show the opposite of the earlier result.

Probabilistic evaluation of F-test results

F-tests are based on probability basis using a null-hypothesis. They can, therefore, be evaluated using a group of metrics. Namely: accuracy, precision, recall and the F1-score [35]. When the F-tests are conducted in groups the overall capabilities of the test can be determined using these metrics. F-test results can be grouped in four groups:

1. TP (True Positive). These are situations where the F-test result deemed to be positive, accepting the null-hypothesis and the data also shows visually that this is the case. Both show that no GC is present [35].

2. TN (True Negative). Here both the F-test result and the actual condition reject the null-hypothesis. Both show that GC is present [35].

3. FP (False Positive). The F-test accepts the null-hypothesis while the actual condition rejects it. In the GC environment this means that the F-test find that there is no GC while the data clearly shows visually that causality is present [35].

4. FN (False Negative). F-testing rejects the null-hypothesis showing that GC is present while upon inspection the two parameters are clearly do not show causality [35].

Both the TP and TN show correct results of the F-test while the FP an FN could be considered mistakes. Figure 2.5 shows the confusion matrix which is built from the four numbers. The accuracy is calculated using equation 2.4. All values are used here. The total amount of true accounts are simply divided by the total amount of accounts, both true and false. Accuracy ranges from 0 to 1, or 0% to 100% with 1 being ideal[35].

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.4}$$

Evaluating accuracy only does not always suffice. For instance, the model could be unbalanced. Unbalance arises when there is a large discrepancy between True positives and negatives or False positive and negatives [35]. In this case the model can be evaluated

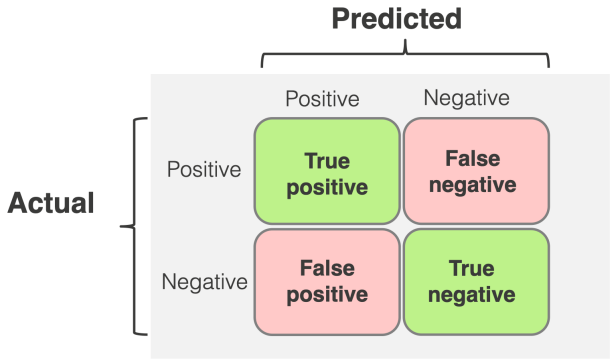


FIGURE 2.5: A figure depicting the confusion matrix used for accuracy calculations [6].

too using the Precision and Recall metrics. These are shown in equations 2.5 and 2.6 [35]. Precision (positive predictive value) determines the share of successfully predicted positives. It thus shows how successful the model is in accepting the null-hypothesis. In the GC environment it determines how good the model is in showing that no GC is present.

The recall (true positive rate) metric does the same but now focuses on the capability of showing that GC is present. Both again range from 0 to 1 (0% to 100%) with 1 being most favourable [6].

$$Precision = \frac{TP}{TP + FP} \tag{2.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.6}$$

Finally, the F1-score uses both the precision and recall to more completely describe the performance of the test [35]. It is shown in equation 2.7.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.7}$$

2.2.4 Additions to the Granger causality method

Addition of parameters

In these examples there are only 2 parameters that are analysed at one time. However, using this method more terms could be added. The current monitoring systems on the OPV class include numerous parameters which could all be added in this method making it very suitable to see the relationships between sensor values [22].

In more advanced models GC may be defined using more than one parameter or sampling rate. In practice this means that parameter A could be defined by parameters B,C,D etc. In essence, a network is created to explain the relationships between all parameters [22].

In figure 2.6 the influence of a third parameter on two parameters is shown. In this model the addition of a fraction of parameter X_1, expressed in C ranging from 0 to 1 determines the causal relationship between X_2 and X_3. The relation is Granger casual from $X_2 \rightarrow X_3$ when $C = 0.7$ while this is the other way around when X_1 is left out of the model entirely, when $C = 0$.

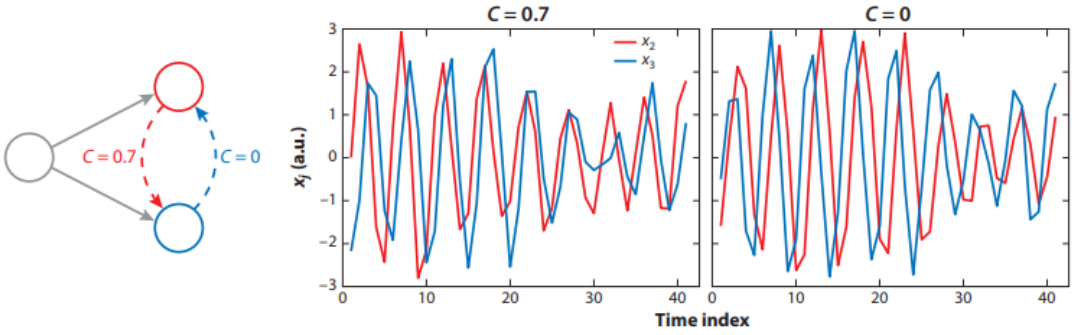


FIGURE 2.6: Example showing the results of the vector auto-regressive model (VAR) model where GC is stated in both directions ($x_2 \rightarrow x_3$ and back) depending on a degree of X_1 expressed in C [22]

Lag selection

Not only the increase of parameters but also the addition or selection of lags may be elaborated upon. As stated earlier the relationship between parameters may be non-existent when looking in short term. However adding lags, to the computation could show GC in a longer term. The selection of lags is therefore a vital step in understanding the relationship between parameters [22].

Two choices may be made when selecting GC analysis. The first one involved the employment of a lag selection criterion [15]. Multiple methods exist in this regard varying in complexity. All of them quantify the maximum amount of lags relevant for the dataset [15]. Here, length of the sample, variance of data and other metrics influence the result. Most commonly used are the AIC, or Akaike Information Criterion [15] and the SIC, or Schwarz Information Criterion [15].

After the implementation of the criterion, testing of the data will start limited to the maximum found lag. Equations 2.8 and 2.9 show the definitions of the two criteria. In both formulas k stands for the amount of parameters, or lags involved and the value $\hat{L}$ is the likelihood or likelihood function. This is the product of likelihoods of all values in the model to fit in the regression model [15]. Therefore it is a measure of strength of the model. The SIC method also uses the n or total amount of data points in the set.

$$AIC = 2k - 2ln(\hat{L}) \tag{2.8}$$

$$SIC = k\,ln(n) - 2ln(\hat{L}) \tag{2.9}$$

The AIC and SIC methods result in a penalty value for every lag added to the test. The lower the penalty value the better the lag selection. Since adding a lag adds complexity, the value k becomes larger and larger. However, adding lags could substantially increase the likelihood function as the model becomes better in fitting data. The methods result in a balance where, the lowest resulting penalty indicates the maximum complexity or lag used for the application of GC.

A study done on Malaysian companies aimed to find a relationship between market share performance and the increase of sustainability practices (environmental, social, etc.). The data was fitted in regression model where market share was fitted onto perceived increase of sustainability practices [29]. A lag selection was done using both AIC and SIC. Table 2.1 shows the results of the various tests. Annually, units of lag (time) where defined

and tested using these criteria. From 2017 to 2020 the minimum number, or minimum penalty is selected [29]. These values are marked by an asterisk and are the same for both methods. Data was stored over multiple years. For every year the calculation was redone since quality of data differed between the years [29]. In 2017 using a maximum of three lags is most appropriate while in the other years four lags is found to be better. The main benefits here are reduction of computing time as lags outside the criterion are left out. Another benefit is the avoidance of misinterpretation of GC test results. If lags outside the criterion are tested, they are statistically illegitimate [26].

ARDL Lag Selection

Year	Lag 4		Lag 3		Lag 2		Lag 1	
	AIC	SIC	AIC	SIC	AIC	SIC	AIC	SIC
2017	−6.08	−6.64	−6.14 *	−6.69 *	−6.07	−6.41	−5.99	−6.23
2018	−6.45 *	−7.06 *	−6.37	−6.85	−6.26	−6.62	−6.29	6.54
2019	−6.25 *	−6.82 *	−6.23	−6.71	−6.19	−6.53	−6.12	−6.35
2020	−6.19 *	−6.76 *	−6.17	−6.62	−6.16	−6.5	−6.1	−6.36

* Optimum lag.

TABLE 2.1: Results of various lag selection criteria resulting in a maximum lag of three to four [29]

In the second method, the selection may also be based on the results of the tests where at first a large amount of lags are analysed. In this case, knowledge of the data and its origin are necessary as it will be used to set the maximum amount of lags and predict its possible outcomes [23]. Also, averaging the amount of useful delays may also be implemented [23]. This could, however, result in the loss of information and prevents further selection of data. In a new method lags are selected based on their individual GC test result [23].

Application of a LASSO-function

To aid lag-selection a LASSO (Least Absolute Shrinkage and Selection Operator) function may be applied. LASSO is an regression analysis method where the minimum error $E(t)$ is calculated using a number of lags. The lags are scaled in the function resulting in a list of most prominent lags. It essentially means that one lag may be left out entirely while others are added to a magnitude of 0.9 or 0.4 [38]. LASSO results in a listing of lags where some are found to be suppressive (magnitude 0) and others to be active in the relationship between two variables [38]. It can, thus, be viewed as a first line selection of time-lags as the magnitude determines the selection of the lags added. Active lags may be analysed further decreasing computational time in the following F-tests [38]. In total, the results of the LASSO function do not say anything about the possible GC yet, LASSO simply estimates which lags could contribute numerically to the relationship between two parameters [38].

2.2.5 Applications of Granger causality

The overall theory behind GC is relatively straight-forward. This means that applications are not limited merely to the field of econometrics. Some possible applications, including a naval one, are discussed below.

In bio-informatics

The development of diseases is characterised by the change in density of certain genes. Relationships may be discovered to further understand and identify the progression over time [23]. Using GC progression can be identified over time. The relation between two genes is found including a given time lag. For instance, the concentration of a gene can be identified to Granger cause the increase of concentration of another gene given a time-lag of two [23]. The relevant time-lags are found using a lasso-function [23].

In the naval field

Research conducted in Canada proves the potential use of GC in a naval application. Research conducted by the S & T Organisation involved the analysis of sensor data on-board Canadian Frigates at times of operational deficiencies [24]. Operational deficiencies, or OpDef's are collections of reports made by operators regarding a mechanical malfunction. An OpDef is opened at the detection of a problem and is closed when repairs are eventually concluded. In between, new developments or intermediate repairs are documented as well. In an earlier stage, indicator sensors were identified and linked to a list of these deficiencies [25].

Figure 2.7 shows a timeline with this information. Here, the red lines mark a number of OpDef's. The purple points mark new entries in the Opdef. During an operational deficiency the indicator sensors are read and analysed using GC. The exact dates of the GC analyses are marked in black. Results show that with an overall accuracy of about 99% sensory indicators are found for the deficiencies. The accuracy is calculated by dividing the number of 'true', T, test results by the total test results meaning that the number of 'false', F, alarms are only at 1% The accuracy is calculated using equation 2.4 [24].

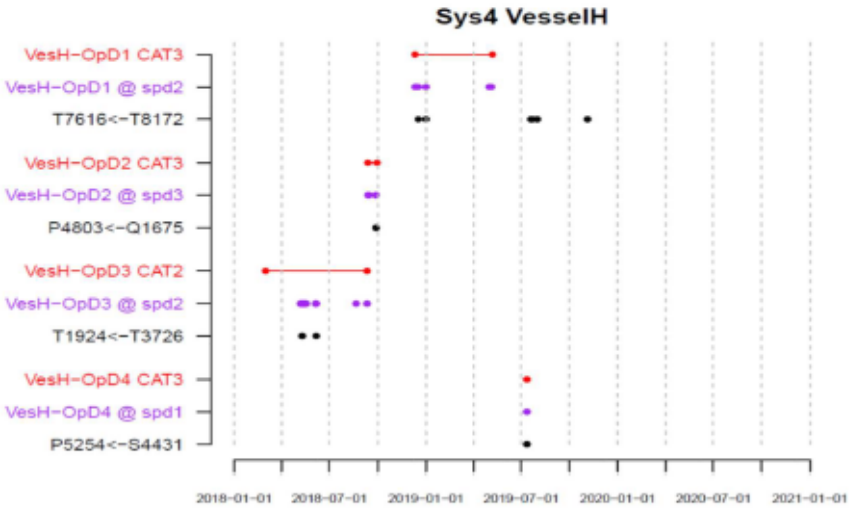


FIGURE 2.7: A timeline showing analysed Opdef's (red & purple) with their respective indicator sensors (black) [24].

As an example, the GC relationship from sensor T8172 to T7616 can with 99% accuracy predict deficiencies. In total the paper lists multiple sensors per deficiency resulting in, for instance, a list of 33 relationships that could predict a deficiency[24].

Sensor data was gathered using the onboard Integrated Platform Management System (IPMS) [24]. This system is present on the Dutch OPV as well. Similar to the Dutch system, data is logged at a $1Hz$ frequency. Data collection is found to be sufficiently comprehensive to enable CBM. A prior study to the one described here found that in a wide array of components, sensor parameters are present to create maintenance indicators [25]. Next to this, available data on maintenance history and scheduled maintenance tasks can be added to the sensor data to improve the overall picture a components health [14].

2.2.6 SCADA alarm analysis

In section 2.1 the addition of diagnostics of a maintenance system was touched upon. Using probability models and alarm documentation a model can be created which puts forward the most probable cause of alarms in a system. Research has shown that on complex systems, such as a wind turbine, the sequence of alarms can be combined with anomalies in sensor data to create a prescriptive maintenance system [18]. Figure 2.8 shows how a sequence of alarms lead to a diagnosis for the wind turbine system.

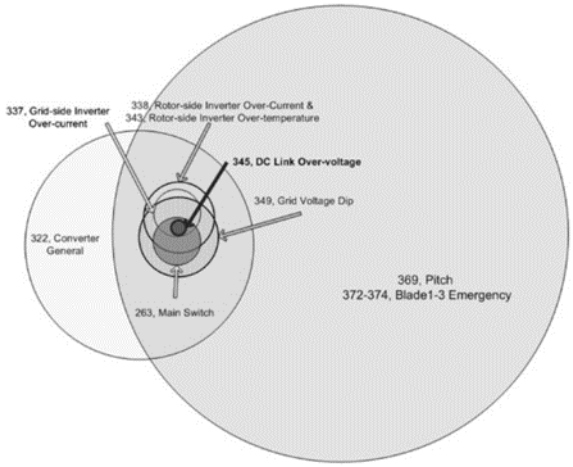


FIGURE 2.8: A VENN diagram showing the most probable cause of a defect in a Wind turbine system using alarms [18].

2.2.7 Outlier detection using a variable window

Another challenge in the detection of faults is the pre-selection of data, meaning, the partitioning of a larger dataset to more accurately define faults in the system [11]. An example of this is a study in the application of outlier detection. The study found that one of variables dictating the number of false positive or negative detections was the chosen window of data analysed at the time. Figure 2.9 shows that changing the chosen window of the analysis from 12 to 16 and eventually 20 hours improves the analysis as less false detections are made. The algorithm thus profited from larger pieces of data [11].

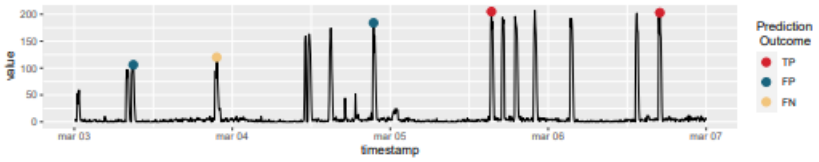
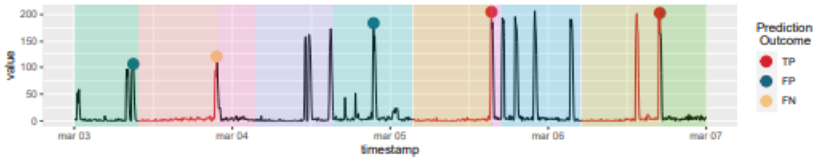
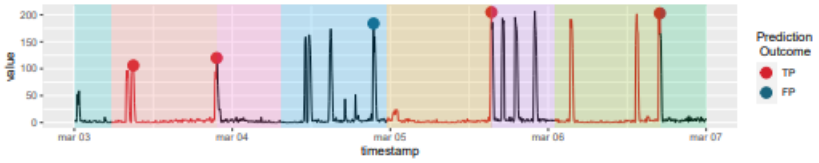


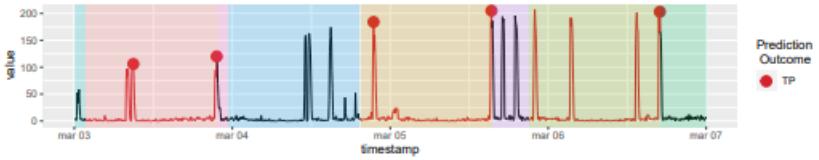
Figure 2: Anomaly scores and algorithm predictions



(a) 12-hours window



(b) 16-hours window



(c) 20-hours window

FIGURE 2.9: Multiple graphs showing the effect of the chosen data window on the amount of true positive, TP vs false positive or negative, FP & FN detections [11].

2.2.8 Data correlation and dimensionality reduction

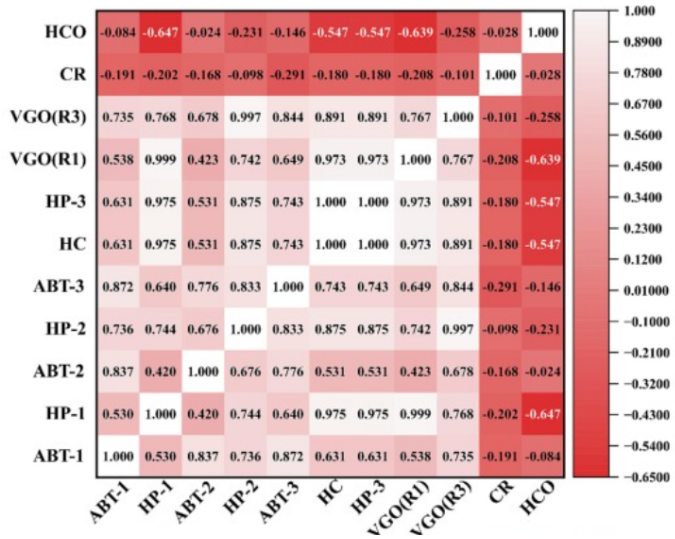
Data correlation

As mentioned before, the goal of calculating correlation is to conclude the strength and direction of relationship between two sets of data [9]. Most commonly, the Pearson's correlation (PC) method is used for this. As shown in equation 2.10, correlation between parameter x and y is expressed in r_{xy}. The result is based on the covariance and variance of the two variables [9]. The resulting value has a range of $[-1:1]$ where the magnitude indicates the strength of the relationship. The more negative the value, the stronger the negative relationship while the higher the value the more strongly positive the relationship. A value which is small, thus closer to zero indicates a weak relationship [9]. In contrary to GC calculations, correlation results is the same independent from the direction of calculation. Typically, values between 0 and 0.3 describe weak, 0.3 and 0.6 average, 0.6 and 0.9 strong and 1 perfect relationships [9]. The same goes for the negative values.

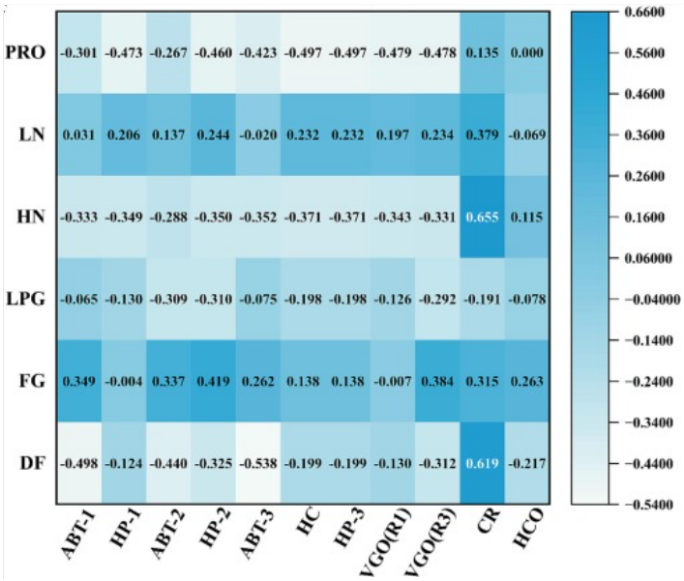
$$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}} \tag{2.10}$$

Dimensionality reduction

In complex systems multiple variables may correlate in such a high degree that they can be combined into one, or left out entirely [37]. In a paper where data-driven PM is achieved the first step is to simplify the analysis by mapping data in a correlation matrix. Figures 2.10a and 2.10b show such matrices. Here, all input parameters are correlated with each other in the first matrix. The second matrix shows the correlation with output parameters. Leaving out parameters based on their correlation with one-another decreases computing time in later steps of the analysis [37].The threshold for removing data is set to be 0.9 meaning that the input dimension is reduced from 11 to 7 variables [37].
Alongside the diagonal all values are 1 in figure 2.10a since the variables are correlated with themselves. The research paper uses the PC method for this. It should be noted that with the use of this method both values are equal opposite of one another. For instance, in figure 2.10a the resulting correlation value between HCO and $ABT-1$ is both the same on both sides of the diagonal.



(A) A correlation diagram where all input variable are correlated with each other [37]



(B) A correlation diagram showing the correlation between out- and input variable [37]

FIGURE 2.10: Correlation matrices used to select variables for further analysis [37]

Chapter 3

Selection of testing data

This chapter will explain the method of data selection for this research. As the volume of available data is large, a methodical approach to its selection is needed. To begin, the data acquisition systems at the OPV are explored. Using this knowledge, meaningful selection criteria are listed to find the most promising datasets for this research. Finally, the selection is made and the final datasets are listed.

3.1 Data acquisition on the OPV

In order to create a meaningful selection of data it is important to know how data at the OPV is collected, sorted and stored. In this section the two main systems used for this selection are elaborated upon.

3.1.1 IPMS

During operation all systems are managed using the IPMS. The system is used to monitor the function of components during operation but collected data may also be extracted for later analysis. This is the main source of data for this research.

BSMIs

As data is logged it is labeled using a coding system. This system is named the Basis Standaard Materieel Indeling (BSMI). The OPV has a large number of these codes. Everyone of these codes is linked to a part, or system in the ship. From propulsion systems to pumps or even the hull, all have their own code. To give a complete list is not particularly meaningful in itself but the BSMI coding system does allow necessary filtering later on.

Sensors

The number of available parameters is listed for each BSMI code. This is, for the purpose of this research, very useful. These can be divided into in- and outputs. Conventional sensors such as temperature, pressure or power sensors are all logged as outputs of the numerous systems onboard. Next to this, set-points, engine settings and running hours are also found.

3.1.2 SAP system

The System Analysis Program Development (SAP) system is used by the RNLN to manage all logistical flows of components, list failures (events) in systems and document other

business processes. The listing of failures is important to the analysis as it determines which dataset has the most potential to test GC on. The failures are manually documented onboard by operators and classified. Factors such as priority, type of failure and time are all listed.

Types of SAP events

At the RNLN all events are classified with a coding system. Generally, M events are related to instances where a single component is in need of preventive or corrective maintenance actions, WP events make up a grouping of actions to be undertaken at one time and Z events lead to alterations of the system. In reality, most events are classified as M events or part of WP events. The total list of definitions per type of event is listed in appendix A.

Priority numbers

Depending on component, safety or degree of failure a priority number is given to the event as well. The number will not be taken into account when selecting datasets. Mainly because of the various aspects it is build from. Damage done instantly due to human-made errors may end up having a much higher priority score than damage done over a longer period of time which is the type of damage most interesting for this study. Also, some components are far more important on the ship. This will always be given higher priority numbers distorting the selection further.

3.1.3 DINO

As of now , all IPMS data is saved onto hard drives onshore. Using an application on the defence network sensors can be selected for further analysis. When system codes are selected with a corresponding time frame the DINO app is used to download this data. The DINO app is created by the DVO group as part of their SM project.

3.2 Selection Criteria

To select useful data from the database criteria are defined. All systems onboard or BSMI codes, are subjected to these selection criteria to find the highest potential candidates. The criteria are the following:

3.2.1 Corrective maintenance actions

The final goal of this thesis is to verify whether GC could be implemented on sensor data to detect problems onboard. This means that for datasets to be useful they must include anomalies caused by failures. Looking at section 3.1.2 the type of events that would provide these anomalies are most likely linked to events leading to corrective maintenance actions. In M_1, M_2 and M_3 events this type of maintenance is applied. Systems, in this case, have experienced faults that need to be addressed earlier than planned. It is therefore assumed that these are the only events interesting for the analysis. M_3 events are left out in the final selection method as these are only administrative in nature.

3.2.2 Sensor count (S)

The second criterion is the number of sensors present in any given BSMI system code. Some of these systems will have only one or two sensors and are therefore not complex enough to analyse. Moreover, the chance is higher that no anomalies or events are found here.
The BSMI codes are therefore divided up into three main groups. For each group, the system with the highest potential will be picked first to analyse. The groups are the following;

Simple systems: These range from 2 to 9 parameters. When applying the to be determined data analysis method, testing it on a less complex system will be ideal to prove the principal theory.

Intermediate systems: Ranging from 10 to 24 parameters. Following the proof of concept in the simple group this stage will add complexity.

Complex systems: Finally, systems with 25+ parameters are selected. These are the most complex systems and will only be tested if the method is functioning well enough.

3.2.3 Sensor types (P)

Not only the total number of sensors is taken into account. To add to this, all selected BSMI codes need to have at least two different sensor types. Since some of these BSMI codes only comprise of less rich parameters these are selected out. As an example, one BSMI code only has three sensors all regarding running hours. Finding anomalies and linking them to each other will become very hard in this case.

The BSMI codes are selected based on three main aspects:

1. Number of corrective maintenance actions (M_1 and M_2)
2. Total number of sensors, in three groups (S)
3. Number of types of sensors (P)

3.3 Method of selection

The method of selection is listed below in three main steps. Figure 3.1 shows the method of selection. The entirety of the selection steps have been conducted using excel while information on SAP events and BSMI codes was found internally.

Step 1: Collection of data

The first step is simply to combine known data into a single excel file. All BSMI's are split into the three groups; simple, intermediate and complex. Their corresponding SAP event data is listed for all BSMI codes as well. These are split into the M_1 and M_2 events and finally the total number of events per system is registered as M_{tot}.

Step 2: Calculation of selection index (I)

A index is created with the aim of showing the higher potential systems. The index, I is calculated using equation 3.1 and is made up of two components: Firstly, the degree of

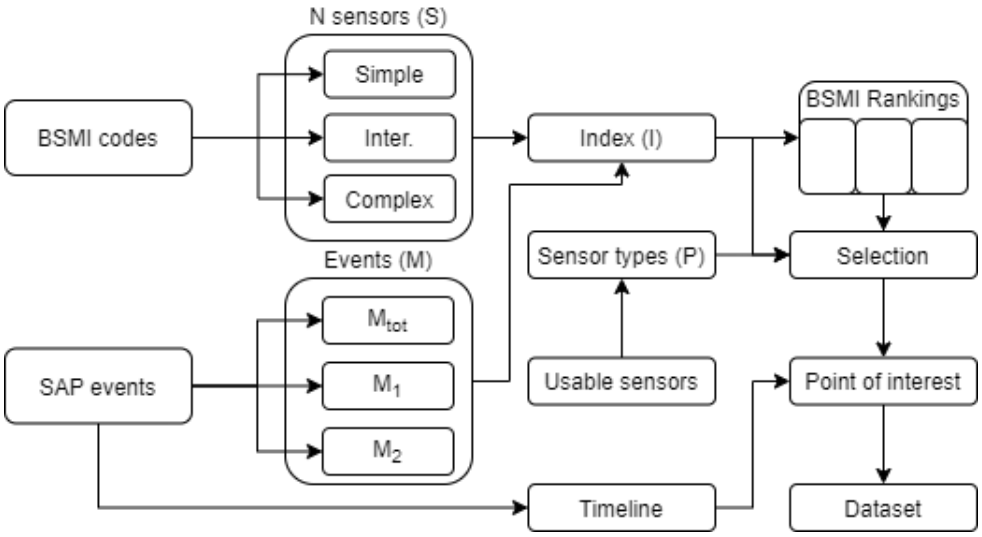


FIGURE 3.1: A diagram showing the data selection method applied

corrective maintenance actions is found by dividing the M_1 and M_2 corrective maintenance events by M_{tot}, the total number of events found per BSMI code. Secondly the amount of corrective maintenance is measured by degree of system complexity where S is the number of sensors per BSMI code. Due to the large variation in systems complexity, the second term is raised to the power of 1/2. This ensures that the index of both simple and complex systems can be calculated using the same equation.

$$I = \frac{M_1 + M_2}{M_{tot}} \sqrt{\frac{M_1 + M_2}{S}} \tag{3.1}$$

Step 3: Ranking and sensor selection

For the simple, intermediate and complex all BSMI codes are ranked based on the resulting index. The BSMI codes ranked highest on the lists now should have the appropriate number of sensors and possible anomalies in their datasets. The quality of sensor data is now assessed. Using a sensor list, all sensors are categorised on type of parameter, named P. The Index I is multiplied by the number of sensor types per BSMI resulting in the final ranking, named R. This is due to the goal of in this research showing GC between different types of data.

Periods of interest

One candidate BSMI is picked from each of these groups. A timeline is created showing all useful events in a chosen time span. To narrow down the eventual dataset, periods of interest are picked. All event data comes from the SAP system and is requested per BSMI code.

3.4 Selection results and potential datasets

The method of selection is applied and the results are shown below. Data on SAP events and a sensor list were collected from an extended period of time. All data was collected from one of the OPV's.

3.4.1 BSMI Rankings

Step one and two are applied and result in the three lists of BSMI codes. Tables 3.1 to 3.3 show the top five results per group. The number of unique parameters is shown column 'P' and the total number of sensors is shown in column 'S'. The product of these is shown in column 'R'; pressure, temperature and power, etc. If the BSMI selection does not result in meaningful datasets when performing the actual analysis the next best BSMI code will be picked instead. The full list of all BSMI codes and all types of sensors can be found in appendix B

Case 1: Simple

From table 3.1 BSMI 12129 is picked having a much higher R score as opposed to the other results.

TABLE 3.1: Selection of BSMI codes for the first case

BSMI	Definition	$M_1 + M_2$	M_{tot}	S	I	P	R
12129	KVDM INST BRANDSTOFINSTALLATIE	24	33	2	2,52	1	2,52
1471	DRINKWATERINSTALLATIE	36	305	3	0,41	2	0,82
1517	WATERMIST BRANDBLUSINSTALLATIE	29	115	4	0,68	1	0,68
1452	STUURMACHINE INSTALLATIE	21	236	4	0,2	1	0,2
1645	DIEPGANGMEETINSTALLATIE	3	12	6	0,18	1	0,18

Case 2: Intermediate

In the second case the second result is picked instead of the first. Here the reasoning is that the first option, BSMI code 1641 is linked to the central management system. Faults in this system could be linked to any other system onboard which will make analysis hard here. Therefore, the second, BSMI 1214 is chosen.

TABLE 3.2: Selection of BSMI codes for the second case

BSMI	Definition	$M_1 + M_2$	M_{tot}	S	I	P	R
1641	PLATFORM MANAGEMENT INSTALLATIE	29	152	16	0,26	3	0,77
1214	VOORTSTUWINGS ELEKTOMOTORINSTALLATIE	13	51	19	0,21	3	0,63
1511	ZEEWATERBRANDBLUSINSTALLATIE	32	284	18	0,15	2	0,3
1591	ZEEKOELWATERINSTALLATIE	31	381	12	0,13	2	0,26
1575	HELIKOPTER BRANDSTOFINSTALLATIE	18	208	11	0,11	2	0,22

Case 3: Complex

BSMI 1331 is picked for the complex case. While having the second highest I score, the amount of different parameters present in the BSMI code (five) makes it outscore the second pick BSMI 1521.

TABLE 3.3: Selection of BSMI codes for the third case

BSMI	Definition	$M_1 + M_2$	M_{tot}	S	I	P	R
1331	VERDELING HOOFDVOEDING 440V	13	37	36	0,21	5	1,05
1521	TRIM-, BALLAST- EN ONTBALLASTINSTALLATIE	28	74	46	0,30	2	0,6
1571	BRANDSTOFLAAD-, TRANSP- EN AFGIFTEINST	10	73	30	0,08	1	0,08
1231	TANDWIELKASTINSTALLATIE	12	327	40	0,02	4	0,08
12121	KVDM INST KRUISVAART DIESELMOTOR	7	183	122	0,01	4	0,04

3.4.2 Dataset selection

Figure 3.2 shows all M_1 and M_2 events registered in BSMI in a considerable period. The blue dots, or events, are shown with priority numbers on the y axis with number 1 being the highest and 4 the lowest. While no selection is made on the priority numbers , its helps to visualise events when situated in close proximity. The periods of interest are plotted in green, these are the points in the timeline where events are registered more closely to each other. The assumption is that failures recorded in data are more likely found here. The three yellow highlighted areas indicate the periods where datasets are extracted. These areas are chosen mostly before and during periods of interest. Appendix C visualises all timelines for the chosen BSMI codes.

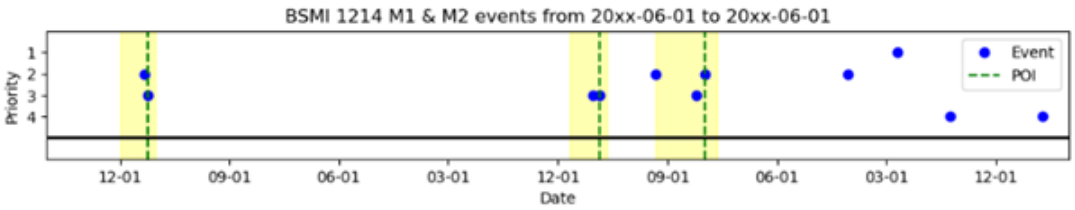


FIGURE 3.2: The timeline of events resulting in the periods of interest for the 1214 BSMI system.

Chapter 4

Methodology

Resulting from chapter 3 datasets are created which would, in theory, contain anomalies. However, before detection of such anomalies is applied the method of GC must first be tested. This is why the method chapter is divided up into two main parts: Firstly an exploratory analysis will be conducted on a full dataset with the goal of comparing GC and correlation as well as a more in-dept test of GC on a found in- and output parameter in the dataset. The second section contains a method of testing the eventual goal of this research; the fault detection capabilities of GC. Finally, the software and other DATs used in the analysis are discussed.

4.1 Exploratory analysis

The main goals of the exploratory analysis are as follows;

1. Initial data-processing
   (a) Successful manipulation of data for the GC tests
   (b) Visualisation of data in the dataset, including registered events
   (c) Creation of correlation matrix for all parameters present
   (d) Creation of GC matrix for all parameters present
2. First in-dept application GC
   (a) Identification of in- and output parameter
   (b) Performance of GC test with an identified in- and output parameter
   (c) List of initial settings for later steps in the analysis

4.1.1 Initial data-processing

Data extraction and visualisation

The first step in the process is the collection and arranging of data for the analysis. As discussed in chapter 3, the resulting datasets are extracted from the DINO application. To get a first overview of the data at hand, a plot is made showing all parameters present in the set. Using the known event-data from the SAP system events are also plotted vertically on the x-axis.

Data manipulation

To reduce file sizes this data has been compressed by the IPMS system to only show changes in registered values. In practice this means that for a parameter like a set-point only one value is registered per day while temperature is measured every second. Since the GC test uses time-series data all parameters have to be reshaped to have the same length. In the end, a dataset is created with equally sized time-series for each parameter present.

Creation of correlation matrix for all parameters present

A correlation matrix is made from all parameters in the dataset. This will give some initial insight in the relationship between parameters and could, in combination with the GC matrix lead to the selection of an in- and output parameter for the in-dept application of the GC test.

Creation of GC matrix for all parameters present

Using the time-series data, GC tests are again done on all parameters in the dataset. The GC matrix will then also be compared with the correlation matrix to show the potential added value of the application of GC on data.

4.1.2 First in-dept application to find GC

Identification of an in- and output parameter

From the dataset the in- and output variables are identified. Next, to verify that the GC test is indeed applicable in this environment a small time frame is selected to perform the analysis onto. One single in- and output variable pair is GC tested. A potential candidate is picked using the correlation and GC matrices.

F-test

GC will be tested over a small time frame. Here the result will show whether the input, I, Granger causes the output, O. Equation 4.1 shows that in order to conclude that GC is present, a F-test as described in section 2.2.3 should result in $P < 0.05$. If results of the test end up being higher than this, no conclusion can be made regarding causality between the two parameters. The test is also performed in reverse to verify that no Bidirectional GC is present. Finally, the F-statistic will be calculated to verify the test results.

By default the significance level for the test is picked to be 0.05

$$P^{I \to O} \leq 0.05 \to GC \tag{4.1}$$

Findings for later testing

In this stage information on effects of resampling data or application of data filtering methods will be gathered as well. The reasoning behind the possible application of filtering is the reduction of file sizes and increasing the speed of computations. However, if loss of information is found to be too large this step will be skipped.

4.2 Further application of the method

Following from the first application of GC a fault detection method will be created and applied to a wider range of data.

4.2.1 Application on larger datasets

The exploratory analysis will only be conducted on a very small piece of a larger dataset. Still, when using a larger set, the basis of the method stays the same. Again, GC will be tested using equation 4.1.

Now, the test will only be conducted when an input value changes. As the F-test is based on finding the reduction of resulting predictive error over time, a test may only be done with a non-static array of data-points. Using a static one will, by definition of the F-test result in the conclusion of no GC. Also, in this way computing time is reduced drastically enabling testing over larger datasets. Figure 4.1 shows the proposed method. Here it can be seen that at the test locations a part of the dataset has been marked in the testing area. The F-test will be done in this area only. The result is that more selective testing is done. The main variables that are relevant here are the size of the testing area before and after the change, and the settings of the F-test itself, so the number of, or selection of lags taken into account.

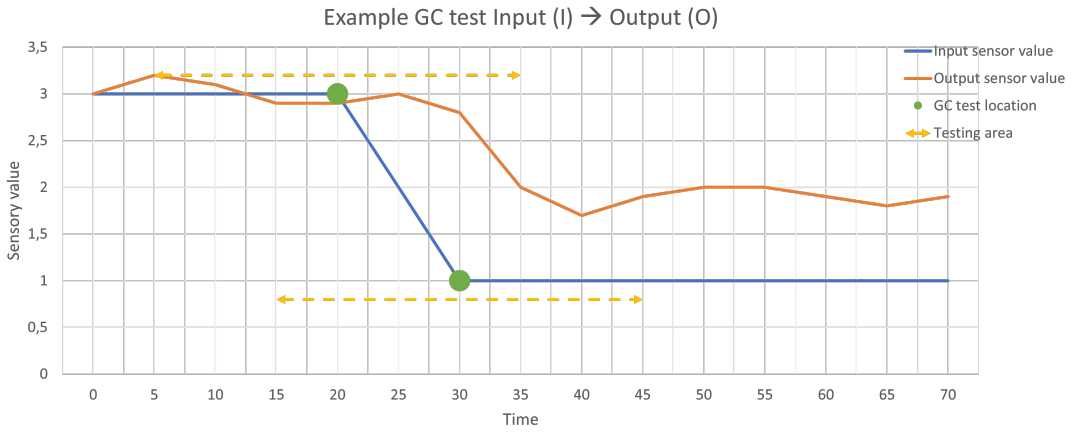


FIGURE 4.1: The proposed GC testing method for larger datasets. Testing only at changes of the input sensor values.

4.2.2 Fault detection

The fault detection method is based on the same confusion matrix as shown in chapter 2. Figure 4.2 again shows this confusion matrix. Where using a F-test, positives (acceptance of null-hypothesis) and negatives (rejections of null-hypothesis) are determined for the prediction. The real, or actual state is then verified visually in the dataset resulting in the TP, TN, FN and FP values.

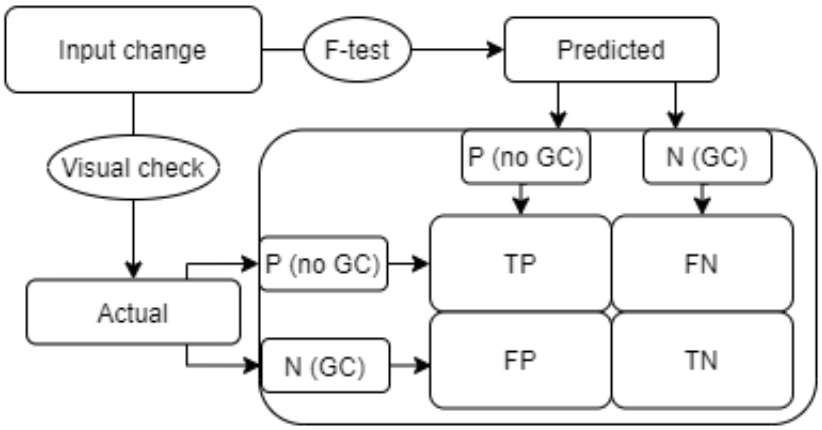


FIGURE 4.2: The proposed method for determining negatives and positives using the F-test and visually.

Determining Negatives and positives using the F-test

For every test conducted at a input sensor change the following statements are proposed:

$$N = P^{I \to O} \leq 0.05 \tag{4.2}$$

$$P = P^{I \to O} > 0.05 \tag{4.3}$$

Formula 4.3 defines any resulting P value lower than 0.05 (indicating GC) will be counted as a Negative or N. This means that here the system is acting normally. The assumption here is that in regular behaviour there should be GC from the input to the output. When the resulting P-value is higher than 0.05 (indicating no GC) a positive is registered. Using the same assumption, this positive registration can be seen as a potential fault in the system. Again, in figure 4.2 it can be seen that the P's or positives show no GC is present while the N's or negatives show that GC is present. This is the same for the visual check.

Use of more than one unit of lag

Most likely, more than one unit of lag will be used for the test. Therefore, a choice has been made regarding the array of results for the f-test for all selected lags. If any of the selected lags show that no GC is present, thus, showing a P-value higher than the selected significance level it will be counted as a positive result. The reasoning for this is that the exploratory analysis will have pointed out what lags are to be used and which are not. F-testing this selection of lags should always point to P-values lower than the significance level showing that GC is present. This, in turn would mean that any diverging value will mark a problem at that time and will therefore be counted as a potential fault.

Accuracy, precision, recall and F1-score calculations

The same metrics are again employed as described in chapter 2. All formulas are again shown below in equations 4.4 to 4.7

$$Acacuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.6}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.7}$$

4.3 Software and other data analysis techniques

4.3.1 Software

The entirety of the code is written in a Python environment using Jupyter notebook. Within Python the *statsmodels* package is used specifically for the F-test. *Statsmodels* includes an built-in module for this purpose and includes the following inputs:

1. Two columns of data, being the in-and output of the time-series
2. Lags: the selection of lags taken into account for the test
3. Resampling rate: Different resampling rates will be tested to find its effect on the accuracy of the results

Finally, a small study on filtering of data will be conducted in the exploratory part of the research.

4.3.2 Other data analysis techniques

As stated in section 4.1 other DATs may have to be included in the test to improve its quality. In the application on a larger scale, the method of testing on GC is done with the testing area. This testing area was introduced following the findings made by the study on anomaly detection using a varying window size [11].

Chapter 5

Results

5.1 Exploratory Analysis: BSMI 1214

Following chapter 3, datasets were selected for further analysis. Early examination of the data found that any of the selected BSMI codes in the simple case had little to no value. The sets did not contain coherent data to test further. Mainly, input parameters could not be identified here. Therefore, the choice was made to start with the first selected BSMI code from the intermediate case: BSMI 1214. This case was found to be rich in data and will thus be used for testing. Moreover, the complex case was assessed to be redundant in further testing since the intermediate case contained enough data for thorough testing of GC.

5.1.1 BSMI 1214: Electric engine propulsion system

The system used for the execution of the proposed method is the drive-train system on the OPV. It consists of two electronic motors which are abbreviated as PEM or 'patrouille-evaart elektrische motor'. All sensors are linked to one of the motors defined by either the SB (stuurboord/ starboard) and BB (bakboord/ port side) of the ship. The two engines are positioned next to each other ensuring redundancy if one engine were to fail. In normal use, both engines will be used in tandem [32]. Figure 5.1 shows such a setup. A diesel engine generates power which is directed to a converter and finally to the electric engines. The system shown is used by the Finnish Defence forces for their newly designed corvette-class ships [32]. Although not entirely identical it does depict a similar setup.

5.1.2 Initial data processing

Data extraction, visualisation and manipulation

This means that the first selected dataset is that of BSMI code 1214 as part of the intermediate case. Below, in figure 5.2 the full dataset to be analysed is first shown. The starting and ending dates are listed on top and all parameters are listed in the legend. Also, vertical striped lines show the events in red at that time. The dates are directly imported from created timelines in section 3.4. The set contains periods of operation, but also idle time.

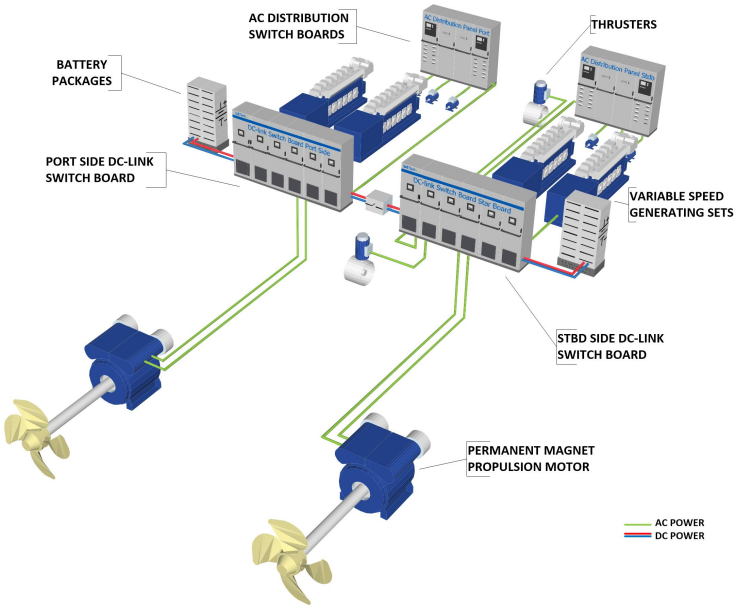


FIGURE 5.1: A visual representation of a double engine propulsion system for a ship using an electric engine powered by diesel-generators [32]

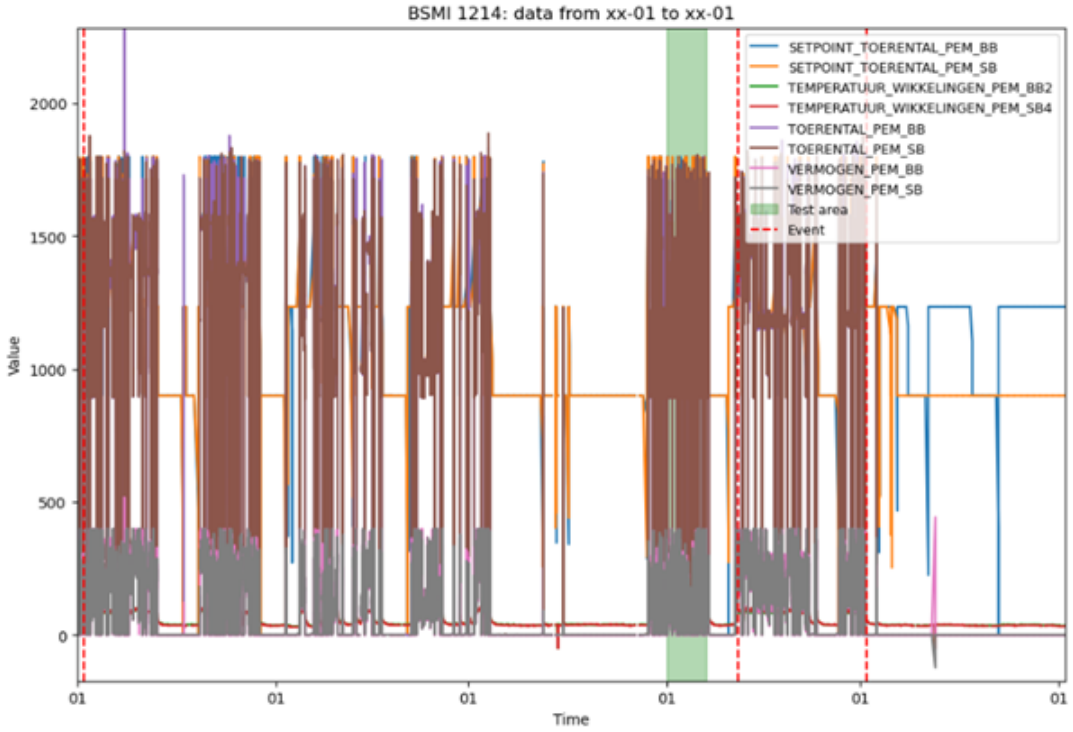


FIGURE 5.2: The BSMI 1214 data gathered from the DINO application.

Within this set a number of parameters exist; Since BSMI 1214 is part of the drive-train on-board the OPV it contains sensor readings on output power (VERMOGEN) and Rotations Per Minute (RPM), (TOERENTAL) reached by the motor. Next to this, the temperature of the motor is monitored as well (TEMPERATUUR). Finally, a RPM set-point (SETPOINT-TOERENTAL) is logged containing input information for the drive-train.

Figure 5.2 shows the full dataset. For the exploratory testing a much smaller size dataset is needed. figure 5.5 therefore, shows a week long part of the data. This part is also highlighted in figure 5.2 in green. This section of data is then reshaped such that all parameters are equal in size. This is done by forward filling any of the compressed sections of IPMS data. The logic here is that only data which had been static is compressed by the IPMS system and thus forward filling is the most simple approach. A correlation and GC matrix is constructed with this data.

Correlation matrix

Figures 5.3 and 5.4 show the results of both the correlation and GC analysis. The correlation matrix shows that between the SB and BB pairs, parameters correlate perfectly in most cases. This is logical as both engines are designed to be used together. The SETPOINT-TOERENTAL parameters, both SB and BB, seem to only correlate with the output power, or VERMOGEN with an average strength. The output parameters regarding temperature, RPM and power correlate strongly with one another.

GC matrix

Next is the GC matrix. At first glance it becomes noticeable that, contrary to the correlation matrix, this one is not symmetric along the diagonal. As stated before in section 2.2.2, GC may differ depending on the direction of analysis. The matrix should therefore be read as follows: Parameters listed on the y-axis are tested to cause the parameters on the x-axis. This is also depicted in the bottom left corner of the figure. The direction of GC testing is shown going from in- to output with the resulting P value filled in the matrix. For instance, the temperature parameters do not GC the set-point parameters while this does occur the other way around. Namely: $TEMPERATUUR \rightarrow SETPOINT > 0.05$ while $SETPOINT \rightarrow TEMPERATUUR < 0.05$.

Note that the F-test is applied here with a significance level of 5% or 0.05. In the GC matrix differences between the SB and BB parameters also become clear. The set-point-SB parameter does not GC the RPM parameter on BB while it does on SB. Again power output on the screw can not be traced back with another parameter, all GC relationships to the power outputs are valued higher than 0.05.

In many cases bidirectional GC is found, this could be the result of problems with frequency or sheer size of the recorded data. Or, a common-cause problem where both parameters are caused by a third one [16]. Lastly, it could also be caused by an case of instant GC [16].

Added value of the GC matrix

In all, the correlation matrix is able to quickly distinguish which parameters are related closely to each other. The GC is then able to show whether a causal relationship could exist between these as well. Due to the size/quality of data however, bidirectional GC is found in the majority of the cases. This means that in further applications a smaller dataset will be needed to clearly explain the causal relationships.

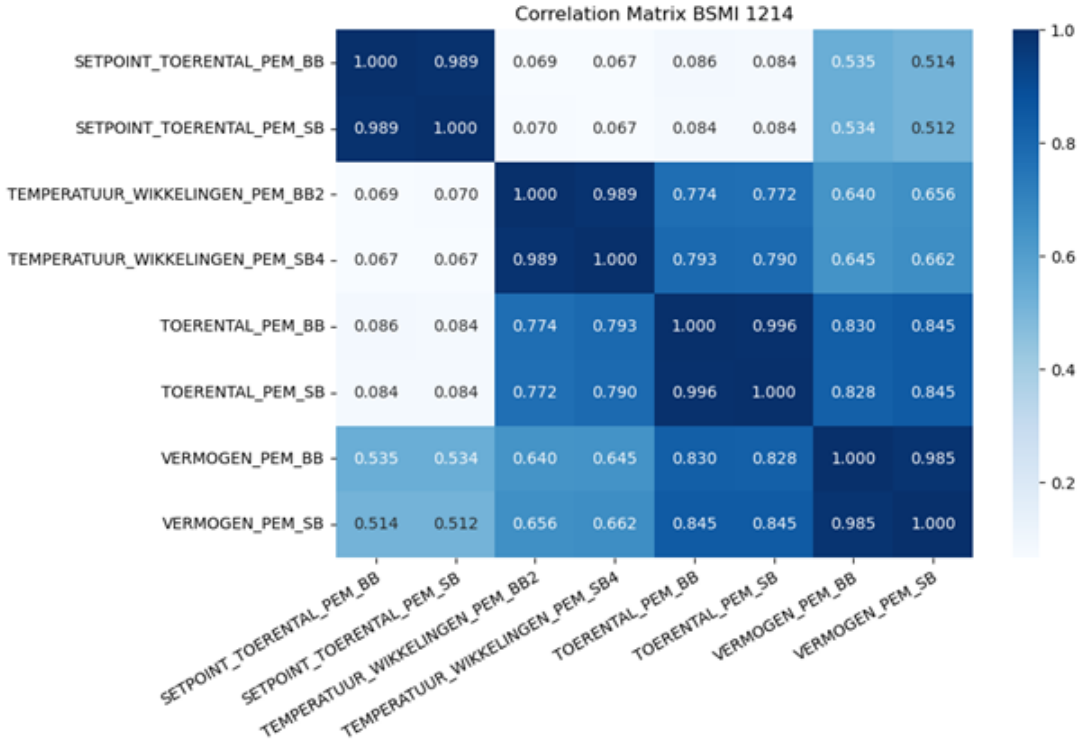


FIGURE 5.3: correlation matrix for the BSMI 1214 data

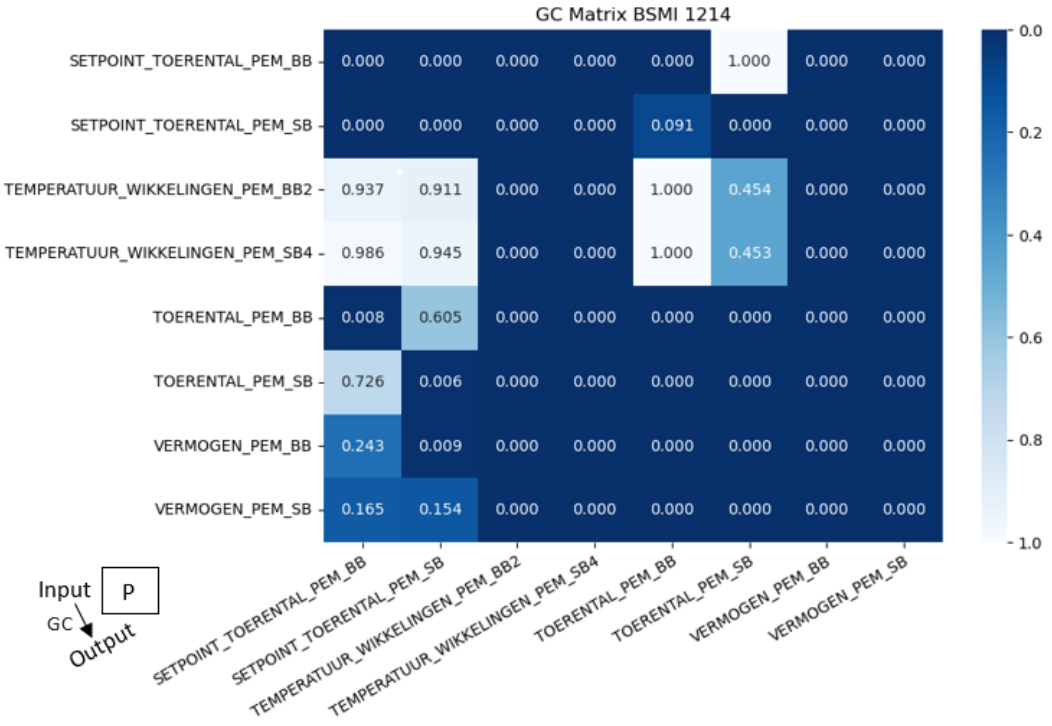


FIGURE 5.4: GC matrix for the BSMI 1214 data

5.1.3 Selection of in- and output parameter

Upon further inspection of data, it can be concluded that from all available parameters a suitable selection would be to test the SETPONT-TOERENTAL parameter as the input for the actual output TOERENTAL. These are the in- and output RPM values of the motor. Here, either SB or BB would be correct. They must, of course, both be parameters from the same side of the ship as the GC matrix show that no causality exists between SB and BB parameters. For the exploratory research the SB data was used. In the further analysis temperature sensors could be addressed as they are clearly caused by RPM parameters. Also, the power parameters could be considered as again, a causal relationship between the two could exist.

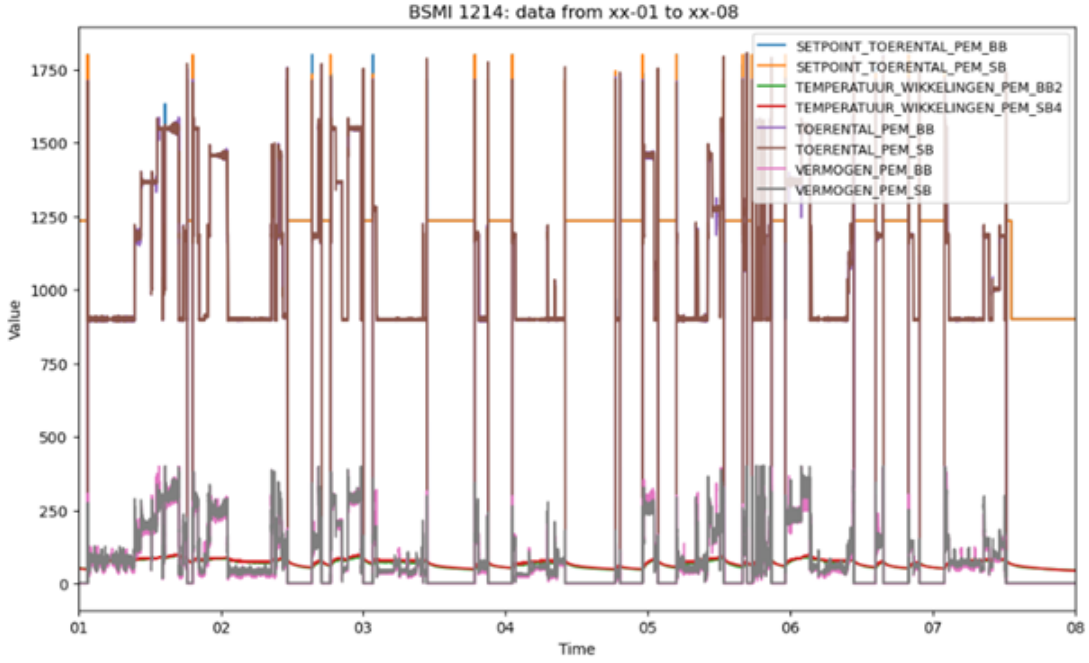


FIGURE 5.5: A week long piece of the BSMI 1214 dataset

5.1.4 Application of Granger causality test of in- and output data

Resulting from the selection of parameters, Figure 5.6 isolates only the in- and output parameters discussed earlier. During operation, the assumption is that whenever the set-point changes on the ship, the actual output of RPM will follow quickly.

Visually, this can be confirmed quite easily. However, during some phases where the set point value stays stable at around 1250 the actual output falls to 0. The conclusion is made that this is a glitch in the software and no actual error occurs in the system. Still, it does offer an opportunity to detect anomalies like this later on in the full analysis. Zooming in to a single set point change, figure 5.7 shows the data in detail.

Again, this particular part of the dataset is highlighted in green in figure 5.6. The reduction in set point results in an almost immediate decrease of actual delivered RPM's. Next to this, the data shows some visible oscillations which could be explained by possible noise in the sensor readings or the motor is continuously controlling to follow the set point.

Nevertheless, the use of filters could prove useful as it would allow resampling, and thus, condensing data without resulting in loss of information.

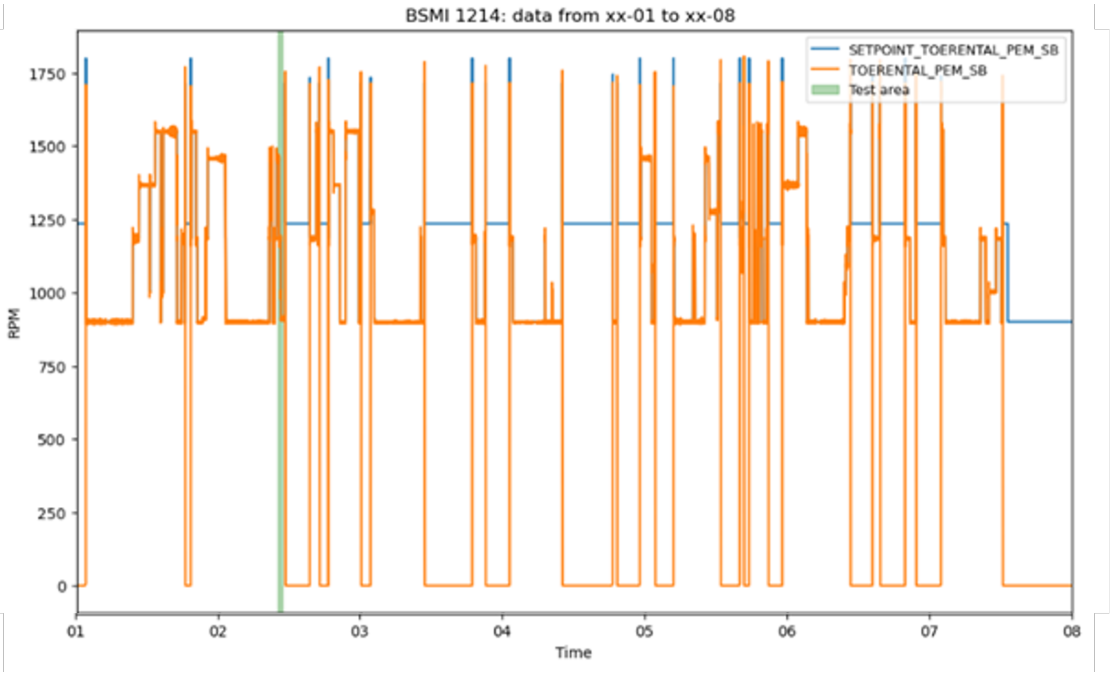


FIGURE 5.6: A week long piece of the BSMI 1214 dataset showing only the in- and output parameters used for initial GC testing

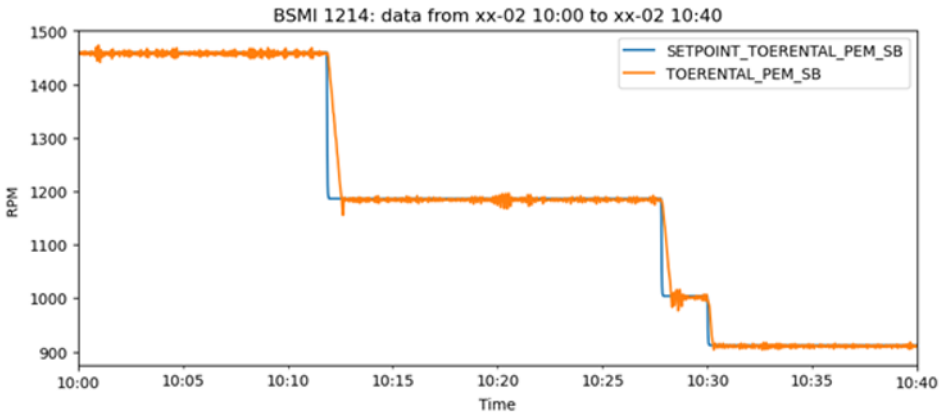


FIGURE 5.7: A graph showing the in- and output data during a set point change.

Inclusion of filters

The following Three filters were applied on the in- and output:

1. A MAF (moving average filter) may be used on time-series data. As the name implies, it revolves around smoothing the data at hand by plotting values that average out within a set time frame. In larger sets of data the MAF filter can be tuned such that the time frames are dynamic such that information loss is minimal [2]. Generally, including an MAF filter does add a delay over the original data.

2. A LPF (low pass filter) is designed to remove noise at pre-determined cut-off frequency and can thus be tuned such that noise is removed. Other than the MAF, it does not result in an added delay [30].

3. A LOWESS (locally weighted scatterplot smoothing) filter, this type of filtering is based on linear regression. It is very effective in removing outliers from time-series data [33].

The application of these filtering techniques is shown in figure 5.8. Here, the set point, original and filtered versions of the data are plotted together to see which filter method best captivates the original data. Quite easily, the LPF looks to be the most suitable filter. It removes the oscillations while also mimicking the decrease in RPM the best. The MAF does remove noise. It does however, clearly result in an added lag. For the GC application, this is particularly bad. Finally, the LOWESS can be disregarded, the application takes too much computing time and results in changing the overall relationship between the two parameters. For instance, the LOWESS filtered data drops off earlier than the setpoint does, thereby creating the impression that it would cause the setpoint change instead of the other way around.

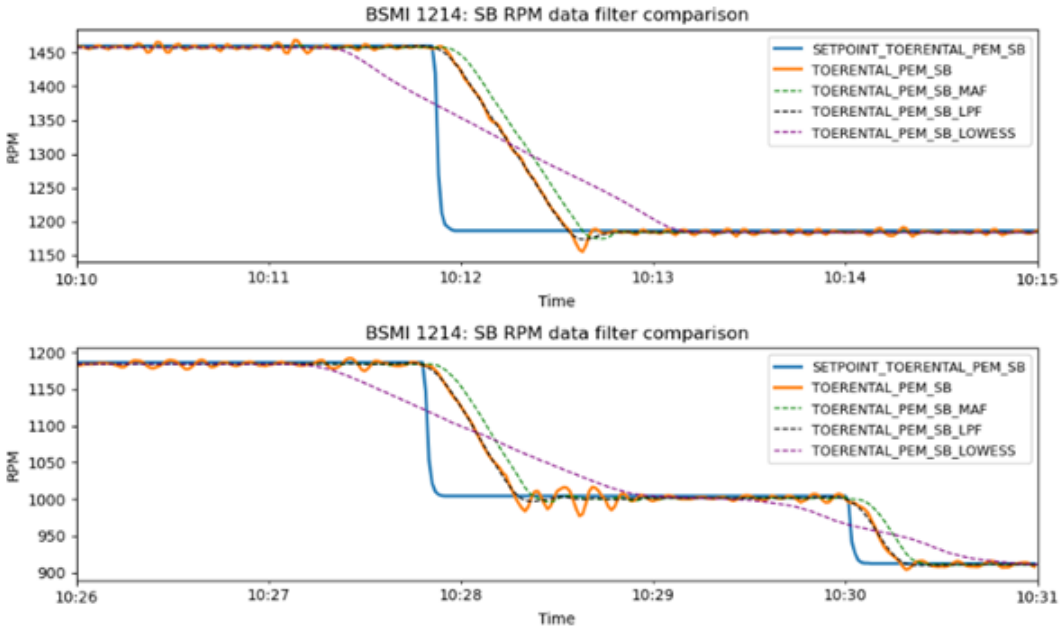


FIGURE 5.8: The application of filters on the exploratory data showing the MAF, LPF and LOWESS filtering techniques.

Choice of maximum resampling rate and amount of lags tested

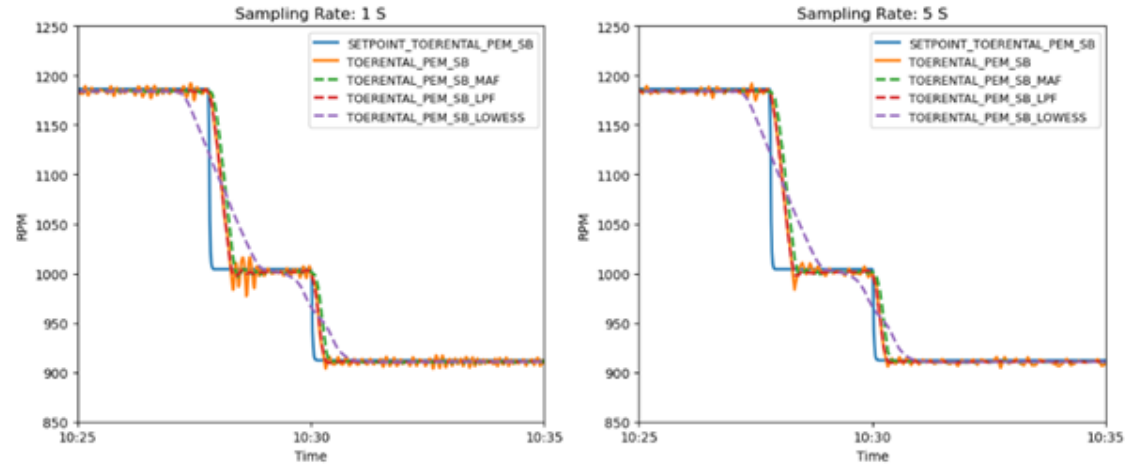
Overall, the goal of resampling is to condense data such that in larger sets computing time is decreased. The resampling is done in multiple steps such that the effect on the GC results is made clear. The GC test is done on the week long data portion as seen in figure 5.6. Tables 5.1 to 5.5 show that resampling is done on 1 (original), 5, 10, 15, 30 and finally 60 seconds. As seen in figure 5.7, the change of set-point is followed by actual RPM's in about 30 seconds to maximum of about. This is the reason a maximum of 30 seconds is chosen for this variable in this test.

The effects of this resampling on the data can be seen in figures 5.9a to 5.9c where the resampled segment of data from figure 5.7 is shown. Upon further inspection it can be concluded that from resampling at 15 seconds or higher, the quality of the data is reduced significantly. Here, real RPM data seems to lower before the set point actually does. In theory, this would mean that the GC test results should show a measure of loss of GC from this point onwards as well. The 10 second resample rate is therefore considered as the balance between loss of information and condensing of data

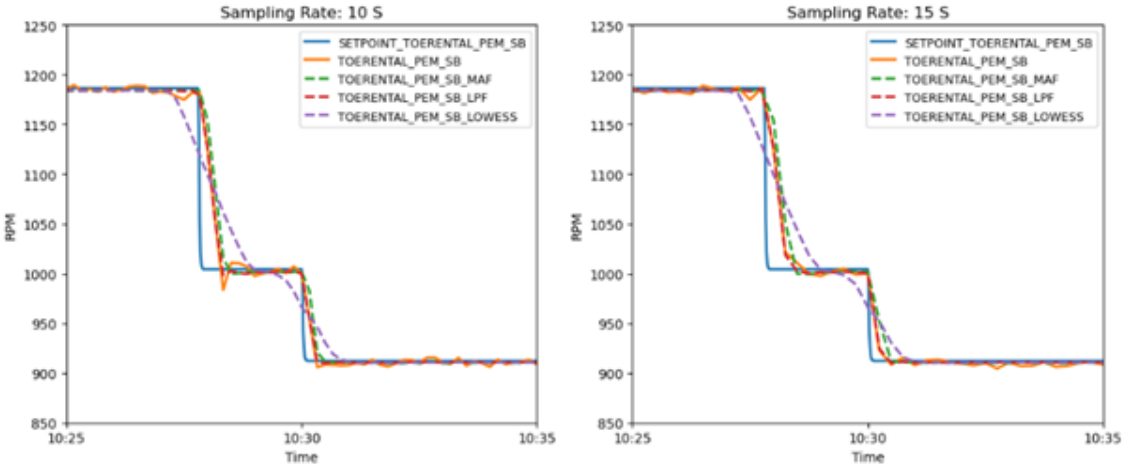
For this part of the analysis the choice was made to include four lags in the analysis. At the 10 second resampling rate it will ensure that the full transition of output RPM is analysed with a margin of one lag. At the highest resampling rate (30 seconds) it would mean that the fourth lag will determine GC from in- to output two minutes away. This is done to find what the effect is of looking beyond the perceived relationship. Lags may be added or removed in later stages.

Choice of length of the test

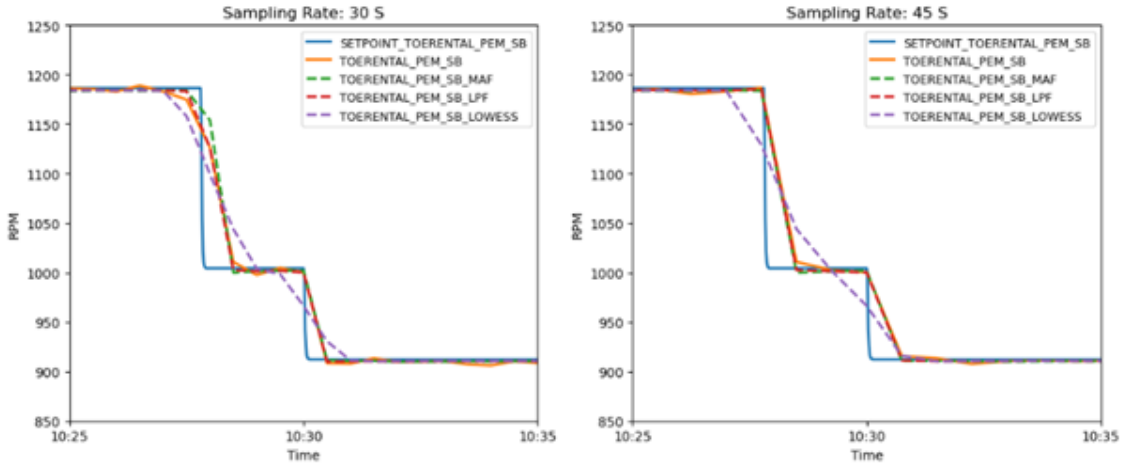
All tests were conducted between $10:25$ and $10:35$ which means that the total length of the tested area amounted to ten minutes. Depending on the resampling rate this means that the length of the test varies between 600 samples at 1 second resample rate to 20 samples at the 30 second resampling rate.



(A) Resampling at 1 and 5 seconds



(B) Resampling at 10 and 15 seconds



(C) Resampling at 30 and 60 seconds

FIGURE 5.9: Resampling of sensory data with rates from 1 to 60 seconds

GC testing results

All GC results are listed in the tables 5.1 to 5.5. A significance level of 0.05 is defined as discussed in section 4.1.2 . Again, if the test results in a value lower than this threshold the hypothesis that a GC relation exists can not be rejected. Therefore, a Granger causal relation may exist.

The first table, table 5.1, shows the relationship between the original data and the set point. Then, table 5.2 to 5.4 show results with the used filters and then finally, table 5.5 shows the relationship the other way around. If the theory works it should show p-values exceeding 0.05.

Table 5.1 shows that an GC relation exists. All values up to an resampling rate of 10 seconds are zero. Then, looking at resampled data at 15 seconds, the results increase marginally. At 30 seconds, the results in table 5.1 are the highest. Interestingly, on lag four the apparent relationship is stronger again. This could be explained by the overall loss of coherence in data [16]. An indicator for this is that in table 5.5 the p-value at lag four in the 30 second resample column is lower than the set threshold as well.

TABLE 5.1: GC from SETPOINT_TOERENTAL_PEM_SB to TOERENTAL_PEM_SB

	resampling rate (s)				
Lag (j)	**1**	**5**	**10**	**15**	**30**
1	0.0	0.0	0.0	0.000011	0.004005
2	0.0	0.0	0.0	0.000133	0.024278
3	0.0	0.0	0.0	0.000836	0.026435
4	0.0	0.0	0.0	0.002533	0.000325

Looking at the filtered results, table 5.2 shows the effect of the MAF filter. Here, the p-values are considerably lower in the 15 second resampled column. The added lag caused by the MAF filter is the explanation of this. In table 5.3 the LPF filtered results are very similar to the original data while the LOWESS filter results in table 5.4 are different. The before mentioned effect of the LOWESS filter on the original data has increased p-values in columns resampling data a 5 and 10 seconds. While the values are still well below the threshold it should still be noted.

TABLE 5.2: GC from SETPOINT_TOERENTAL_PEM_SB to TOERENTAL_PEM_SB_MAF

	resampling rate (s)				
Lag (j)	**1**	**5**	**10**	**15**	**30**
1	0.0	0.0	0.0	0.0	0.000063
2	0.0	0.0	0.0	0.0	0.000455
3	0.0	0.0	0.0	0.000004	0.000520
4	0.0	0.0	0.0	0.000028	0.0

TABLE 5.3: GC from SETPOINT_TOERENTAL_PEM_SB to TOERENTAL_PEM_SB_LPF

	resampling rate (s)				
Lag (j)	**1**	**5**	**10**	**15**	**30**
1	0.0	0.0	0.0	0.000006	0.001882
2	0.0	0.0	0.0	0.000318	0.009825
3	0.0	0.0	0.0	0.002512	0.013558
4	0.0	0.0	0.0	0.007108	0.000401

TABLE 5.4: GC from SETPOINT_TOERENTAL_PEM_SB to TOERENTAL_PEM_SB_LOWESS

	resampling rate (s)				
Lag (j)	**1**	**5**	**10**	**15**	**30**
1	0.0	0.0	0.0	0.000024	0.002920
2	0.0	0.0	0.0	0.007178	0.016843
3	0.0	0.000108	0.000046	0.078602	0.019577
4	0.0	0.000223	0.000439	0.168411	0.0

As a confirmation of the shown data, table 5.5's results are much higher, often exceeding the set threshold. Only on the first lag at the 1 second resample a p-value lower than the threshold is found. The explanation of this could be the following; As seen in figure 5.7 the change in set-point takes about 30 seconds to be followed in the actual RPM data. In the first lag of a 1 second resampled GC test this relation is not correctly portrayed. This leads to the conclusion that, for this relationship a GC test should include lags higher than the first. Or, resampling rates should be made higher such that the first lag is distanced further away.

TABLE 5.5: GC from TOERENTAL_PEM_SB to SETPOINT_TOERENTAL_PEM_SB

	resampling rate (s)				
Lag (j)	**1**	**5**	**10**	**15**	**30**
1	0.002235	0.292826	0.387168	0.869672	0.547745
2	0.188880	0.624912	0.716901	0.730004	0.801005
3	0.751243	0.691054	0.941682	0.722649	0.913367
4	0.704881	0.408959	0.963906	0.776231	0.046003

5.1.5 Preliminary conclusions

Section 4.1 laid out four goals for the exploratory analysis. With the visualisation and acquisition of data the first two of these are met. Goal number three, regarding the performance of the GC test showed that a GC relationship can be found between an observed in- and output. Furthermore, a distinction can be made between a GC test from in- to output and the reverse of this. Resampling of data is possible, however, resampling at rates higher than 15 seconds may be unwise as information is lost.

Therefore, in further application, which is the application of fault detection over larger periods of time, a resample of the data with a rate of 10 seconds or lower is taken. From all used filters, the LPF's results mimic the original data the best. However, since the differences are minimal the choice may be made to not implement a filter at first.
Following from the reverse tests, the first lag at a 1 second resampling rate may not be useful. The same can be said for the use of lags too late or at too high resampling rates. The resampled dataset showed that when far enough GC can be found going both sides. In abnormal behaviour, higher GC test results are expected within about 30 seconds.

The number of samples taken in each test depended on the resampling rate chosen. Here, again the conclusion is made that the resampling rate had most influence on the results of the test. F-statistics where all sufficient for the test, however, where low for the 30 second resampling test indicating that the hypothesis, although rejected, was done so with little statistical certainty. Therefore, a minimum of 40 data-point is taken, derived from the 15 second resampling rate tests. The conclusions on proposed settings, resampling rates and filters only apply on this application with this particular in- and outputs. As section 5.2 includes more applications the settings of the test may change. However, for this application these findings are applied initially.

5.2 Fault detection using Granger causality

The following section addresses 3 applications of the fault detection technique:

1. BSMI 1214 set-point RPM to RPM: The first application should lead to the best performance of the fault detection method. The in- and output parameters are related most closely. Also, the exploratory analysis showed that a strong Granger causal relationship exists from the input towards the output.

2. BSMI 1214 set-point RPM to POWER: In the second application the goal is again to achieve a well functioning fault detection method. Now the power (VERMOGEN) parameter is used. Although assumed to be causal, it can already be seen in figure 5.10 that the data is much more unstable. Therefore, it is assumed that performance of the detection method could lead to a lower level of performance.

3. BSMI 1214 RPM to POWER: The correlation matrix showed that these parameters are highly correlated. To verify that correlation is indeed not the result of a close causal relation, the final application is meant to verify that both of the output parameters indeed follow from the single input parameter. If the detection method is working correctly the test should show a low accuracy and low performance level (low $F1$-score).

In this section again data is used from BSMI 1214. The same week-long dataset as seen in figure 5.5 is now subjected to a fault detection method as previously laid out in section 4.2.1. Figure 5.10 shows only the parameters tested in this section. In all F-tests conducted in this chapter resulting P-values have been verified using the F-statistics.

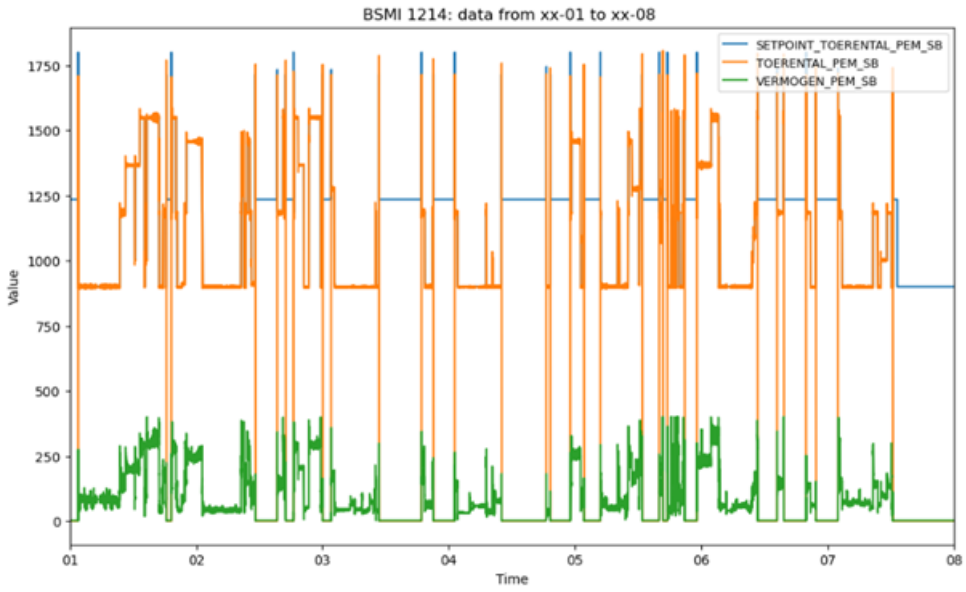


FIGURE 5.10: Data used for the various test cases

5.2.1 Visual inspection of the dataset

Before fault detection can take place with the proposed method, a visual inspection of the data will point out what number of faults exist in the dataset. In total, 177 set-point changes are detected of which eventually 40 are found to show a visual loss of GC. The number is significant, however, a large portion of cases can be attributed to an error in the IPMS's data logging. This means that two kinds of loss of GC exist:

1. Set-point 0 errors: 32 of these errors are found in the data. These are the moments the set-point seems to stall at 1250 RPM during idle time while the actual output drops to 0. The loss of GC is very prominent here as the difference between in- and output parameters becomes very large. A error is identified at the start and end of an idle moment.

2. Potential faults: 8 are found in total. The potential faults are cases that are of most interest. At these moments GC is lost due to a possible malfunction. To give an example, figure 5.12 shows three examples of moments of potential fault in the system. In the first two cases input parameters diverge from the input parameter during or around a set-point change. In the third case a fault happens at a moment when the input value remains static, still both the power and RPM parameters show either a spike/dip at that particular moment. All remaining potential faults can be found in appendix D.

Figure 5.11 shows the same, before mentioned dataset now including all 40 cases of loss of GC. The set-point 0 errors are plotted as black-dotted lines and are relatively equally divided over time. Potential faults occur less often, more notably, in very close proximity at one point in time listing 5 of the 8 in total between a time frame of 2 hours at the end of day five. The potential faults are depicted in red, stripped lines.

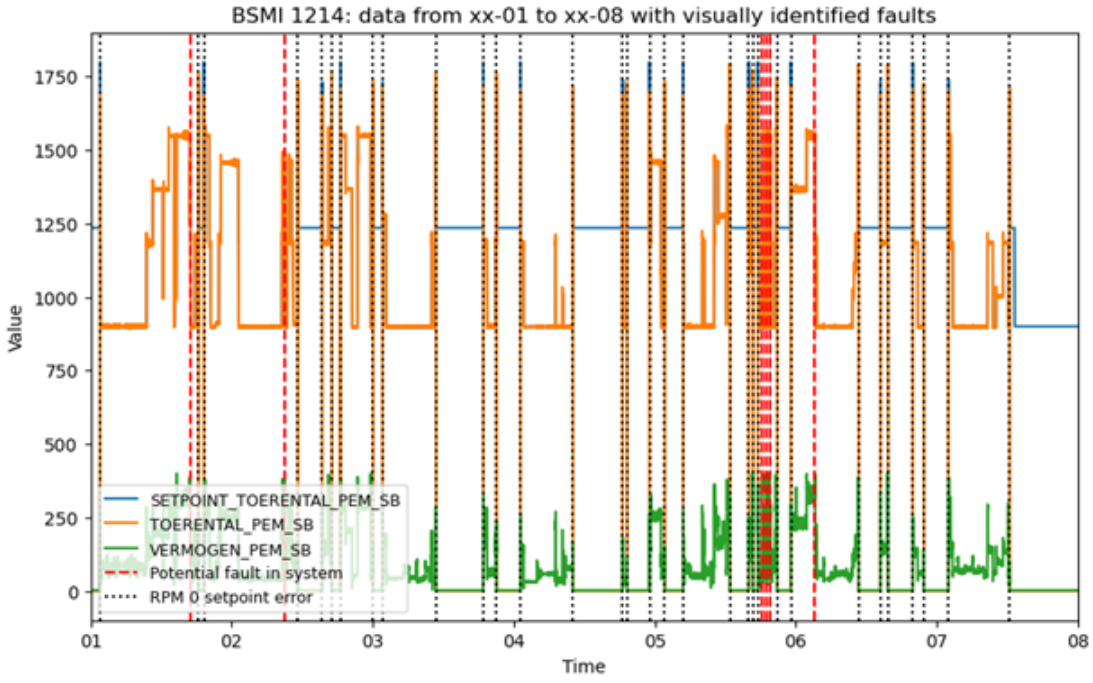


FIGURE 5.11: A plot showing all visually identified cases of loss of GC

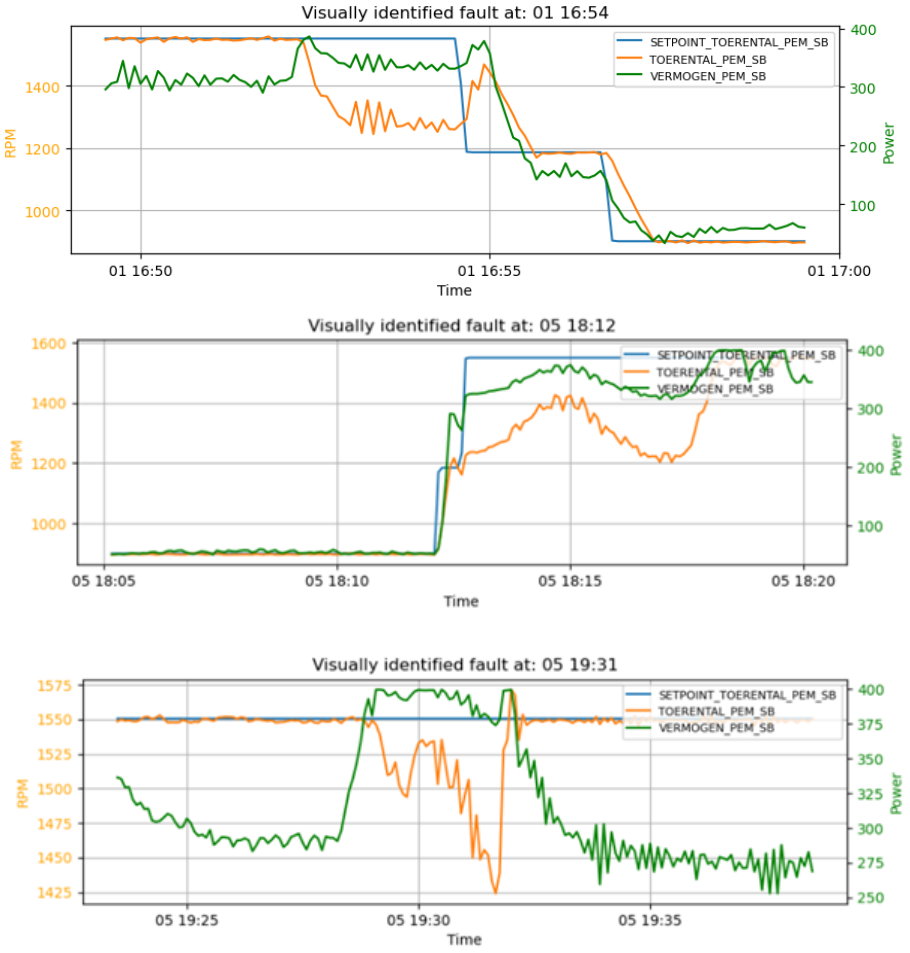


FIGURE 5.12: Three examples of visually identified faults in the dataset.

5.2.2 Application 1: BSMI 1214, set-point RPM t RPM

Figure 5.13 offers a visual representation of the results. Due to the effectiveness of the test only the unidentified or falsely identified problems in the dataset are plotted.

Test settings and results

All used settings for this test are listed in table 5.6. All testing parameters are shown in the top section of the table. The total length of the test is expressed by both the total amount of time (in seconds) and the amount of data points (n). Most interesting in the table are the overall accuracy and performance (F1-score) of the fault detection method. An accuracy of 98% (0.98) is concluded and a F1-score of 96% (0.96), Also notable, the precision reaches the maximum of 100% (1) showing that the detection method never marked normal behaviour as an anomaly.

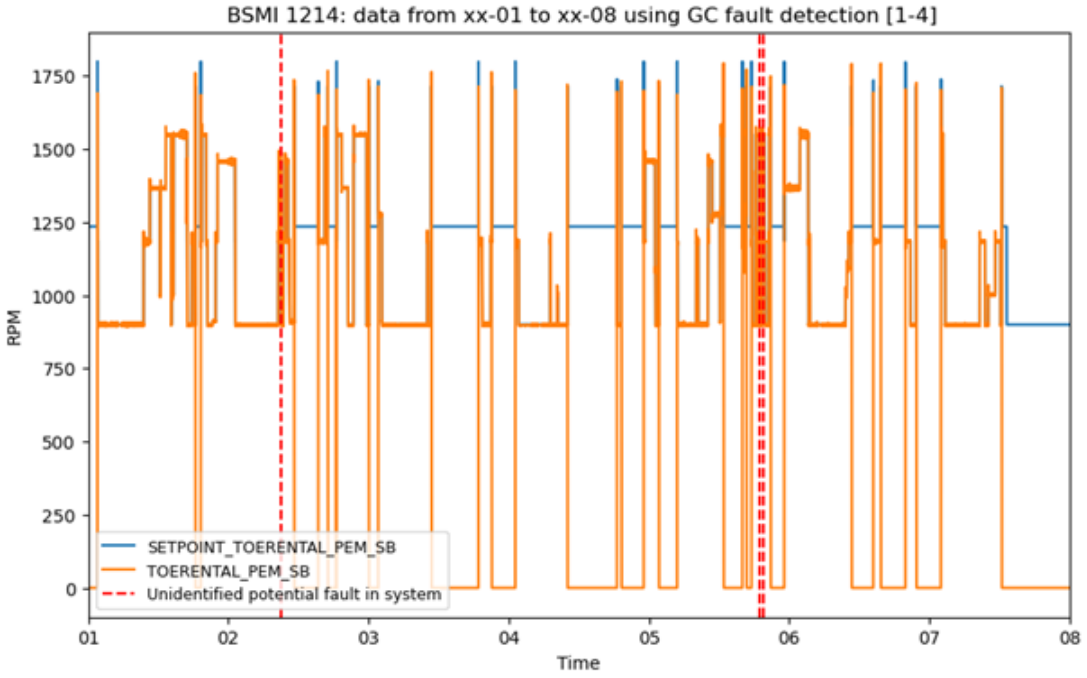


FIGURE 5.13: Results of the first fault detection test showing only the unidentified cases of loss of GC

TABLE 5.6: Settings, results and evaluation metrics from test 1 in application 1

Fault detection: BSMI 1214 Setpoint RPM to RPM (test 1)			
Test setting		Value	
Resample rate (s)		5	
Data points before set-point change (n)		30	
Data points after set-point change (n)		30	
Total length of test (n, s)		60, 300 seconds	
Filter		no	
Lags used (j)		[1, 2, 3, 4]	
Significance level (-)		0.05	
Test results	Total	Potential faults	Set-point 0 errors
P_{visual}	40	8	32
P_{F-test}	37	5	32
N_{visual}	137	-	-
N_{F-test}	140	3	-
Confusion matrix	result	Evaluation metrics	result
TP	37	Accuracy, A	0.98
TN	137	Precision, P	1
FP	0	Recall, R	0,93
FN	3	F1-score, F1	0.96

Analysis of unidentified faults

Three potential faults remain undetected: one halfway day two and two at day five. Looking at figures 5.14 and 5.15 both sections are shown in more detail. At day two, at 09 : 02, shown in figure 5.14, Input RPM drops just before the change in set-point to which it spikes to the before set level to then act normally again. The figure also shows the boundaries of the F-test conducted at that time. One possible reason for the method to no detect this fault could be the reaction to the set-point change. At the moment of the change, the causal relationship is regained again very quickly which could lead to the F-test resulting in the conclusion that actually a causal relationship does exist. Further analysis is needed to draw conclusions in this regard.

The other two are found on day five in the same area as the one pointed out in section 5.2.1 as the location where five potential faults are found. A section of the dataset is shown in figure 5.15 where two potential faults remain undetected. First of all it should be noted that the potential faults at around 18 : 12, 18 : 21 and 18 : 55 are detected successfully, therefore these are marked in green in the figure. At 19 : 01 the same reason for identification could be concluded as at day two. Here, the output RPM value regains a causal relationship almost instantly at the set-point change. Again, further research is needed to draw conclusions. At 19 : 30 the reasoning is much more simple: Due to the fact that only a small area is included in the F-test at the moment of a set-point change, no test contains area of time at which this fault occurs. The fault occurs more than ten minutes before or after a set-point change while the tests only include five minutes of data. This shows the limitation of the method by only testing for GC at times of an input change.

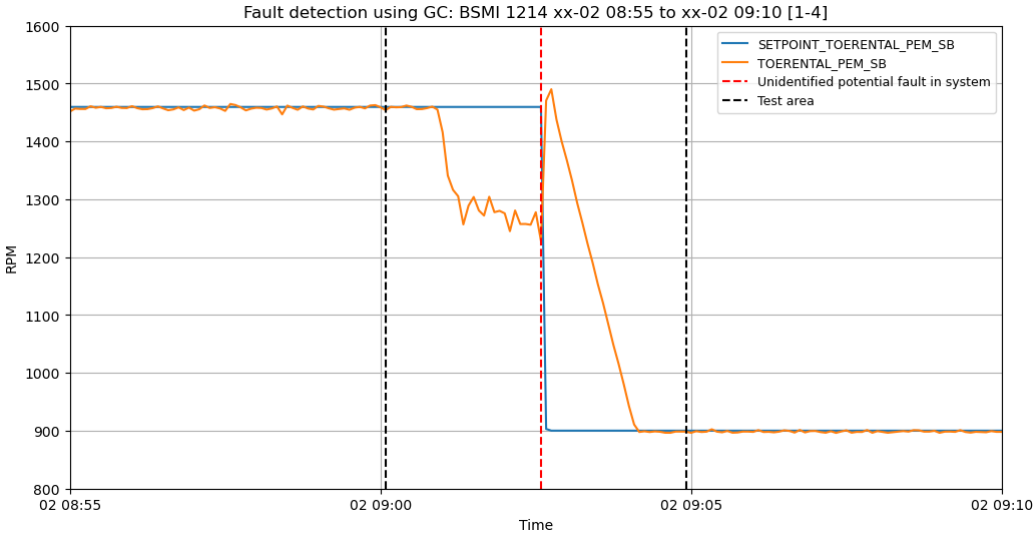


FIGURE 5.14: Zoomed in section of the data showing an undetected fault

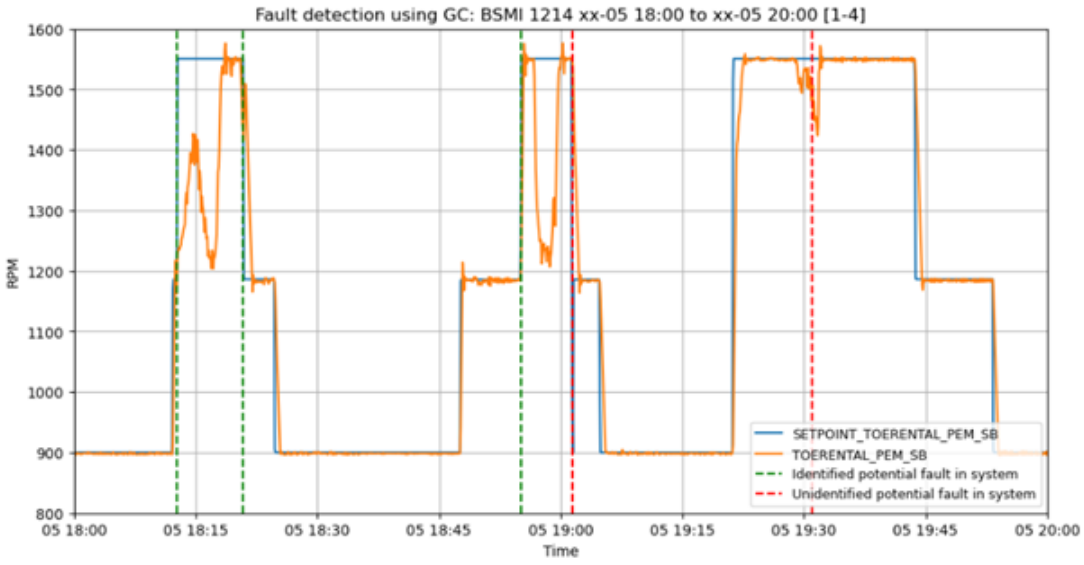


FIGURE 5.15: Zoomed in section of the data where two potential faults remain undetected.

Analysis of resulting P-values

To get a better understanding of the test results an calculation is made on the resulting P-values for all tested lags at times of a detection of loss of GC. Table 5.7 shows the average resulting p-value per analysed lag for both the set-point 0 errors as well as the potential faults. The goal is to see how these two cases of loss of GC differ in their F-test result.

First of all, for the set-point 0 errors all resulting values are very high. This means that the F-test concludes with a high amount of certainty that no GC relation exists between the in- and output. For every analysed lag the P-value seems to increase considerably. At lag it is about six times smaller than at lag four. A possible explanation for this could be that the system is functioning properly at the start of the set-point change. However, over time, as the in- and output parameters diverge further due to the error, the loss of GC becomes increasingly more apparent.

For the potential faults, a large difference exists between the P-value in the first lag as opposed to the later ones. The average of the first lag is about 3 times higher than later ones. This could be the result of potential faults being an a-typical reaction to a set-point change. Normally, the relation should be almost instant as was found in the exploratory analysis. At potential faults this is clearly not the case anymore.

Two of the three undetected potential faults are listed in the table as well. Here it can be seen can again, P-values are highest in the first lag with the notably high P-value of lag one at $02\ 09:02$. Here a decrease of significance level to 0.01 could ensure that this fault is detected as well. At $05\ 19:01$ the results are very low, all well below the significance level of 0.05. It seems that at that moment, the F-test shows that there is a causal relationship.

TABLE 5.7: average resulting P-value per type of detection in application 1: Set-point RPM to RPM

Average lag value	Lag (j)			
Type loss of GC	1	2	3	4
Set-point 0 error	0.0835	0.2924	0.4178	0.5406
Potential fault	0.3162	0.1102	0.1417	0.1261
02 09:02	0.0304	0.0080	0.0003	0.0000
05 19:01	0.0002	0.0000	0.0000	0.0000

Adjustment of test settings for a second test

Of the eight potential faults five where detected in the first test. Considering the goal of the research the determination was made that this is result is not yet sufficient. The analysis of the test results found that potential faults are found most prominently in the first lag. This follows from the high average P-values in this lag. With this knowledge a second test will be conducted which is focused on the first lag. Also the significance level will be lowered to 0.01 from 0.05 to increase the possibility of finding new potential faults. Furthermore, only taking the earlier lags in the second test and retaining the same (or higher) F1-score will also result in an reduction of computation time without cost of performance.

Results of second test

Table 5.8 shows the change in setting and the new results. only lags 1 and 2 are taken into account. Figure 5.16 shows the dataset again, now with the improved test settings applied.

With the changes made to the test parameters, the $F1$-score has increased to 97% (0.97). The only remaining undetected faults observed at $19:01$ and $19:30$ in figure 5.15. At $19:01$ the decrease of the significance level did not result in the detection of the fault. since the length of the test was not altered it again meant that the fault at $19:30$ did not occur in the scope of any of the F-tests and was therefore, again not detected. Nonetheless, the increase of performance concludes that the second test is more successful.

Length of the test

In test one and two the total length of the F-test is 60 samples. A trade off exists in this regard: The more samples are taken in a single test the lower the resulting $P - values$ are overall. Because the anomaly in any cases is only a fragment of the total test, the longer the test the smaller the relative part of this anomaly. On the other hand, the shorter the test, the higher the chance an anomaly is not found at all. Also, shortening the test has the increased chance of reducing the F-statistic to a degree that F-tests are made invalid.

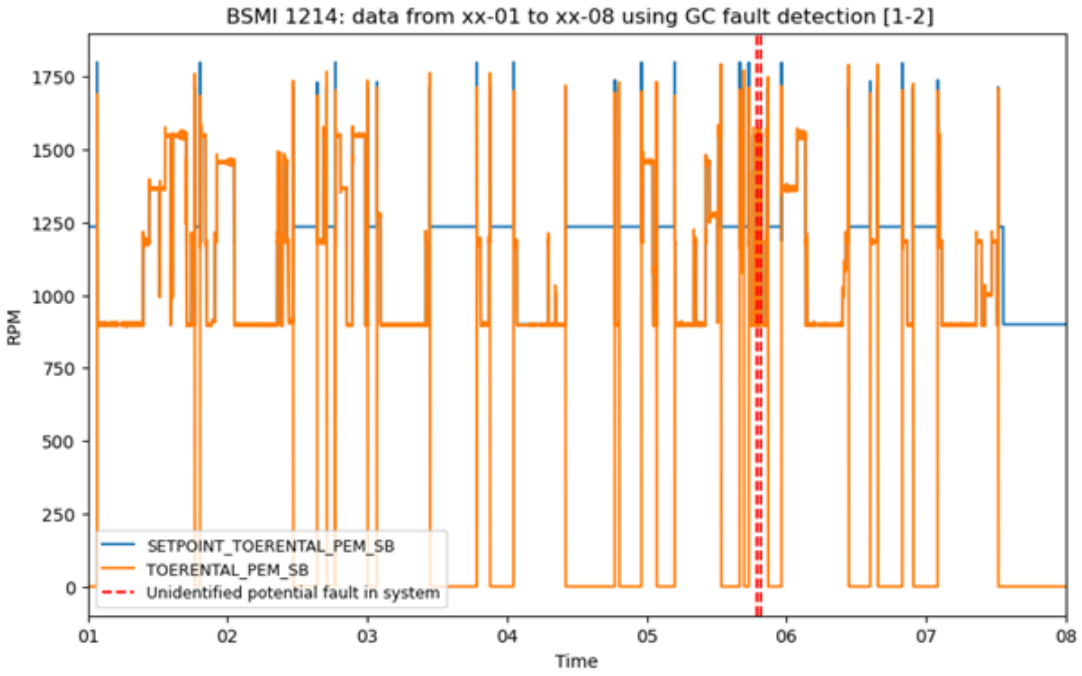


FIGURE 5.16: Results of the second fault detection test for application 1 showing only unidentified faults

TABLE 5.8: Settings, results and evaluation metrics from test 2 in application 1

Fault detection: BSMI 1214 Setpoint RPM to RPM (test 2)			
Test setting		Value	
Resample rate (s)		5	
Data points before set-point change (n)		30	
Data points after set-point change (n)		30	
Total length of test (n, s)		60, 300 seconds	
Filter		no	
Lags used (j)		[1, 2]	
Significance level (-)		0.01	
Test results	Total	Potential faults	Set-point 0 errors
P_{visual}	40	8	32
P_{F-test}	38	6	32
N_{visual}	137	-	-
N_{F-test}	139	2	-
Confusion matrix	result	Evaluation metrics	result
TP	38	Accuracy, A	0.98
TN	137	Precision, P	1
FP	0	Recall, R	0,95
FN	2	F1-score, F1	0.97

Conclusions: Application 1

The following conclusions may be drawn from the first application of fault detection.

1. Fault detection is very much possible in the chosen dataset using the initial settings found in the exploratory analysis. As expected, the strong causal relationship meant that losses of GC could almost all be discovered.

2. Further insight in resulting P-values showed that potential faults where most visible in the first time lag. The probability of concluding no causal relationship was highest there. This led to the conclusion that selection of lags is possible as the test's performance was improved while decreasing computing time.

3. Faults that are undetected initially are characterized by output values regaining causality quickly at the moment of a set-point change.

4. The reduction of the significance level could improve the overall performance of the fault detection test.

5.2.3 Application 2: BSMI 1214, set-point RPM to POWER

In the second application, data from BSMI 1214 is again taken. Figure 5.17 shows the data used in the second application. Contrary to the first test conducted in the first application, a considerable number of faults are left undetected, While most 0 RPM errors are detected a number of these errors are not found my the detection method. Moreover, only one potential fault is found correctly. Two are marked falsely leaving seven undetected faults. All results are seen in figure 5.17. Table 5.9 shows the test settings and results.

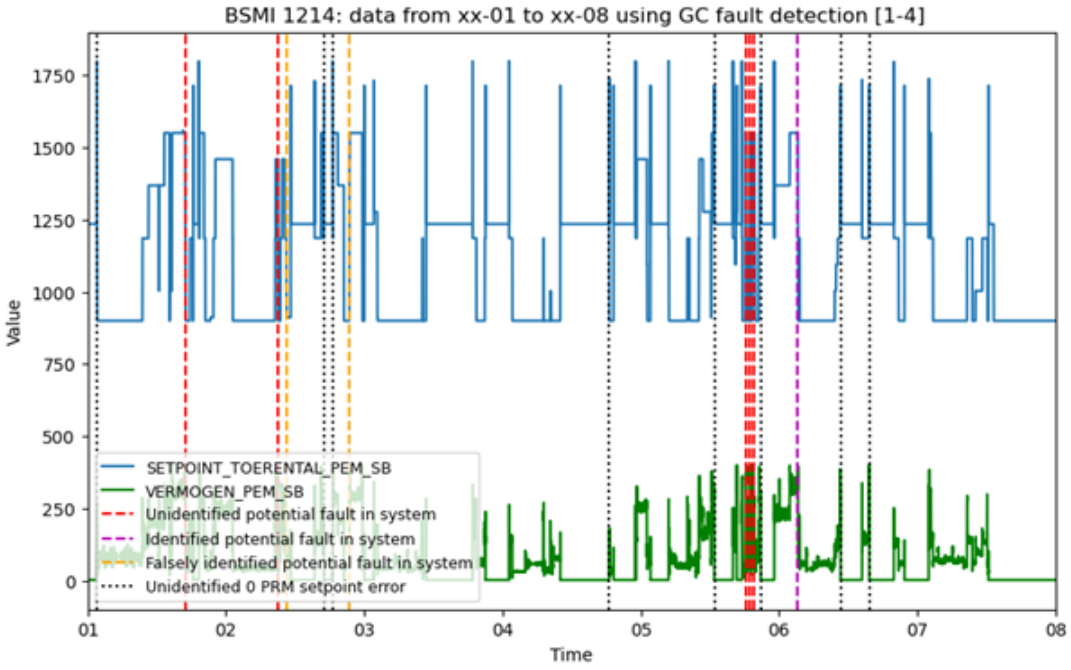


FIGURE 5.17: Results of the first fault detection test for application 2

Test settings and results

Two settings where chosen differently from the first application. Both the number of data points taken for the test before and after a set-point change where increased. Also, a low pass filter was applied to the power output data. The reasoning for this is that upon visual inspection of the data it was seen that the power parameter was much more unstable and less overall dependent on the input parameter. The added length and application of a filter, would, in theory prevent the unintentional marking of potential faults caused by the aforementioned variance of the power data.

With an overall $F1$ score of 78% (0.78) it can quickly be concluded that the test is less effective than the ones conducted in the first application. The loss in performance can be attributed mostly by the low recall score of 68% (0.68).

TABLE 5.9: Settings, results and evaluation metrics from test 1 in application 2

Fault detection: BSMI 1214 Setpoint RPM to Power (test 1)			
Test setting		Value	
Resample rate (s)		5	
Data points before set-point change (n)		40	
Data points after set-point change (n)		40	
Total length of test (n, s)		80, 400 seconds	
Filter		yes, LPF	
Lags used (j)		[1, 2, 3, 4]	
Significance level (-)		0.05	
Test results	Total	Potential faults	Set-point 0 errors
P_{visual}	40	8	32
P_{F-test}	29	3(1)	26
N_{visual}	137	-	-
N_{F-test}	150	7	6
Confusion matrix	result	Evaluation metrics	result
TP	27	Accuracy, A	0.92
TN	135	Precision, P	0.93
FP	2	Recall, R	0.68
FN	13	F1-score, F1	0.78

Analysis of falsely and unidentified losses of GC

The current performance levels are considered insufficient. Therefore, a second, adjusted test will be set up. To correctly change the settings of the test more insight must be gained on the precise reason for failure of correct detection. Figures 5.18 and 5.19 show two, very different, moments where the detection method has failed.

Figure 5.18 shows an undetected 0 RPM set-point error found around 15 : 40 on day six. The power output generally follows the input parameter until about 15 : 39. Then, the power output is seen to decrease. Afterwards is spikes shortly before falling to zero. The RPM input stalls at around 1250. Apart from the decrease in power output seen prior to the short spike it can be concluded that actually the power output does seem to be Granger caused by the RPM set-point. Before falling to zero, a small spike is seen almost

identical to the set-point input, also the fall of power output is just as pronounced as the fall in input RPM. Over the entirety of the test area causality is visually present almost the entire time except for a small moment before the visually perceived error.

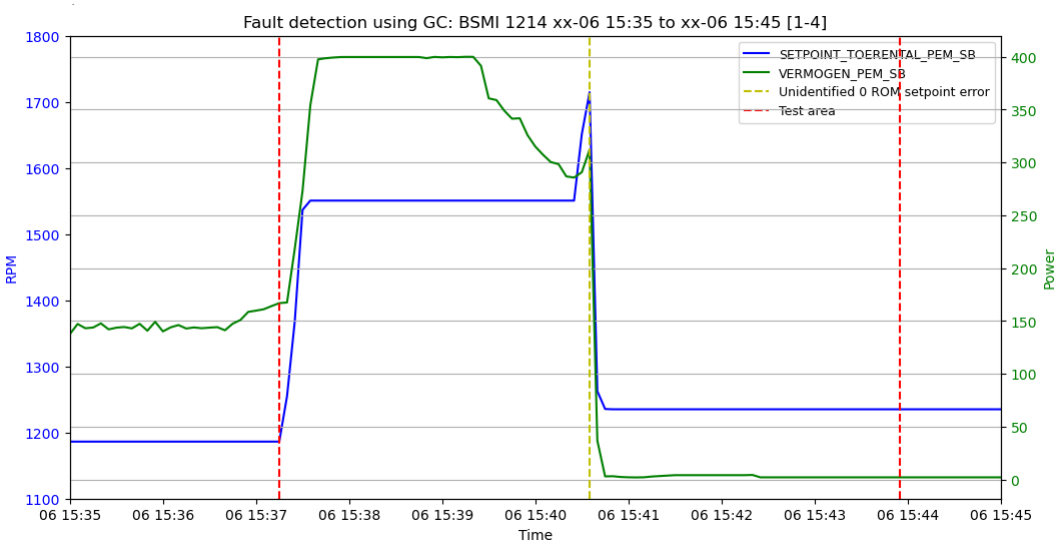


FIGURE 5.18: Zoomed in section of the data showing an undetected RPM-0 error

In the second case, in figure 5.19 a potential fault was marked at 02 21 : 27, visually, no record of an potential fault is made. Upon inspection it was concluded that indeed no fault is found here. As is notable with the power output in the entire dataset, values tend to to take longer to find a new equilibrium after a set-point change. This is seen between 10 : 23 to 10 : 27 where power output falls considerably after a very substantial increase of RPM input. As the test area shows, This reaction of the power output is taken into account by the F-test while the increase of RPM input is falls outside the boundary. Possibly, reducing the length of the test will reduce this phenomenon.

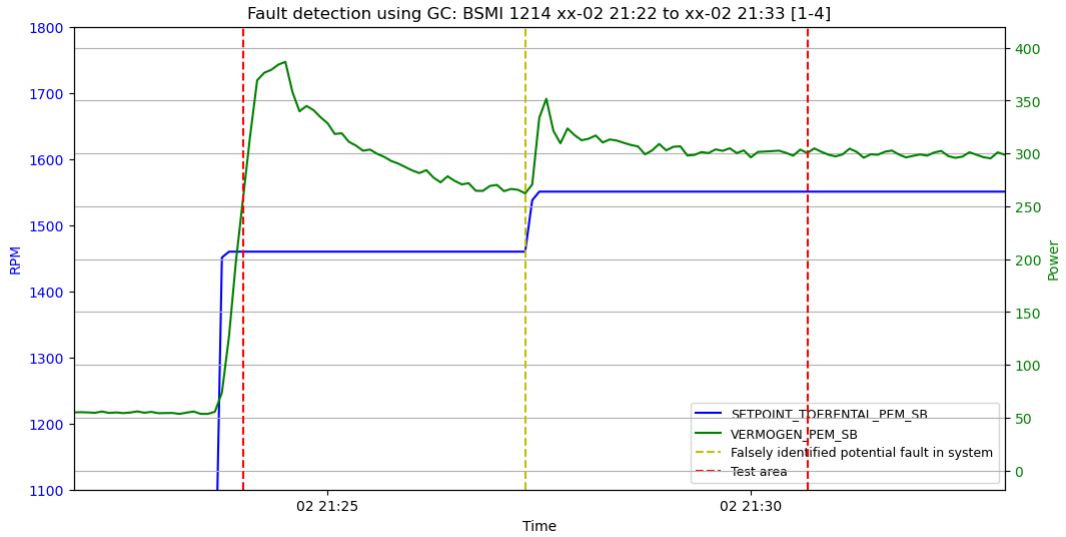


FIGURE 5.19: A section of the data showing a falsely detected potential fault.

Analysis of resulting P-values

Just like in application one, an analysis of resulting P-values is needed to improve performance of a second test. Table 5.10 shows that again, potential faults are detected mainly because of high P-values found in the first lag. This time however, the same is the case for the set-point 0 errors. The resulting P-values for the undetected error at 06 15 : 40 are almost all zero except for in lag three where it is minimal nonetheless. The F-test is thus unable to find this error which is not very surprising as figure 5.18 showed that generally the behaviour looks to be causal. In the second test, the significance level is again lowered to 0.01 which could improve detection for other cases. At this moment, the assumption is that the reduction in significance level will not result in detection of the error.

Upon further inspection of the data the conclusion is made that the power parameter reacts differently to a change in set-point as opposed to the RPM output parameter and thus likely will not show GC in the first lag as the power output parameter rises or falls to an equilibrium after an set-point change. This is the conclusion made from the falsely identified fault in figure 5.19. Therefore, lag one is removed from the test and a fifth lag is added, thereby shifting the window of the analysis one lag.

Table 5.10 also shows an undetected fault at 02 03 : 19 which was (correclty) detected in application one. The highest resulting P-value is found in lag four which might indicate that anomalies in the data are found at places where the power output is unable to find an equilibrium at a longer distance from the set-point change.

TABLE 5.10: average P-values per type of detection in application 2

Average value	Lag (j)			
Type loss of GC	1	2	3	4
Set-point 0 error	0.5788	0.3633	0.3463	0.2257
06 15:40	0.0000	0.0000	0.0002	0.0000
Potential fault	0.1967	0.0115	0.0053	0.0080
02 21:27 (falsely detected)	0.4973	0.0000	0.0000	0.0000
06 03:19 (undetected)	0.0264	0.0083	0.0256	0.0470

Application of adjustments on second test

The following changes in test setting where applied for the second test: The F-test was applied to lag two to five instead of one to four shifting it one lag. The length of the test was decreased to a total number of data-points again being 60 and the significance level was again decreased to 0.01.

With the application of adjustments on the F-test settings an improved result may be seen in table 5.11. Also figure 5.20 shows the results. Still, an considerable number of potential faults are not detected by this test. However, none are detected falsely anymore following from the shortening of the test. Also the amount of undetected 0-RPM errors has fallen.

Overall, the $F1$-score of the test was increased from 78% (0.78) to 88% (0.88). This is mostly resulting from the increase of true positives (TP) going from 27 to 31. Also,

precision was increased again to 100% (1) showing that the test no longer results in false detection of loss in GC.

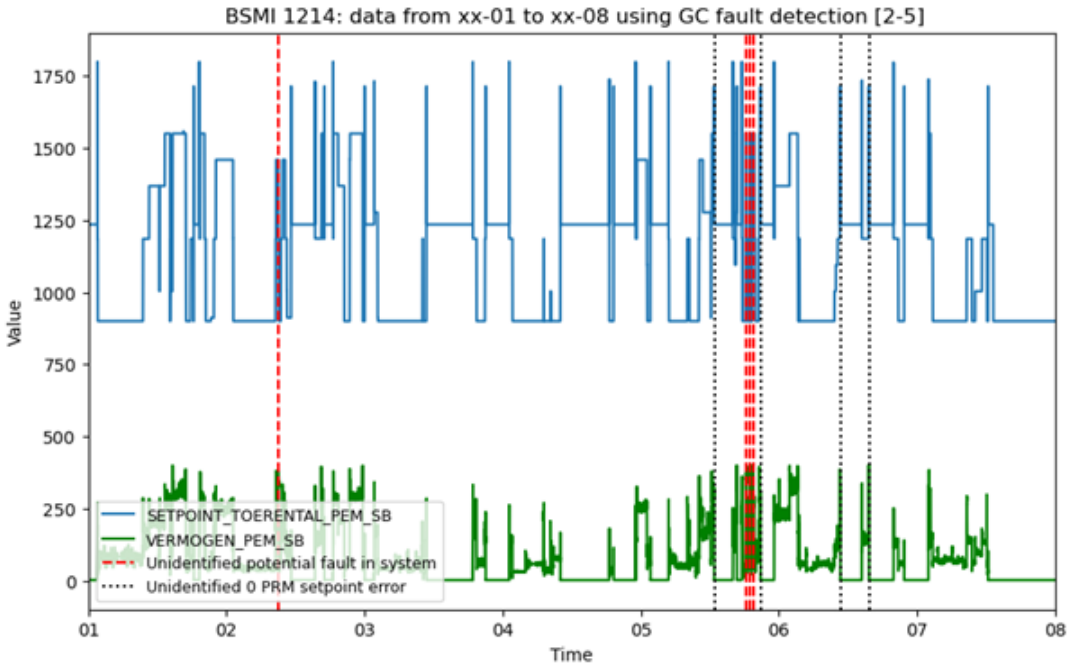


FIGURE 5.20: Results of the second fault detection test for application 2

TABLE 5.11: Settings, results and evaluation metrics from test 2 in application 2

Fault detection: BSMI 1214 Setpoint RPM to Power (test 2)			
Test setting		Value	
Resample rate (s)		5	
Data points before set-point change (n)		30	
Data points after set-point change (n)		30	
Total length of test (n, s)		60, 300 seconds	
Filter		yes, LPF	
Lags used (j)		[2, 3, 4, 5]	
Significance level (-)		0.01	
Test results	Total	Potential faults	Set-point 0 errors
P_{visual}	40	8	32
P_{F-test}	31	3	28
N_{visual}	137	-	-
N_{F-test}	146	5	4
Confusion matrix	result	Evaluation metrics	result
TP	31	Accuracy, A	0.95
TN	137	Precision, P	1
FP	0	Recall, R	0.78
FN	9	F1-score, F1	0.88

Conclusions: Application 2

Again, some conclusions may be drawn from the second application:

1. A reduction of performance in the fault detection test is concluded in the second application, the reason for this is most likely the weaker link between the in- and output parameters. The power parameter showed an tendency to take longer to reach and equilibrium after an set-point change with an visual spike of power output when for instance, the RPM set-point was increased.

2. The removal of a lag of the analysis (number one) improved the precision of the test. Due to the less strong link between the two parameters a higher degree of non-GC was found in this lag, even at moments of normal functioning.

3. The length of the test was taken to be longer in the first test but proved adverse in localizing faults in the system. It resulted in taking data 'out-of-context' on its fringes.

4. New selection of lags improved performance as well as the reduction of test length. A filter was applied in both tests which prevented the false detection of more faults in the system.

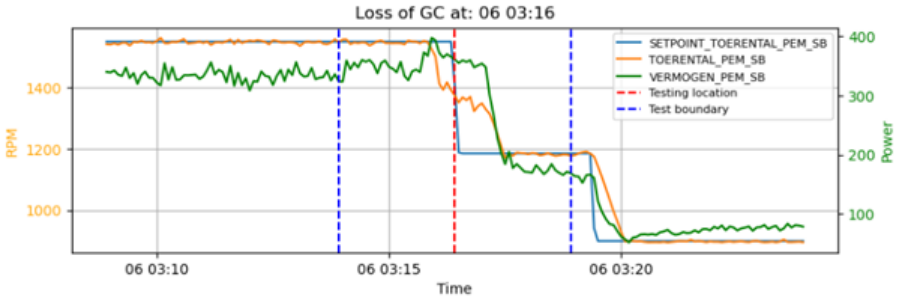
5.2.4 Similarities between undetected faults and 0 RPM errors in application 1 and 2

Now that both applications of the fault detection method have been concluded, a comparison between the detected and undetected faults is made to show where the F-test, and more broadly fault detection using GC succeeds and falls short. In figure 5.21 shows three cases where the three parameters in analysed are shown at locations of visually detected faults. The test boundaries of the F-test are shown (blue lines) at the right location (red line). Figure 5.21a and 5.21b show that faults are detected in both applications. Here, it can be concluded that loss of GC is mostly detected by an a-typical reaction at the location of a set-point change. In figure 5.21a both parameters are already decreasing and show only a late response to decrease in set-point. This is the same case for the second figure. At the set-point increase both output parameters react at a much slower rate.

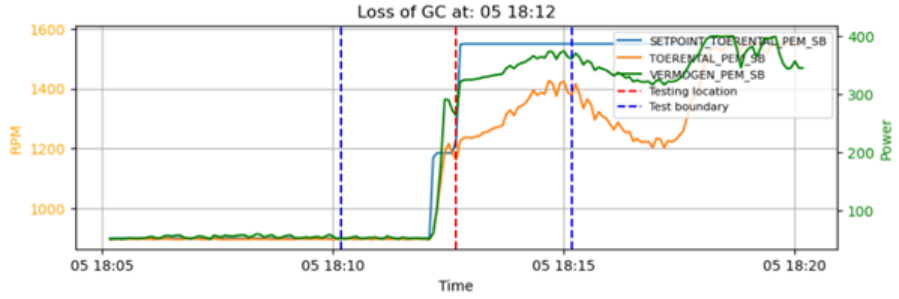
In the third case (figure 5.21c) the change of set-point causes both parameters to react correctly. The power output even follows the short dip in RPM input. This location was marked however, since both parameters show unusual behaviour between $18:56$ and $19:01$. even tough this is shown in the test's boundary to a considerable extent still no loss of GC is found. Therefore, the following can be stated: If the output variables react correctly to an change in input, no loss of GC will be registered, even when prior to the set-point change a-typical behaviour is shown.

This effect may come from both the structure of the regression model used to test GC as well as the type of data used as the input parameter. Because the F-test measures the improvement of prediction between the univariate vs the bivariate autogression model, having a static parameter such as the set-point RPM parameter means that the error remains static, moreover does not increase. During the test, this means that even tough, the output RPM and power show an anomaly, if the input parameter remains static during that time, the F-test will not show any improvement nor degradation of the prediction. Then, when a set-point change does occur and the output parameters do show a Granger

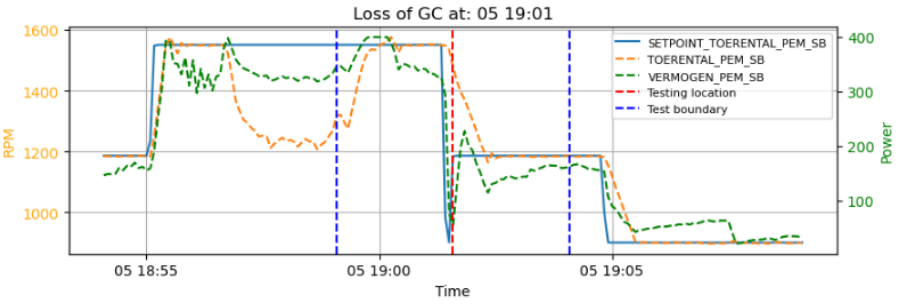
causal relationship as expected, overall GC is concluded and the anomaly before the set-point change remains undetected. F-statistics are low in this cases which does show that the strength of the GC is not strong. Still the null-hypothesis is rejected.



(A) Fault detected in both applications



(B) Fault detected in both applications



(C) Fault detected by neither applications

FIGURE 5.21: Three examples of identified faults either detected (figure 5.21a, 5.21b or not detected (5.21c) by the fault detection method

5.2.5 Application 3: BSMI 1214, RPM from/to POWER

The third and final application is done between output RPM and output POWER. The goal of this application is to verify that GC between two output variables is most likely not present. The correlation matrix showed a strong relationship between the two parameters. Moreover, the GC matrix showed bidirectional GC. The assumption here is that a common-cause issue exists. A single test is done on the data in both directions of testing. Both with the RPM parameter as input and again as output. Below in table 5.12 both these tests are shown. No set-point 0 errors can be detected as the set-point is not tested in this application.

F-tests are again only conducted at the moment of a set-point change. This is done to make sure results can be compared to the visual analysis of the dataset as well as to simplify the exact locations of where the F-tests must take place.

Looking at the results in table 5.12, both setting and results from the tests (in both directions) are shown. The overall conclusion is that performance of the test is very low. By sheer luck, almost all faults are found by both tests which result in high recall values. Nonetheless, resulting $F1$-scores are low resulting from hte very loww precision of the tests.

TABLE 5.12: Settings, results and evaluation metrics from test 1 in application 3

Fault detection: BSMI 1214 RPM from/to Power			
Test setting		Value	
Resample rate (s)		5	
Data points before set-point change (n)		30	
Data points after set-point change (n)		30	
Total length of test (n, s)		60, 300 seconds	
Filter		no	
Lags used (j)		[1,2, 3, 4, 5]	
Significance level (-)		0.05	
Test results	Total	Potential faults	Set-point 0 errors
P_{Visual}	40	8	32
$P_{RPM \rightarrow Power}$	161	88	78
$P_{Power \rightarrow RPM}$	143	76	67
N_{Visual}	137	-	-
$N_{RPM \rightarrow Power}$	16	2	-
$N_{Power \rightarrow RPM}$	34	3	-
RPM→ Power			
Confusion matrix	result	Evaluation metrics	result
TP	38	Accuracy, A	0.26
TN	14	Precision, P	0.23
FP	123	Recall, R	0.95
FN	2	F1-score, F1	0.37
Power → RPM			
Confusion matrix	result	Evaluation metrics	result
TP	37	Accuracy, A	0.39
TN	42	Precision, P	0.27
FP	95	Recall, R	0.93
FN	3	F1-score, F1	0.41

Figure 5.22 shows the relationships between the three analysed parameters. The figure shows that as an input variable, GC exists from the set-point RPM variable to the output RPM and power variables while the other way around this is not found. Hence the one-way pointing arrows. Between the output power and RPM variables correlation exists. Naturally in both directions. However, no GC is present here as both are caused by the same input parameter. This example shows how GC can be applied in a mechanical environment and can successfully show the relationships between variables.

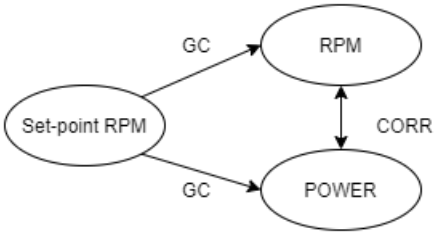


FIGURE 5.22: Concluded relationships between variables in BSMI 1214

5.2.6 Linking SAP events to the results of the fault detection method

The fault detection method has found a number of cases where the in- and output parameters have found to be functioning in an atypical manner. Mostly around the fifth day where, using the test results from application 1, three faults where detected. To find out if these cases of loss of GC have actually led to a maintenance action on the ship, the SAP database is consulted. One entry within a week after the dataset a entry is made concerning the inability of the system to react to changes in input settings. Although insufficient information is given on the particular dates and components involved, it can be concluded that a link can be made between the events and the faults detected by the method. Next to this, the possible link also demonstrates that datasets can indeed be selected using the selection method in chapter 3, however, more information is again needed on the SAP event to find out to what extent this is true.

Chapter 6

Conclusions, Discussion & Recommendations

Following the results, conclusions, a discussion and recommendations will be given. The conclusions aim to answer the identified sub- and main research question. A discussion will put forward the scientific relevance and achievements of the thesis. Finally, recommendations are given for further research as well as for application of the study in the RNLN.

6.1 Conclusions

As the RNLN undergoes a large modernisation effort the need for new SM techniques with the goal of improving the availability of the new generation of ships. A part of its fleet consists of the so-called OPV, which is a vessel used in a wide range of applications. This thesis's main research question therefore addresses the functionality of one technique;

To what extent can Granger causality be applied on available data from current oceangoing patrol vessels at the Royal Netherlands Navy with the goal of implementing smart maintenance?

Before conclusions can be drawn on this question itself, first, all five sub-questions are addressed:

What data is currently gathered and stored on the oceangoing patrol vessels? An onboard system called the IPMS is used to manage operations. Data logged during operations may also be saved in one central area using this system. A large number of sensors are outfitted on the ship providing enough data for further analyses. Using a coding system called BSMI sensors are labeled based on their system code. Every code, or system, comprises of a number of sensory output but may also contain input values.
Maintenance actions are registered in the SAP. Faults in a particular system are registered here by the crew. SAP information comprises of documentation during failure, however, adjustments to systems are also recorded. The BSMI system is again used for the coding. Data is stored on hard-drives onshore. Using an in-house created application called DINO datasets can be extracted for further analysis.

How can possible use-cases to test Granger causality be selected on the ocean-going patrol vessels based on their maintenance history and sensor data availability?
SAP and IPMS data was collected for all BSMI codes on the one of the RNLN's OPV's over an extended period of time. An newly defined index was created including, sensor as well as maintenance data which guided the selection of datasets to analyse. A timeline highlighting crucial areas was created as well showing periods with high number of corrective maintenance actions. The three use-cases selected where BSMI 1471, 1214 and 1331 as the simple, intermediate and complex use-cases respectively.

How is the Granger causality test applied in this data environment and what variables are involved in its application?
Resulting from the selection of use-cases a dataset was analysed on GC, namely BSMI code 1214 which is part of the drive train. Proving to be rich in data this was the only analysed use-case. For the GC test to be applied correctly, the following has to be taken into consideration:

1. Two time-series have to be selected of which one is a input and the other the output. It should be clear before testing what relation is expected, thus which is which. This will determine whether the GC test is applied successfully as it allows the testing of GC in both directions verifying the result.

2. Selection of lags. GC is based on the prediction of values based on previous values in a time-series. Selection of the number and location of these lags is vital in the analysis and is based on the perceived relationship between the input and output sensor.

3. When testing a longer dataset on GC, bidirectional GC was found in many cases. The GC test is therefore not suitable to test on a large dataset. Its use can be demonstrated when tests are done on a smaller scale with a minimum of 20 data-points. However, in practice a standard of 60 points was taken for each test to ensure statistical significance.

Testing GC done using F-tests. The F-test calculates the reduction of error predicting values of one time-series when added previous values of the same and another time-series. Here the initial time-series is the perceived output sensor time-series and the added is the perceived input sensor time-series. Two measures determine the GC relationship:

1. The P-value: As the F-test is based on a null-hypothesis that describes no Granger causal relation is present, a probability value expressed in percentage which determines whether to accept or reject the hypothesis. A significance level must be picked which determines this. If the result is lower than the significance level the hypothesis may be rejected and GC is concluded.

2. The F-statistic: This is the measure of reduction of error when calculating the next value in the time-series of a bivariate regression model as opposed to the univariate regression model. It is based on the respective resulting errors, length of the test (number of data-points) and number of time-lags, or delays, taken into account. The higher the F-statistic the stronger the reduction of error which means the Granger causal relationship is stronger.

In the applications put forward in this thesis the following testing method is proposed:

1. A in- and output sensor parameter is chosen from a dataset, visually, anomalies are found between these parameters based on the assumed relationship. In the discussed applications, GC is assumed at all times.

2. A input variable is used as a marking-point for the F-test. If the input value changes a F-test is conduced over a set number of data-points before and after the input change. As said, 60 data-point were taken as a default.

3. A range of lags are selected for the first test based on visual inspection of data. Results of the test will indicate whether a change of lag-selection is needed.

4. In normal working behaviour GC is assumed. Therefore, if the result of a F-test results in a P-value higher than the chosen significance level indicating no Granger causal relationship, an anomaly in the dataset is marked.

5. The method is repeated over a longer period of time resulting in a list of F-tests performed at each change in input value showing either GC is, or is not present at that particular moment in time.

6. Performance of the tests are expressed in accuracy, precision, recall, and $F1$-scores.

What requirements have to be met in a dataset in order to successfully conduct a Granger causality test?
When starting the eventual analysis. It was concluded that in order to start meaningful testing a list of requirements in the dataset would have to be met:

1. Identifiable in- and output sensors: simple use-cases did not contain input sensor data which prevented testing in those systems.

2. All sensor parameters should start and end at the same point in time as the method needs two equally sized time-series.

3. The length of the to be analysed time-series should be the same as well. If length varies between sensors they may be resized. This thesis used simple forward filling method for the set-point input parameter as only changes in values were logged.

4. This thesis used a constant, set-point based sensor parameter as the input for its F-tests. This made it easy to crease smaller F-tests at set-point changes to detect faults. However, as discussed in the results chapter also had its downsides. A continuously logged input parameter is preferred over the set-point based input as the regression model functions better with determining the error over time in this environment.

5. When using a continuously logged input parameter, F-test may be done periodically instead of set-point change centered. More research is needed on the actual application of a testing method like this.

6. Periods of use: Selected datasets may contain large portions of idle-time providing little data to test the theory with. The selected dateset used in this thesis were relatively long extending for a number of months including sometimes periods of weeks comprising of idle-time. New datasets could be selected based on times of use instead of just on the selected date to date.

7. To perform one test only a short time-series is needed. As discussed this could start from only 20 data-points depending on the contents of the data. However, to demonstrate the fault detection capabilities, more tests are needed in one dataset. This thesis used a week long dataset resulting in 177 tests performed over that period. Therefore, the length of the dataset should not be expressed in time. Also, the thesis used downsampling to reduce computing time per F-test meaning that the 60 data-point F-tests performed generally in this thesis were actually 300 seconds of time as data was resampled to $0.2\,Hz$.

With what performance can faults in the system be detected using Granger causality?
Three applications where looked at within the BSMI 1214 case. The first application tested GC from input to output RPM,the second tested input RPM to output power and a third tested output RPM to and from power output. The following is concluded:

1. 32 accounts of a IPMS logging error were marked by visually where in input RPM parameter stalls at times of non-operation. At 8 moments potential faults were detected visually of which 5 occured in close succession.

2. As expected, application one performed best with a $F1$-score of 0.97. In- and output RPM were expected to show a strong Granger causal relationship hence the high performance of the test. More importantly no false identification of problems in the dataset were made.

3. Application two was successful as well, a $F1$-score of 0.88 showed that between RPM input and power output a Granger causal relationship may also be found, although, less strongly so. A number of faults were left undetected, but again, the fault detection method marked no faults at wrong locations.

4. The third application showed that between the output sensor parameters no GC exists. A low average performance of 0.39 is found on average showing that although, the two parameters correlate strongly with one-another, they are not Granger causal.

5. Unidentified faults in the dataset are attributed to the fact that the constant values found in the input RPM parameter result in the F-test's inability to show non-Granger causal behaviour at times.

6. Looking back at the dataset selection, a SAP event was recorded mentioning the exact behaviour detected by the fault detection system. While promising, more research is needed to definitively prove whether this means that the detection system has found an actual problem in the system.

Overall, the conclusion is made that, depending on the link between in- and output parameter, faults may be detected at a high rate.

Following from the conclusions made for all sub-questions, a conclusion can be made for the main research question.

To what extent can Granger causality be applied on available data from current oceangoing patrol vessels at the Royal Netherlands Navy with the goal of implementing smart maintenance?

The RNLN OPV's contain the right data gathering and storage infrastructure to visualise data in various systems. The available data can be tested on GC successfully using a created method testing an identified in- and output parameter in a system. The testing method is able to identify anomalies in the data where a direct link between these parameters is defined with an performance of 0.97 to 0.88 depending on the relationship between the in- and output proving that, in combination of the known subsequent SAP event, SM could be implemented using GC.

6.2 Discussion

This thesis has achieved to study the feasibility of GC based SM for the RNLN. In doing so, a number of discussion points are elaborated on further.

Use-case selection
The thesis was able to pick use-case datasets for further analysis. With the use of SAP records and the sensor data coming from the DINO application. Next to this, all sensors where listed based on their type and number which improves the overall selection of new datasets.

Prove of non-GC for fault detection
In other research, the goal of the study was to prove GC existed between two time-series. In this paper the same was done with the addition of detecting when this was no longer the case. This method of finding the loss of the Granger causal relationship could be considered a novel application of the GC method as a means of anomaly detection.

Exclusion of lag selection method
Applications of GC testing are based on visual inspection of data. During the exploratory segment of the results chapter a in-dept view of RPM data was used to determine what lags are used in the analysis. The choice of selecting lags like this was made over a method like the AIC method described in the literature study because the thought was that it would further the understanding of the data and the GC method. If AIC was applied without having extensive experience using the F-test, a multitude of mistakes in its application could be made without proper reasoning behind them.

Creation of test setup
Because the input parameter used was constant with only the changes in set-point being monitored, this called for a testing method based around these set-point changes. Initially the thought was that, like in similar research, a test would be executed over period leading to insights in GC between the two parameters. In this case, F-test had to be applied at very localized areas as the constant input value would otherwise distort the outcomes of the F-test.

Application of GC monitoring over time
If a non-constant input paramerter were selected it would allow for the F-test to be conducted periodically. In this way a test could be done every x minutes resulting in a constant

evaluation of the Granger causal relationship over time. Again with the assumption that GC should be present at all times between the in- and output parameter, a loss of GC may be detected at the moment it happens. This thesis had picked the constant set-point based input parameter which constrained it from applying this theory. It therefore had to rely on the localized F-tests instead.

6.3 Recommendations

6.3.1 Further research

Application on more parameters
In the current application GC tests were conducted between Input RPM and output RPM and power. Another output parameter is temperature. GC testing could be applied here as well.

Application on more datasets/BSMI codes
In extension of the first point, the research could be extended by analysing more datasets in the same BSMI code or by analysis new parameters in another BSMI code. An extra analysis of the Granger causal relationships between parameters in another year could prove the use of GC as a means of anomaly detection further. The addition of testing the same types of parameters in a different BSMI code may also prove the versatility and usefulness of the method. For instance, a BSMI code containing RPM and power output data could be tested upon to show GC there as well.

Extension of lag selection
In order to create a more generic method of testing GC in other data environments in the RNLN, the application of a lag selection method as mentioned in the literature study should be considered.

Indentification of failure indicators-sensors
At this point no link has been made between failure modes and sensors within the systems on the OPV. With the use of the SAP records in addition to further research on the systems' components, indicator sensors could by identified. This would result in a similar approach as used in a study put forward in the literature study [25].

Analysis of Periodic GC testing
For testing in larger datasets the choice was made to test GC at every set-point change of the input parameter. If, in a different system no such constant input-parameter is present it could be wiser to apply a GC test periodically. For instance, if the input parameter is the power output of a drive-train system and the output parameter is the temperature a test could be performed every ten minutes to verify that the Granger causal relationship is still correct.

Creation of a multivariate regression model
At this stage only a bivariate regression model was used of the fault detection method. With the creation of a multivariate regression model, more than one input parameter could be used to predict the output parameter. To implement this, a different, most likely complex, system is needed which included these multiple input parameters.

6.3.2 Recommendations for the RNLN

Improvement SAP records
To aid in further research the contents of the SAP record could be expanded. More documentation should be provided to correctly determine the cause of failure after it has happened. Examples of improvements could be: Reason of replacement of a certain part, documentation of damage endured by the system, exact date of first detection or progression of failure. This added information could lead to a more focused GC analysis. Also, datasets may be selected more precisely as more insight is available in the failures involved.

Further case-studies
The current applications are all conducted in the BSMI 1214 system. The RNLN could apply the GC testing method on a number of other systems. The systems are already identified and ranked by this study.

Data management on the OPV's
When extracting data from the DINO application the following additions to the data management system could be made:

1. Visualisation of idle-time on-board. To more conveniently select datasets for further inspection it could be useful to highlight moments of use of the ship over time. Data is available from 2016 onwards, however, large parts of data are filled with 'empty' data.

2. Frequency of logging: The GC test are generally conducted on data which was downsampled from 1 Hz to 0.2 Hz. All of the parameters involved are logged at 1 Hz which, in case of temperature sensors, for example is very frequent. To safe storage space and improve processing speed of analyses this could be addressed.

3. Creation of timelines involving SAP data: The selection of use-cases based on maintenance actions put forward in this thesis is already applied to all BSMI codes on the OPV. The timelines combining including the SAP data is only created for the selected systems. Timelines such as these are found to be an useful tool in picking datasets for further analyses. The DINO application could benefit from visualisation of SAP data where the entirety of SAP documentation is listed in the timeline as well.

Bibliography

[1] K. S. Anil. Granger causality. *Scholarpedia*, 2:1667, 2007. URL: https://api.semanticscholar.org/CorpusID:27480653, doi:doi:10.4249/scholarpedia.1667.

[2] Y. Chen, D. Li, Y. Li, X. Ma, and J. Wei. Use moving average filter to reduce noises in wearable ppg during continuous monitoring. *International Summit on eHealth*, 181:193–203, 06 2016. doi:10.1007/978-3-319-49655-9_26.

[3] Modernisering voor 4 patrouilleschepen marine. URL: https://www.defensie.nl/actueel/nieuws/2020/01/01/modernisering-voor-4-patrouilleschepen-marine.

[4] Patrouillleschip (opv). URL: https://www.defensie.nl/organisatie/marine/materieel/schepen/patrouilleschepen.

[5] Voorkomen efficiënter dan genezen. URL: https://magazines.defensie.nl/allehens/2019/07/06_data_gestuurd-onderhoud.

[6] How to interpret a confusion matrix for a machine learning model. URL: https://www.evidentlyai.com/classification-metrics/confusion-matrix.

[7] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. URL: http://www.jstor.org/stable/1912791.

[8] preventive maintenance vs predictive maintenance. URL: https://www.hanarasoft.com/preventive-maintenance-vs-predictive-maintenance/.

[9] Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988. URL: https://www.taylorfrancis.com/books/mono/10.4324/9780203771587/statistical-power-analysis-behavioral-sciences-jacob-cohen, doi:10.4324/9780203771587.

[10] Frost J. Understanding significance levels in statistics. URL: https://statisticsbyjim.com/hypothesis-testing/significance-levels/.

[11] C. Jacinto, , D. López, I. Aguilera-Martos, D. García-Gil, I. Markova, M. Garcia-Barzana, M. Arias-Rodil, J. Luengo, and F. Herrera. Anomaly detection in predictive maintenance: A new evaluation framework for temporal unsupervised anomaly detection algorithms. *Neurocomputing*, 462:440–452, 2021.

[12] V. J. Jimenez, N. Bouhmala, and A. H. Gausdal. Developing a predictive maintenance model for vessel machinery. *Journal of Ocean Engineering and Science*, 5(4):358–386, 2020.

[13] J. Krinitz, S. Alfano, and D. Neumann. How the market can detect its own mispricing -a news sentiment index to detect irrational exuberance. *Hawaii International Conference on System Sciences*, pages 1412–1421, 01 2017. doi:10.24251/HICSS.2017.170.

[14] I. Lapin. Using sensor data for the development of digital twins in support of condition-based maintenance. In *Conference Proceedings of iSCSS*, 2022.

[15] V. Liew. Which lag selection criteria should we employ? *Economics Bulletin*, 3:1–9, 01 2004.

[16] M. Maziarz. A review of the granger-causality fallacy. *The Journal of Philosophical Economics*, 8(2):6, 2015. URL: https://EconPapers.repec.org/RePEc:bus:jphile:v:8:y:2015:i:2:n:6.

[17] P. Nunes, J. Santos, and E. Rocha. Challenges in predictive maintenance – a review. *CIRP Journal of Manufacturing Science and Technology*, 40:53–67, 2023. URL: https://www.sciencedirect.com/science/article/pii/S1755581722001742, doi:10.1016/j.cirpj.2022.11.004.

[18] Y. Qiu, Y. Feng, P. Tavner, P. Richardson, G. Erdos, and B. Chen. Wind turbine scada alarm analysis for improving reliability. *Wind Energy*, 15(8):951–966, 2012. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/we.513, doi:10.1002/we.513.

[19] C. Rijsdijk, A. da Silveira, N. Nale, and T. Tinga. Using ship sensor data to achieve smart maintenance? In *Proceedings of the International Naval Engineering Conference 2020*, pages 1–10, United Kingdom, 2020. Institute of Marine Engineering, Science and Technology. International Naval Engineering Conference, INEC 2020. URL: https://www.imarest.org/events/inec-2020, doi:10.24868/issn.2515-818X.2020.047.

[20] J. M. Rohrer. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1):27–42, 2018. arXiv:https://doi.org/10.1177/2515245917745629, doi:10.1177/2515245917745629.

[21] M. Sherif, L. Sang-Heon, D. Jantanee, C. Nicholas, and H. Soltan. Lean thinking for a maintenance process. *Production and Manufacturing Research: An Open Access Journal*, 3:236–272, 01 2015. doi:10.1080/21693277.2015.1074124.

[22] A. Shojaie and Emily B. Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9(1):289–319, 2022. URL: https://www.annualreviews.org/doi/pdf/10.1146/annurev-statistics-040120-010930, doi:10.1146/annurev-statistics-040120-010930.

[23] A. Shojaie and G. Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):16–23, 2010. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2935442/, doi:10.1093/bioinformatics/btq377.

[24] B. Spencer, B. Khaddaj, S. Luong, T. Mitchell, and Z. Erdenebaatar. Surveillance and investigation of faults on canadian frigates with pecan. *DRDC and RDDC*, pages 1–7, 2022.

[25] B. Spencer and S. Luong. Influence between sensors on halifax-class frigates. *DRDC and RDDC*, 2020.

[26] P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004. doi:10.1109/MSP.2004.1311138.

[27] W. W. Tiddens, J. Braaksma, and T. Tinga. Decision framework for predictive maintenance method selection. *Applied Sciences*, 13:3, 02 2023. doi:10.3390/app13032021.

[28] Transitie koninklijke marine. URL: http://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2023Z08715&did=2023D20829.

[29] B. Uzair and S. Noralfishah. The impact of sustainability practices on share performance with mediation of board members experience: A study on malaysian listed companies. *International Journal of Financial Studies*, 11:4, 12 2022. doi:10.3390/ijfs11010004.

[30] J. van Driel, C. N. L. Olivers, and J. J. Fahrenfort. High-pass filtering artifacts in multivariate classification of neural time series data. *Journal of Neuroscience Methods*, 352:2, 2021. URL: https://www.sciencedirect.com/science/article/pii/S0165027021000157, doi:10.1016/j.jneumeth.2021.109080.

[31] E. VanDerHorn, Z. Wang, and S. Mahadevan. Towards a digital twin approach for vessel-specific fatigue damage monitoring and prognosis. *Reliability Engineering & System Safety*, 219:108222, 2022. URL: https://www.sciencedirect.com/science/article/pii/S0951832021007006, doi:10.1016/j.ress.2021.108222.

[32] Finnish defence forces: Reference case. URL: https://wetech.fi/references/finnish-defence-forces/.

[33] W. Wettayaprasit, N. Laosen, and S. Chevakidagarn. Data filtering technique for neural networks forecasting. In *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, pages 225–230, 2007.

[34] S. Xiao-Sheng, W. Wenbin, H. Chang-Hua, and Z. Dong-Hua. Remaining useful life estimation – a review on the statistical data driven approaches. *European Journal of Operational Research*, 213(1):1–14, 2011. URL: https://www.sciencedirect.com/science/article/pii/S0377221710007903, doi:10.1016/j.ejor.2010.11.018.

[35] R. Yacouby and D. Axman. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 79–91. Association for Computational Linguistics, 2020. URL: https://aclanthology.org/2020.eval4nlp-1.9, doi:10.18653/v1/2020.eval4nlp-1.9.

[36] I. Zaniletti, R. L. Dirk, D. G. Lewallen, D. J. Berry, and H. Maradit-Kremers. How to distinguish correlation from causation in orthopaedic research. *The Journal of Arthroplasty*, 38(4):634–637, 2023. URL: https://www.sciencedirect.com/science/article/pii/S0883540322010403, doi:10.1016/j.arth.2022.11.019.

[37] P. Zheyuan, H. Pan, L. Cheng, P. Chong, F. Xiangcheng, and L. Honglai. Data-driven prediction of product yields and control framework of hydrocracking unit. *Chemical Engineering Science*, 283:119386, 2024. URL: https://www.sciencedirect.com/science/article/pii/S0009250923009429, doi:10.1016/j.ces.2023.119386.

[38] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Appendix A

SAP events recorded at the RNLN

Melding soort	Omschrijving	Ordersoort	Omschrijving	Subordersoort	Omschrijving	Subordermeldingsoort
M1	OH-Aanvraag: correctief onderhoud uitstelbaar. Melding bevat ILS informatie maar uitvoering hiervan is niet direct nodig. Opdrachten worden vaak in AM periode afgehandeld of doorgeschoven naar BO.	**PM01**	Correctief onderhoud	**SO01**	Correctief onderhoud	**SO**
M2	Storingsmelding: Correctief onderhoud is niet uitstelbaar. Melding bevat ILS informatie en uitvoering hiervan is direct benodigd. Heeft impact op staande planning en moet direct worden uitgevoerd.	**PM01**	Correctief onderhoud	**SO01**	Correctief onderhoud	**SO**
M3	Activiteitenmelding: Correctief onderhoud alleen vastlegging. Melding bevat ILS informatie en uitvoering is niet nodig. Melding wordt alleen lokaal gemaakt voor analyse doeleinden.					
M4	Preventief onderhoud uit PO plan. Equipments kennen een PO strategie en PO plan. Tellerstanden of draaiuren triggeren het starten van een PO taak.	**PM02**	Preventief onderhoud	**SO02**	Preventief onderhoud	**SO**
M5	ILS melding: Aanvraag ILS producten. Een aanvraag voor een ILS product kan overal vandaan komen. Van een schip uit de DMO kant maar ook een uit een werkplaats. Een product kan worden gezien als een stafadvies, modificatie of tekening etc.	**PM07**	ILS	**PM08**	ILS Suborder	**SO**
M6	Overig onderhoud kan voortkomen uit werkaanvraag CZSK maar ook uit het werken voor tweeden, derden en " Eigen Bedrijf". Om de verschillende soorten onderhoud te identificeren, wordt een projectnr. In Sap aangemaakt en aan de melding en order gehangen. Het projectnr. heeft een relatie met Primavera en heeft een eigen bedachte conventie.	**PM05**	Overig onderhoud	**SO05**	Overig onderhoud	**SO**
M7	Modificatie opdracht. Opdracht uit exploitatie. Een product uit het ILS proces. Melding en order moeten gescheiden blijven van overige ordersoorten i.v.m. de juiste financiële en administratieve afwikkeling.	**PM03**	Modificatie	**SO03**	Modificatie	**SO**
M8	Preventief onderhoud direct. Preventief onderhoud handmatige aanvraag. Preventief onderhoud wat uitgevoerd wordt voordat PO plan triggert of dat er alleen een handmatige vraag is.	**PM09**	Preventief OH direct	**SO02**	Preventief OH direct	**SO**

FIGURE A.1: All types of events recorded at the RNLN

M9	Modificatie invest. Modificatie uit staffels. Een product uit het ILS proces. Melding en order moeten gescheiden blijven van overige ordersoorten i.v.m. de juiste financiële en administratieve afwikkeling.	**PM10**	Def Investeringsorder	**SO10**	Investering	**SO**
WP	Werkpakket melding. Melding die wordt gebruikt voor clustering van orders of het overzetten van deelactiviteiten uit Primavera naar SAP. Hierbij wordt ook een revisienummer gebruikt om verschillende meldingen bij elkaar te houden. Daarnaast wordt gebruik gemaakt van een projectnummer om het project te kunnen identificeren (M6).	*	inhoudsafhankelijk	*	inhoudsafhankelijk	**SO**
Z1	uitbouw Z1 melding wordt gecreëerd na uitbouwen equipment.	**PM04**	Reparatieorder	**SO04**	Reparatieorder	**SO**
Z3	OH-beh. Component: Wordt gecreëerd door ASM t.b.v. herstel componenten	**PM04**	Reparatieorder	**SO04**	Reparatieorder	**SO**
Z1	Uitbouw: Z1 melding wordt gecreëerd na uitbouwen equipment.	**SO04**	Reparatieorder	**SO**	Reparatieorder	**SO**
Z3	OH.beh.Component: Z3 melding wordt gecreëerd t.b.v. revisie componenten buiten ASM om, wordt gebruikt i.c.m. SO04 bij gekoppeld componentenonderhoud.	**SO04**	Reparatieorder	**SO**	Reparatieorder	**SO**

FIGURE A.2: All types of events recorded at the RNLN

Appendix B

List of sensors in BSMI codes

M1+2	I	BSMI	Definition	S	Dru	Temp	Toer	Kopp	Verm	Niv	Vocht	Freq	Stro	Span	Uren	Hoev	Verbr	rest	Types (P)
24	2.52	12129	KVDM INST. BRAND-STOFINST.	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1
29	0.68	1517	WATERMIST. BRAND-BLUSINST.	4	0	0	0	0	0	1	0	0	0	0	0	0	0	2	1
1	0.58	1525	BELADINGS COMPUT-ERINST.	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
36	0.41	1471	DRINKWATER INST.	3	1	0	0	0	0	0	0	0	0	0	2	2	2	5	3
21	0.20	1452	STUURMACHINE INST.	4	4	0	0	0	0	0	0	0	0	0	0	0	0	4	1
3	0.18	1645	DIEPGANGMEET INST.	6	2	0	0	0	0	0	0	0	0	0	0	4	2	1	3
15	0.17	1522	LENSINST.	3	0	0	0	0	0	1	0	0	0	0	0	0	2	1	1
3	0.14	2411	INFORMATIE RETRANSMISSIE INST.	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1
7	0.12	1532	AFVOERINST.	8	0	0	0	0	0	2	0	0	0	0	0	4	2	1	3
16	0.10	1413	VERWARMINGS INST.	4	0	0	0	0	0	0	0	0	0	0	0	0	0	3	
7	0.09	1584	VUILE OLIE TRANSPORT- EN AFGIFTEINST.	6	0	0	0	0	0	2	0	0	0	0	0	4	2	1	
2	0.04	1352	OMZETTING EN VERDELING 24V/28V	3	0	0	0	0	0	0	0	0	0	3	0	0	0	3	1
2	0.03	1213	VOORTSTU. REGEL- EN AU-TOMATIEKINST.	8	0	0	2	0	0	0	0	0	0	0	0	6	2	1	3
5	0.02	1582	SMEEROLIELAAD-TRANSP-AFGIFTEINST.	9	0	0	0	0	0	3	0	0	0	0	0	6	0	9	2

TABLE B.1: List of BSMI codes selected for the simple case with all types of sensors involved

M1+2	I	BSMI	Definition	S	Dru	Temp	Toer	Kopp	Verm	Niv	Vocht	Freq	Stro	Span	Uren	Hoev	Verbr	rest	Types (P)
29	0,26	1641	PLATFORM MANAGEMENT INST.	16	4	5	0	0	0	0	5	0	0	0	0	0	2	14	3
13	0,21	1214	VOORTSTUWINGS ELEKTOMOTOR INST.	19	0	2	4	0	6	0	0	0	0	0	0	0	4	12	3
32	0,15	1511	ZEEWATERBRAND BLUS INST.	18	8	0	0	0	0	0	0	0	5	0	0	0	0	13	2
31	0,13	1591	ZEEKOELWATER INST.	12	3	2	0	0	0	0	0	0	0	0	0	0	3	5	2
18	0,11	1575	HELIKOPTER BRANDSTOF INST.	11	0	0	0	0	0	3	0	0	0	0	0	0	3	3	6
52	0,11	1561	ZOETWATER BEREIDINGS INST.	21	0	0	0	0	0	6	0	0	0	3	0	0	6	9	2
30	0,11	1414	VENTILATIE- EN LUCHTBE-HANDELINGS INST.	18	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
17	0,05	1592	ZOETKOELWATER INST.	12	3	6	0	0	0	1	0	0	0	0	0	0	2	10	3
15	0,04	1492	KOEL- EN VRIES INST.	10	0	5	0	0	0	0	0	0	0	0	0	0	5	5	1
7	0,04	1415	KOUDWATER MAKER INST.	10	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
6	0,04	1314	NOODDIESEL GENERATOR INST.	11	0	5	0	0	1	0	0	1	1	0	0	0	1	9	5

TABLE B.2: List of BSMI codes selected for the intermediate case with all types of sensors involved

M1+2	I	BSMI	Definition	S	Dru	Temp	Toer	Kopp	Verm	Niv	Vocht	Freq	Stro	Span	Uren	Hoev	Verbr	rest	Types (P)
28	0,30	1521	TRIM-, BALLAST- EN ONTBALLAST INST.	46	4	0	0	0	0	12	0	0	0	0	0	0	27	16	3
13	0,21	1331	VERDELING HOOFDVOEDING 440V	36	0	0	0	12	4	0	0	12	7	12	0	0	12	47	5
10	0,08	1571	BRANDSTOFLAAD-30 , TRANSP- EN AFGIFTE INST	0	0	0	0	0	0	9	0	0	0	0	0	18	9	1	3
12	0,02	1231	TANDWIELKAST INST.	40	10	22	2	4	0	0	0	0	0	0	0	2	38	4	36
7	0,01	12121	KVDM INST KRUISVAART DIESELMOTOR	122	20	82	6	0	2	0	0	0	0	0	0	10	110	4	6

TABLE B.3: List of BSMI codes selected for the complex case with all types of sensors involved

Appendix C

Timelines of events in BSMI codes

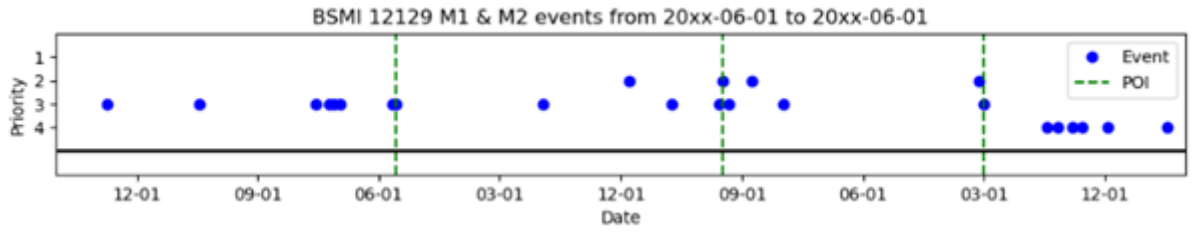


FIGURE C.1: Timeline of BSMI 12129 showing events and points of interest

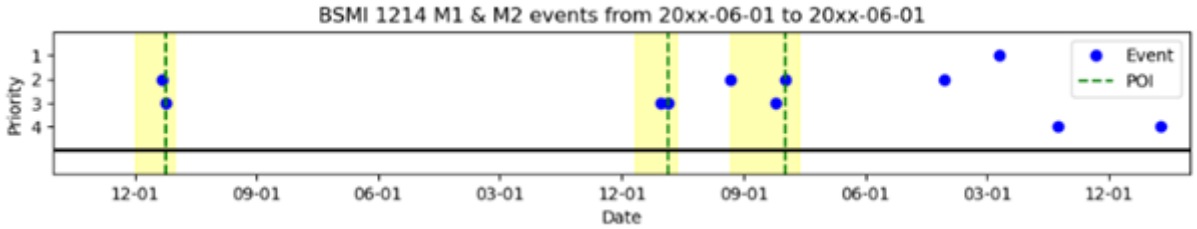


FIGURE C.2: Timeline of BSMI 1214 showing events and points of interest with highlighted locations of datasets

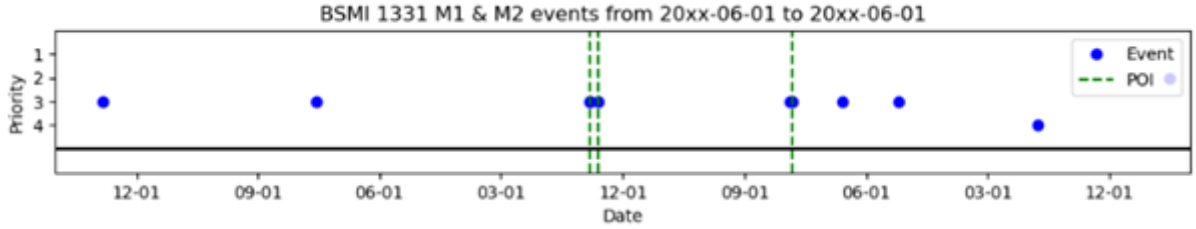


FIGURE C.3: Timeline of BSMI 1331 showing events and points of interest

Appendix D

All visually found potential faults

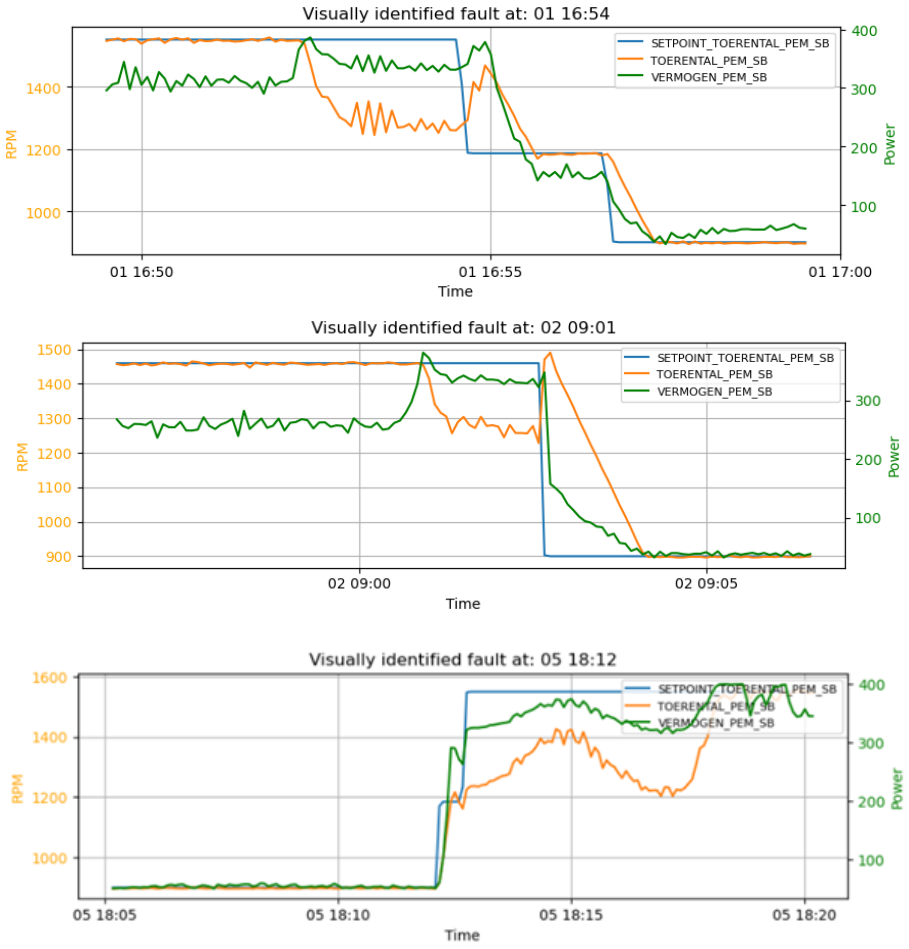


FIGURE D.1: All visually identified faults in the dataset.

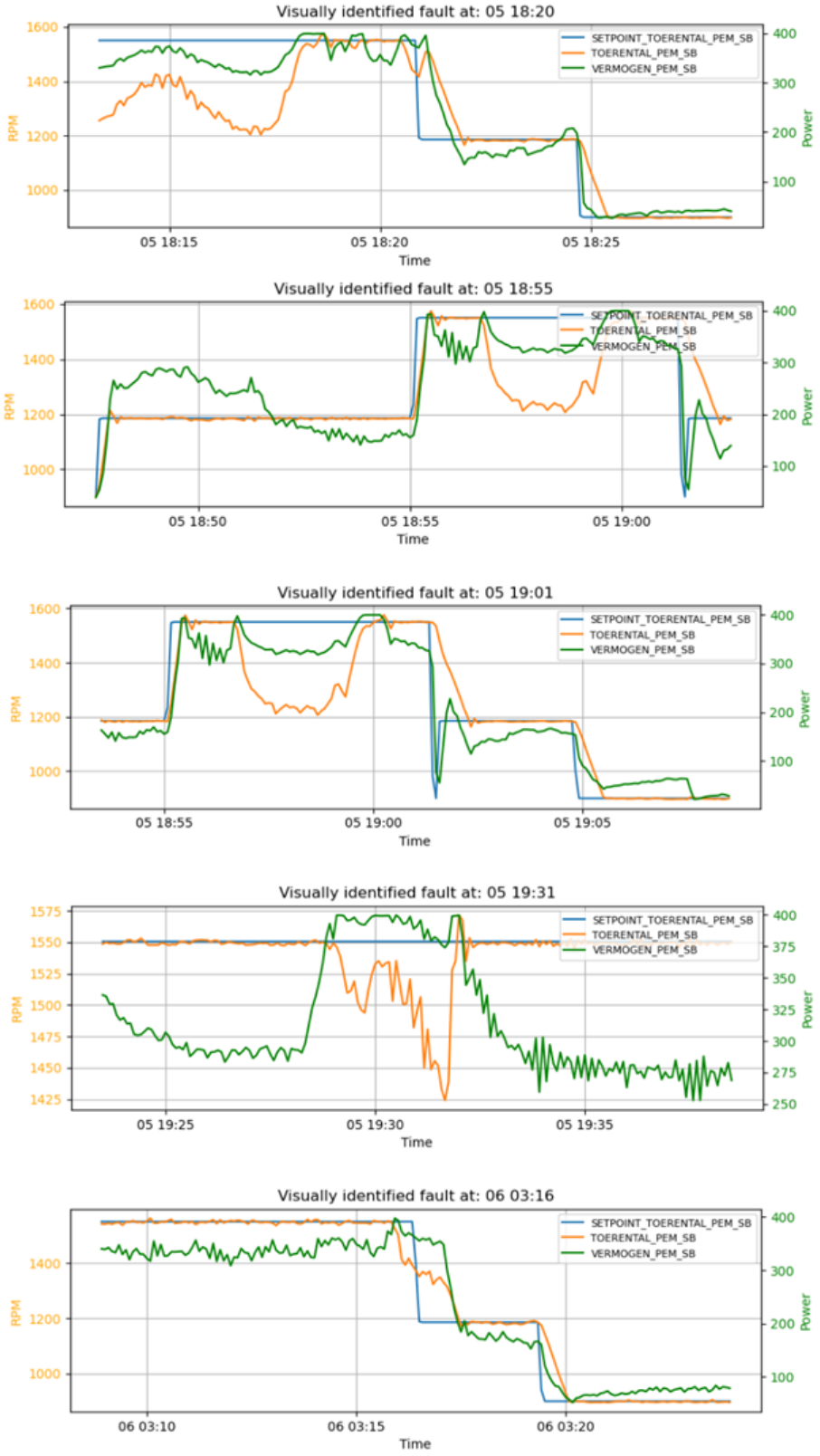


FIGURE D.2: All visually identified faults in the dataset.