Detecting abnormal behavior in lithography machines

Dassen, B.

Award date:
2019

Eindhoven University of Technology

Master Thesis

Detecting Abnormal Behavior in Lithography Machines

Author:
B. Dassen

Supervisors:
A. Di Bucchianico (TU/e)
L. Troisi (ASML)
S. Schepens (ASML)

Industrial and Applied Mathematics
March 24, 2019
Eindhoven

Abstract

ASML produces lithography systems for the semiconductor industry, which are used for the production of chips or integrated circuits. If a failure occurs in such a complex system it is difficult to determine the cause of this failure. The Customer Support Diagnostics Team has as current aim to solve these failures as fast as possible. This is because a failure results in an unscheduled down. Downtime in such a machine is extremely costly. The nominal costs for customer can be $20 per second of unscheduled downtime. In the future they want to use predictive maintenance, where failures are detected before they occur. This could reduce the unexpected costs drastically. Therefore it is needed to understand complex systems in which thousands of signals are measured, such that a distinction can be made between normal and abnormal behavior.

In this thesis the WELLE sticker failures are investigated, these failures only occur in the TwinScan NXT. For this purpose both the measurements and the corrections of the SyCo data are used. This data is transformed to a lower dimension using Principal Component Analysis. This is done for different types of classes, which are based on the type and the measurements of the machines. On this new data self-starting Statistical Process Control methods are applied. These methods are able to detect changes in the data. It turns out that some of these changes occur before a calibration. This means that the proposed method is able to detect changes in the machine for which in the current system a calibration is used to reverse these changes.

Acknowledgments

I would like to thank several people who where involved in this project. I would like to thank my supervisor from the TU/e, Alessandro Di Bucchianico for his guidance and support during the entire project. Especially for providing structure and helping me keeping the project organized.

I would also like to thank Luca Troisi and Sander Schepens for their weekly advice and support during the project, which encouraged me to keep going on and to improve the project. I am grateful that I could develop my presentation skills at ASML and that I learned to communicate with several people the entire project. I would like to thank the Customer Support department of ASML for providing me with the opportunity to conduct this research.

Finally, I would like to thank Claudia Tegelaers for her love, support and interest.

Bart Dassen
March, 2019

Contents

Chapter 1

Introduction

1.1 Introduction to ASML

ASML researches, develops, designs, manufactures, markets and services lithography systems for the semiconductor industry. ASML was founded in 1984 by Philips and Advanced Semiconductor Materials International (ASMI). The machines produced by ASML produce micro chips or integrated circuits for different kinds of electronic devices. ASML has developed different kind of machines, in this research we will focus on the TWINSCAN NXT. First we will explain the general idea of lithography system and the TWINSCAN NXT then we discuss some problems when diagnosing this machine.

Lithography System: TWINSCAN

Making an integrated circuit requires several technological steps. Lithography is one of these steps and requires both high resolution to realize extremely small features and high precision to properly place each layer with respect to the others. Integrated circuits are printed on wafers, which are round thin slices made of silicon. Out of one wafer, hundreds of dies can be obtained. Each of these dies is one integrated circuit. A processed wafer can be seen in Figure 1.1. The pattern that is printed on the wafer will determine the functionality of the chip. This pattern is in the scales of nanometers. To achieve this precision ASML introduced the TWINSCAN platform. This platform uses two stages: the measurement stage and the exposure stage. These stages can be seen in Figure 1.2. In the measurement stage, all kind of characteristics are measured, for example the position of the wafer. This information is used in the exposure stage, where the wafer is exposed to light. The wafer is exposed by light projected through a mask called the reticle. The reticle determines the pattern that is printed on the wafer. This pattern has to be projected very precisely to ensure correct behavior of the chips. To get the pattern as precise as possible on the wafer optics are used. The wafer is coated with a light-sensitive chemical before exposure. When the light is projected onto this chemical the pattern that is projected is printed onto the wafer. After printing any unwanted silicon is etched away, such that the original designed 3D structure is left. Also note that not only silicon is etched, but any material which is used to make the integrated circuits (metals, oxides, etc.). The process of printing and etching is repeated until the desired pattern is printed. The margin of error in these patterns is very small. To obtain that precision ASML uses predictive algorithms and metrology systems. This means that the blueprint of a chip is optimized such that the design is printed correctly and that the system is measured and corrected in real-time.

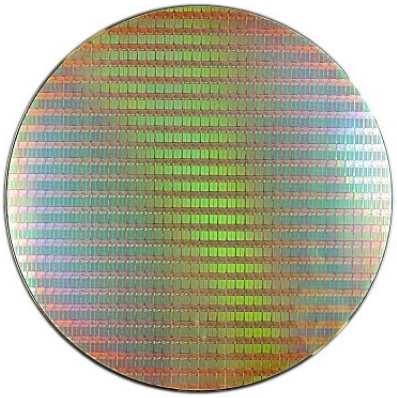


Figure 1.1: A processed wafer.

Figure 1.2: The TWINSCAN System. In blue the measurement side and green the exposure side.

Lithography System: TWINSCAN NXT

The TWINSCAN NXT uses deep ultra Violet light (DUV light). There are the following machines: TWINSCAN NXT:2000i, TWINSCAN NXT:1980Di, TWINSCAN NXT:1970Ci, TWINSCAN NXT:1965Ci and the TWINSCAN NXT:1950i. These machines are designed for different products. For example the 2000i is designed for 300-mm wafers at the sub 7-nm node, while the 1965Ci is designed for 300-mm wafers at the sub 20-nm node. The numbers in the names of the machines represents the lens that is used. This means that each of these machines is different. The machines are even more diverse because of different modular options depending on the customer needs.

Diagnostics

The department Customer Support(CS) is responsible for solving issues customers have with their machines. This has to be done as fast as possible to reduce the costs. When a machine is not working it costs customers on average $20 per second. Solving these issues fast however is not an easy task. The first reason is that the machines are very complex. They have a lot of different parts that could fail. It gets even more complex because of the many differences between machines. As already discussed there are 4 different NXT systems. These systems all have different possible updates and components depending on the customer needs. Furthermore there are a lot of settings which the customer can decide on, these settings influences how the machine works. For example certain measurements or corrections can be turned off. The current diagnosing methods are based on diagnosing after a defect has occurred. One system is based on previous solutions for very similar problems. The engineer can give feedback on this system by sharing his own solution or confirming that the proposed solution worked. There are two kinds of failures. The first failures result in a broken part such that the machine cannot operate anymore. The second failures result in a machine that can still operate but the quality of the products is not optimal. With the current diagnosing method these failures get noticed after something goes wrong. In the future ASML wants to use predictive maintenance. The advantage of predicting a failure is that it is possible to plan around it (i.e. order spare parts, rearrange production plans) to mitigate or completely remove the disruption to the productive process. Detecting and predicting failures will be the focus of this report.

1.2 Problem Description

To improve the diagnostics of the machine there is need for methods to detect defects in a part. This will be the first goal of this project. Automatically detecting if the machine is healthy or unhealthy is important for improving the diagnostics of the machines. When the cause of a failure is known it can be repaired as fast as possible. Even better would be to detect signs of a beginning failure. Such that it can be solved before it causes a serious problem. These failures could be caused by problems like drift, calibration issues or wear. To detect these failures thousands of signals are being measured in the machines. This data could be used for automated monitoring to timely detect failures of parts. To do this it is needed to understand abnormal behavior from the signals. It is needed to investigate ways to preprocess and visualize these signals. Such that a better understanding of abnormal behavior can be achieved. To make this problem more clear we answer the following questions:

- How to distinguish between normal and abnormal behavior?
- How can we detect abnormal behavior in an automated way?
- How can we predict failures due to abnormal behavior in an automated way?

To help answering these questions the following sub questions are formulated:

- How to transform the data for monitoring purposes?
- Is the approach also applicable to other machine types?
  - Applicable over time within the same machine?
  - Applicable over time in different machines?
  - Are there different signals important for different machines?
  - How do we find the important signals for a machine?

To make the goal more clear we want the following outcome:

- find a transformation for the sensor data,
- state clear specified performance of the algorithms,
- have a generalisable method w.r.t. different machines,
- be able to explain the out of control behavior of the machines.

1.3 Report Outline

After the introduction, the case about lens aberrations will be made more clear in Chapter 2. In Chapter 3 the mathematical background used for this project will be discussed. A data description is given in Chapter 4. The used approach is given in Chapter 5. The results of this approach are presented in Chapter 6. Finally we discuss the conclusion and recommendations in Chapter 7.

Chapter 2

Lens Aberrations

In the introduction we already saw that the projection needs to be very precise. A small deviation could mean the difference between a working chip and a defect chip. Therefore it is very important that the lenses are used correctly for the projection. There are systems in the machine to measure how well the projection is done and systems to adjust the lenses. In this chapter we will discuss why the lenses need to be adjusted, which systems are responsible for adjustments and what happens when a defect occurs.

2.1 Lens Problems

A perfect lens is a lens for which the incoming light waves are projected as one point on the image plane. This point is called the focal point. Aberrations cause the light waves to diverge, which causes the image to be blurred. The difference between a perfect lens and a lens with aberrations can be seen in Figure 2.1. There are several problems in the machine that causes aberrations in the lens, the main problem is lens heating. When the machine is producing wafers, energy from the laser heats the projection lenses. This changes the optical properties of the lenses. The lenses will reach a thermal equilibrium. At that point the optical properties and aberrations remain constant. When the machine stops the lenses cool down and the optical properties change again.

To solve these kind of issues the lenses are controlled by a feed-forward and feedback mechanism. When a failure happens the system can compensate for some failure. If a part is slowly breaking it might be that at the start the systems can correct this defect. After some time the part completely breaks and the system cannot correct it anymore. From the moment it starts to fail the machine is not performing optimal anymore. This should be detected as soon as possible to prevent the machine from malfunctioning.

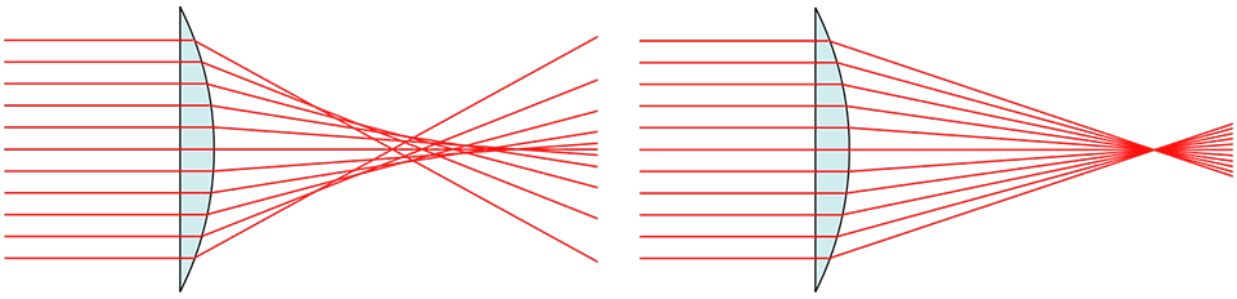


Figure 2.1: The difference between a perfect lens and a lens with aberrations from digitalphotographylive (2012).

An example of a part that could fail is the Wet Exchangeable Last Lens Element (WELLE) sticker. The WELLE is the last lens in the projection lenses. It is protected from the cooling water by the WELLE sticker. It minimizes aberration drift in the lens. This sticker could delaminate, it also suffers from hydrophobicity loss. These issues cause the projection to be inaccurate. A replacement of the WELLE sticker takes around 50 hours. The lifetime of a WELLE sticker is 2 years, but 15% of the stickers delaminate earlier.

Depending on the lens element we can adjust the position, the rotation and the behavior of that lens element. It is also possible to adjust the incoming light waves and the position of the focal point. These adjustments have some constraints, for instance the space to move a lens is limited. The time it takes to adjust each manipulator also plays an important role. This is because when the lenses are adjusted the machine is not in production. The aberration is measured via the Zernike polynomials, these will be explained in Section 3.7.1. In the next sections we will explain in more details how the projection lens works and how the projection lens is adjusted.

2.2 The Projection Lens

The projection lens contains several special lens elements to minimize aberrations. Each machine can have different kind of lenses and different software versions. The projection lens is an evolving system, so in the future new types of lenses and correction methods might be developed. There are a lot of manipulators in the machine, which are used to control different issues. These manipulators do not tell us if a part is broken, even though some manipulators(e.g. MF-EPLE) are used to control degeneration.

The lens elements can be sorted in three categories. These groups are based on the speed of the adjustment. The first group is the semi-active group, which are very slow manipulators and are only used during a lens setup. The second group is the active group which are faster manipulators and are use every lot. The last group is the scanning group, which are very fast manipulators. This group is used during production. The following lens elements are in each category:

Semi-active SAXY, HDM, WELLE

Active MF-BALE, FlexWave, APACE, MF-EPLE

Scanning SNexZ

Combining all the degrees of freedoms of the manipulators there are a lot of parameters to control. To control the lens elements we need to measure the aberrations. This will be explained in the next section.

2.3 Measurements

To use the manipulators as best as possible the aberrations of the lenses need to be known. This is expressed in terms of Zernike coefficients. These are coefficients multiplied with Zernike polynomials such that the linear combination describes the aberrations in the lenses. We denote the ith Zernike polynomial with $Z_i(x, y)$, where (x, y) are coordinates on the unit circle. A more elaborate explanation of Zernike polynomials can be found in Section 3.7.1. The full description of the aberrations uses multiple field points. A single field point is a fixed point in the field of the lens. The field contains of a grid of 13×5 field points at which measurements are done.

For each field point the aberrations are measured. This is done by ILIAS(Integrated Lens Interferometer At Scanner) and PARIS(Parallel ILIAS). PARIS is optimized for measuring the

wavefront at wafer level at several field points in a short time. ILIAS measures at one field point with optimal accuracy. For each fieldpoint it is possible to fit Zernike polynomials to describe the aberrations. Assume for one fieldpoint we get n measurements $\{x_i, y_i, v_i\}_{i=1}^n$, where (x_i, y_i) are the coordinates and v_i the aberration at that point. For these measurements we will fit a set of N Zernike polynomials. So we describe the aberrations with the function

$$w(x, y, \beta) = \sum_{i=0}^{N} \beta_i Z_i(x, y) + \varepsilon,$$

where $Z_i(x, y)$ is the ith Zernike polynomial and we assume that $\varepsilon \sim N(0, \sigma^2)$. The βs are calculated with least squares fitting. This means that we have a matrix $X \in \mathbb{R}^{n \times N}$, where $X_{ij} = Z_j(x_i, y_i)$. We want to find $\beta \in \mathbb{R}^N$ such that $V \approx X\beta$, where $V \in \mathbb{R}^n$ are the measured values. We solve

$$\min_\beta ||V - X\beta||^2,$$

which can be done for all fieldpoints. This gives a grid of 13×5 values for each Zernike coefficient. Then the averages are taken over each 5 values such that 13 averages remain for each Zernike coefficients. Through these 13 points a third order polynomial regression model is fitted with the weighted least squares method. The polynomials that are used for the fit are not orthogonal on these points. After fitting we have four coefficients for each Zernike coefficient which describe the complete aberrations in the field. The four coefficients representing the offset, tilt, curvature and third order in the field. The coefficients at each fieldpoint will be used to control the manipulators in the lens model. The $4N$ coefficients from the fitted linear models are used for diagnostic purposes.

2.4 Lens Model

The manipulators and measurements are combined in the lens models. Lens models are used to adjust the lenses using manipulators to achieve the desired optical performance. A lens model solves the following minimization problem

$$\min_x |Ax - b|^2 + |\Gamma x|^2,$$

where $A \in \mathbb{R}^{m \times n}$ contains the lens dependencies, $x \in \mathbb{R}^n$ contains the manipulator settings and $b \in \mathbb{R}^m$ are the aberrations at specified locations on the lens grid. Tikhonov regularization is used, where Γ is the corresponding diagonal matrix, this is to bound the settings of the manipulators. The value of m is dependent on the amount of Zernike coefficients. For most machines this is 35 or 63. This means that $m = 35 \cdot 5 \cdot 13 = 2275$ or $m = 63 \cdot 5 \cdot 13 = 4095$. There are three different lens models. These are

CLM The Calibration Lens Model is used for a setup of the lenses. This model is used only a few times a year. At this point the machine is not in production so all the possible manipulators can be used.

DLM The Driver Lens Model corrects aberrations measured during lot production.

SLM The Scanning Lens Model is used during exposure. It makes very fast adjustments to reduce the aberrations.

The different lens models use different manipulators. The CLM model uses all manipulators. The DLM uses the active and scanning manipulators and the SLM uses only the scanning manipulators. These lens models are used in a larger system to ensure that the lenses perform optimal.

2.5 Complete System

The complete feedback mechanism that uses the lens models and also measures the aberrations are called System Corrections(SyCo) and Lot corrections(LoCo). For SyCo the production needs to be stopped. This is because SyCo requires the lenses to be cold. Cooling the lenses takes some time, therefore the machine cannot be in production when SyCo is used. The customers of ASML can determine when they want to use SyCo. It is recommended to use SyCo every day once. LoCo is used every lot and it is possible to execute while the lenses are warm. This causes the first measurements (with a cold lens) to be different from the measurements when the lenses are warm. Both SyCo and LoCo report the $4 \cdot N$ coefficients from the fitted linear models. SyCo reports a total drift and a residual drift set. The residual drift set describes the aberrations that SyCo cannot correct. The total drift describes the situation without SyCo corrections. The data is measured periodically, which means that the SyCo and LoCo data is multivariate time series data.

2.6 Problem Description

The focus of this report will be on issues with the Wet Exchangeable Last Lens Element (WELLE) sticker. This sticker could partially detach from the WELLE, which causes the projection to fail. The goal is to timely detect issues with the WELLE sticker, such that issues can be solved before they cause trouble. For this we will use SyCo and LoCo data from different machines. This is as we described in the previous section multivariate time series data. The research questions remain the same as in Section 1.2. For this case we can set a few success criteria:

- We prefer to have a low false positive rate over a low false negative rate. Both rates should be as low as possible.
- We want to be able to quantify the accuracy with confidence intervals.
- Detect the problem in time such that there is time to react and prevent issues.
- Being able to make forecasts preferably 1 month (minimal 1 week) ahead.

Chapter 3

Theoretical Background

In Section 2.6 we have seen that in the lens case we have multivariate time series data for multiple machines. For each machine we want to detect if the machine is unhealthy. Therefore we are going to discuss some methods to detect changes in time series. We will discuss machine learning methods, Statistical Process Control (SPC) and methods to reduce the dimension of the problem. We will introduce a method to correct processes, which is called automated process control (APC). Then we will discuss the combination of SPC and APC. Finally we will explain the orthogonal polynomials in which more background for Zernike polynomials is given.

3.1 Detection

At installation the machine is assumed to be in a healthy state, but after a while the behavior changes and it could happen that the machine is in an unhealthy state. With an unhealthy state we mean that the machine is not performing optimal. It could even be that production has to be stopped. The point of this change needs to be detected as soon as possible. To do this we need to have an understanding of the machines behavior in the healthy state. An approach could be to assume that a machine is healthy at some point. Then after that point try to detect a change in behavior and verify if this is unhealthy behavior or not. This change could be a shift in the mean, a different trend or change in correlation between variables. This change could happen sudden, but it could also be a gradual change. The point at which the change happens is often called a change point. In more mathematical terms: we want to detect the point at which the underlying distribution of the process changes. Another way of seeing this is to detect an anomaly. Which is finding anomalies (outliers) compared to the usual (healthy) signal. A final definition comes from the field of control engineering, which is called Fault Detection and Isolation (FDI). In methods from this field a fault is detected and then the fault is isolated. If this is done correctly the fault can be removed. These three methods (change point detection, anomaly detection and fault detection and isolation) are very similar and are all suitable for our problem. We will discuss them in more depth below.

3.1.1 Change Point Detection

The change point detection problem tries to detect if there is a change in the observed time series and the time of this change. There is a difference in online and offline change point detection. In offline change point detection the whole sequence of gathered data is analyzed to determine the exact location of all the change points. While in online change point detection the change point needs to be detected as soon as possible. For our problem we want to start with offline change point detection to get information about the time series. This information

could be used to find out if the machine is healthy or not. The goal is to design an online change point detection method using the information discovered in the offline method.

Offline change point detection will be introduced based on Chen and Gupta (2011). Let $X_1, X_2, ..., X_n$ be a sequence of independent random vectors with probability distribution function $F_1, F_2, ..., F_n$ respectively. The problem can be written as the following hypothesis

$$H_0 : F_1 = F_2 = ... = F_n$$
$$H_1 : F_1 = ... = F_{k_1} \neq F_{k_1+1} = ... = F_{k_2} \neq F_{k_2+1} = ... = F_{k_q} \neq F_{k_q+1} = ... = F_n,$$

where $1 < k_1 < k_2 < ..., k_q < n$. Here $q > 0$ is the unknown number of change points and $k_1, k_2, ..., k_q$ are the unknown positions of the change points.

Methods that are often used for change point detection are maximum likelihood ratio test, Bayesian test, nonparametric test, stochastic process and information-theoretic approach. It is often assumed that the distribution functions $F_1, F_2, ..., F_n$ are known and that the observations are independent. There are also nonparametric methods for example in James and Matteson (2013) and Capizzi and Masarotto (2018).

3.1.2 Anomaly Detection

Anomaly detection is the problem of finding patterns which are different than the majority of the data. Anomaly detection occurs in many different fields (Chandola et al. (2009), Mehrotra et al. (2017)). In our case we want to detect anomalies in time series data, this happens often in human health (for example ecg data (Chuah and Fu (2007))) and in industries (similar to our case). It detects a pattern in the data that does not behave normal. An approach to detect abnormal behavior is to define a region which represent normal behavior. Any observation which does not belong to this normal behavior can be declared as an anomaly. Defining this normal region is a very complex task. This is because the line between normal and abnormal behavior is often not precise. For the lenses abnormal behavior could be defined as all the aberrations of the lenses that cause the chips to malfunction. Another problem in the field of anomaly detection is the availability of labeled data. It is often very time consuming to label data, because it is not always clear if something is an anomaly or not. Anomalies can be grouped into three classes (Chandola et al. (2009), Mirsky et al. (2017)):

Point anomalies This is an individual case which can be considered as an anomaly.

Contextual anomalies Given a certain context a data instance can be considered an anomaly. Each data instance can be defined by contextual attributes and behavioral attributes. So for instance a time series with temperature measurements. The time is a contextual attribute. Since a measurement of 30 °C in the winter is an anomaly while the same measurement in the summer is normal.

Collective anomalies This happens when a separate point is not necessary an anomaly but together with other points it is considered an anomaly.

Examples of methods that are used for time series anomaly detection are Markov models and ARIMA models (Mehrotra et al. (2017) Chapter 5). These models predict the future values. If the prediction deviates too much from the true value (which is observed after the prediction) it is classified as an anomaly. Another method to find anomalies is to create a database of abnormal patterns (Iturbe et al. (2017)). Then the new patterns are compared to the old patterns. If the database is not complete it does not detect all possible anomalies. Mirsky et al. (2017) uses a method to detect also unseen anomalies. This is based on dissimilarity of the data to the already seen data. Other machine learning methods used for anomaly detection

are neural networks, clustering based methods and one-class support vector machine (Goldstein and Uchida (2016)).

Another term related to anomaly detection is concept drift, which is often used in machine learning. This is the phenomenon of changing patterns and relation in data over time (Žliobaitė et al. (2016)). Models built for this kind of data become irrelevant over time. This drift also occurs in monitoring tasks and predictive maintenance. One approach of these models is a dection method to trigger a model update. An example of such a detction method are statistical change detection tests (Gama et al. (2014)).

3.1.3 Fault Detection and Isolation

Fault Detection and Isolation is applied to systems that are under engineering control. The lenses are being adjusted continuously, so this is very similar to our case. The first step in this method consists of fault detection. When a fault is found it is isolated and identified. A model-based approach could be used, which means that the next value is predicted based on a previous determined model. Example methods are filter-based apporaches such as Kalman filters (Meskin and Khorasani (2011) Introduction). It is also possible to estimate parameters and compare those to the true values of the system when it is fault free. In these methods there is a decision making tool. This is formulated very similar to the change point detection problem. Possible decision making tools include the Sequential Probability Ratio Test (SPRT), CUSUM Algorithm and Generalized Likelihood Ratio Test (Hwang et al. (2010)).

3.1.4 Comparison

For change point detection and FDI the input is time series data. FDI is often applied in control systems, where data is dependent because of the control step. While change point detection often assumes that data is independent distributed. Anomaly detection is often used much broader than the use for time series. The time series anomaly detection is very similar to the change point detection. The difference is that it does not try to find a change point, but it classifies all points after the change point as anomalies. There are some methods which are commonly used all these three methods. These methods are based on the generalized likelihood ratio test and the CUSUM algorithm, which are statistical methods. Another branch of commonly used methods are machine learning algorithms. Therefore we will further investigate machine learning algorithms and statistical process control.

3.2 Machine Learning

Machine learning is a technique which automates procedures, while making as few assumptions as possible. It is data-driven meaning that it tries to find a model based on the data presented. This means that the data set that is used needs to be large and representative. We will consider three different setups of machine learning. These are supervised, semi-supervised and unsupervised (Chandola et al. (2009), Goldstein and Uchida (2016)).

Supervised anomaly detection A fully labeled data set is available. The anomaly detection problem becomes a classification problem where the data is likely to be unbalanced. A problem that could occur is that rare anomalies might not be in the training data. This means that it is unsure how unseen anomalies will be classified.

Semi-Supervised anomaly detection In this case there is a data set available which consists of data without anomalies. The model tries to determine normal behavior, any behavior that is different is classified as an anomaly.

Unsupervised anomaly detection The data set has no labels. The algorithm tries to learn properties of the data set. Often distance and densities are used to make an estimate about normal and abnormal behavior.

In the lens case there are no labels, creating labels is time consuming and requires an expert. Another problem is that there are multiple fingerprints of unhealthy behavior, meaning that it is impossible to collect all the possible fingerprints. Therefore this case can be seen as a semi-supervised or unsupervised machine learning problem.

In supervised machine learning the data is split into three sets: a training set, test set and validation set. The training set is used to fit the models. The validation set is used for model selection. The test set is used to evaluate the chosen model (Friedman et al. (2001), Chapter 7). In unsupervised machine learning this cannot be done, because there are no labels. In unsupervised learning the validation is often done with heuristic arguments. Also the decision of the algorithm that is used is often based on heuristic arguments (Friedman et al. (2001), Chapter 14).

To understand the ideas in machine learning for anomaly detection we will discuss some methods used for the different types of machine learning.

3.2.1 Supervised Methods

Supervised machine learning in the anomaly context is about classifying measurements as an anomaly or normal behavior. For this purpose there are many classification algorithms, such as support vector machines or artificial neural networks (Goldstein and Uchida (2016)). In our case we could also see the process in which the data distribution changes over time. The is called concept drift (Gama et al. (2014)). Although the full algorithm for concept drift is a supervised learning algorithm, the detection is often done by hypothesis testing or statistical process control.

3.2.2 Semi-Supervised Methods

In semi-supervised machine learning only one class is learned. It is often called "one-class" classification (Goldstein and Uchida (2016)). This "one-class" is the class of normal behavior. Semi-supervised machine learning is often used in anomaly detection (Žliobaitė et al. (2016), Iturbe et al. (2017)). This is because anomalies are rare, while there is enough data of normal behavior. Algorithms that are used are one-class support vector machine or methods that model the probability density function of the normal behavior.

3.2.3 Unsupervised Methods

Very often unsupervised learning algorithms create some kind of clustering of the data. Grouping similar objects and finding structures within the data. This method is used, if it is not possible to determine if data is anomalie free (Žliobaitė et al. (2016), Iturbe et al. (2017)). When talking about anomaly detection, clustering could be done based on distances. Then we want a situation where the distance between the anomalies and normal behavior is large. For example the k-Nearest Neighbors algorithm. Another example is the Local Outlier Factor (LOF) (Breunig et al. (2000)). This gives a score to each point based on the distances to the nearest neighbors.

3.2.4 Comparison

A downside with machine learning is that it needs a lot of data. Even worse is that if the process dynamics change the model is often not representative anymore. This could happen when a machine is upgraded. For supervised machine learning we need to determine labels. The labeling of data is not trivial since it is unclear when an anomaly occurs. This might cause errors in the labeling which makes the model less accurate. Semi-supervised and unsupervised methods have the issue that the result is difficult to verify. This makes it difficult to get an idea of how well the algorithm performs. These are reasons to investigate statistical process control.

3.3 Statistical Process Control

In this section we will introduce the concept and some used methods within Statistical Process Control (SPC). SPC is a method to detect if a process is in control. We will distinguish between univariate SPC and multivariate SPC. The first two subsections (General SPC and Univariate SPC) are based on van Stijn (2018).

3.3.1 General Statistical Process Control

Statistical process control is traditionally used to monitor the quality of products. It was first introduced by Shewhart (Shewhart (1925)). Shewhart worked at Bell Telephone Company, where he applied his theory. For each product there was a characteristics to measure and the production happened in batches. In SPC these batches are called rational subgroups. It is desired that the production process is stable, meaning that the quality of the products is similar. An external factor can influence the process, such that the process shows unwanted behavior. To remove this external factor it is important to first detect this unwanted behavior. Traditional SPC is used to monitor the variability in a process, where the products are produced in batches. A process could be either in control or out of control.

In control: There is only common cause variation, which is variation that is natural to the process.

Out of control: There is variation which is caused by external factors.

SPC gives a signal if there is out of control behavior detected.

The primary technique used within SPC is a control chart, which visualizes the variation of the process (Montgomery (2009)). The control chart consists of a quality characteristic from the process in time, a center line (CL) and upper and lower control limits (UCL and LCL). The process is out of control if the quality characteristic falls outside of the control limits. The control chart is an advantage of SPC, because it is often easy to use and easy to interpret. In the traditional setting there is a phase I and a phase II. In phase I, the process is assumed to be in control. In this part a characteristic of the products is measured such that it can be used to estimate the control limits for phase II. In phase II the process is monitored and signals when out of control behavior is found. This is done by comparing the characteristics of each rational subgroup with the control limits derived in phase I. An example of a control chart can be seen in Figure 3.1. It can be seen that in this chart all points are within the control limits, so we can conclude that this process is in control. An out of control signal is produced if the statistic falls outside of the confidence limits. There are also more possible rules that can be used to increase the sensitivity of the control chart. An example of such a rule is: "The process is out of control if a run of eight consecutive points are on one side of the center line" (Montgomery (2009)).

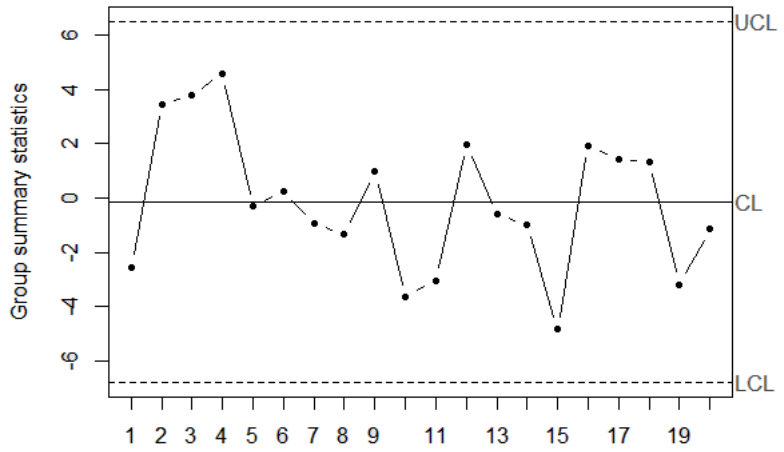


Figure 3.1: An example of a control chart, with the summary statistic and the control limits (UCL, CL and LCL)

These rules should be handled with care since they could cause false positives (Di Bucchianico (2019)).

In more modern applications of monitoring there are often no batches, so the rational subgroup size is 1 (Ferrer (2014)). This is refered to as individual measurements. Nowadays more than one quality charateristic is measured in a system. Therefore we destinguish between univariate and multivariate statistical process control. The volume of data in the world expands. This is because of the improvements on sensors and data storage. That is why there is an increased emphasis on Multivariate Statistical Process Control (Megahed F.M. (2015)). For example for the lenses in the ASML machine the data is described by a total of 124 or 252 variables. For SyCo these are measured on a daily base, while for LoCo this is done approximately every ten minutes. This gives a large data set that can be used. There are no batches in this process, so we need a method that is appropriate for individual measurements.

We will now make more precise what we mean by in control and out of control this is based on Ferrer (2014). Given a process where we have some parameters $\theta = (\theta_1, ..., \theta_n)$ from the underlying probability distribution. In the univariate case it is often assumed that the quality characteristic is normally distributed with mean μ and variance σ^2. Such that we have $\theta = (\mu, \sigma)$. The process is in control if it is likely that the measurements are from the distribution $N(\mu, \sigma^2)$ and the process is out of control if it is unlikely that the measurements are from the distribution $N(\mu, \sigma^2)$. At each time we have $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, for some initially determined θ_0. In the multivariate case with k variables, where we assume that the data is multivariate normal distributed with mean $\mu = (\mu_1, ..., \mu_k)$ and covariance matrix Σ. We have that $\theta = (\mu_1, ..., \mu_k, \sigma_1, ..., \sigma_k, \sigma_{1,2}, ..., \sigma_{(k-1),k})$. It is also possible to use more specific hypothesis. For example a change in the mean, which gives the following hypothesis

$$H_0 : \mu_1 = ... = \mu_n$$
$$H_1 : \begin{array}{lll} \mu_i & = \mu_0, & i = 1, ..., m \\ \mu_i & = \mu_0 + \delta, & i = m+1, ..., n \end{array}.$$

Another example could be to the hypothesis to detect a trend. This is given by

$$H_0 : \mu_1 = ... = \mu_n$$
$$H_1 : \begin{array}{lll} \mu_i & = \mu_0, & i = 1, ..., m \\ \mu_i & = \mu_0 + ci, & i = m+1, ..., n \end{array}.$$

These hypotheses can be tested using the Sequential Probability Ratio Test (SPRT) and the Neyman–Pearson Lemma (Tartakovsky et al. (2014)). The framework of hypothesis testing

in the statistical process context is further explained in Di Bucchianico and Van Den Heuvel (2015). For these methods the assumption is that the variables are i.i.d., this is however not always the case for example in wind turbine data (van Dalen (2018)). A solution to this problem could be to use predictive residuals or recursive residuals.

Performance measures are important for understanding how well a control chart works. The most important performance is to detect an out of control situation as fast as possible with the least amount of false alarms as possible (Tartakovsky et al. (2014), Chapter 1). This results in a trade-off between detection speed and false alarm rate. The most used performance measure is the average run length (ARL). There are two types of ARL's, the ARL_{in} indicating the time it takes until a control chart signals in an in-control situation. The other one is the ARL_{out} which is the time it takes to detect an out of control situation. Very often we want to have ARL_{out} as small as possible. There are other performance measures for example the conditional expected delay (CED), which is closely related to the ARL_{out}. Another measure is the predictive value (PV) which tells how important an alarm is when it occurs. It is important to choose the right performance metric which is dependent on the needs of the application (Frisén (2011)).

3.3.2 Univariate Statistical Process Control

We are interested in methods which work for individual measurements. Other methods can be found in Montgomery (2009) and van Stijn (2018). Self-starting methods are of interest to us, because in the lens case it is difficult to get a good estimate for μ and σ for in control behavior.

Shewhart

The first and simplest control chart is introduced by Shewhart in Shewhart (1925). This method is called the Shewhart $\overline{X}$-chart. It is designed with rational subgroups in mind, it does not work with individual measurements. The traditional Shewhart method can be modified to work for individual measurements. This method is based on the Moving Range defined by

$$\text{MR}_i = |x_i - x_{i-1}|.$$

In phase I the average of the original values $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the average of the moving range $\overline{\text{MR}} = \frac{1}{n-1}\sum_{i=2}^{n} \text{MR}_i$ are calculated. These are used for the following control limits

$$\begin{aligned} \text{UCL} &= \overline{x} + 3\frac{\overline{\text{MR}}}{d_2} \\ \text{CL} &= \overline{x} \\ \text{LCL} &= \overline{x} - 3\frac{\overline{\text{MR}}}{d_2}, \end{aligned}$$

where the constant d_2 denotes the mean of the relative range. This value depends on the way MR_i is calculated. We use consecutive values and therefore $d_2 = 1.128$ (Montgomery (2009)).

The values of the new incoming phase II data can be used in the chart, an example can be seen in Figure 3.2. For this example 25 values are generated following a standard normal distribution. Then 25 values are generated following a normal distribution with mean 5 and variance 1. The control limits are based on the first 25 values (Phase I). The 26th statistic (In phase II) immediately gives an out of control signal. So the change in mean is detected by this method. The MR chart method is good for detecting large shifts in the mean ($> 1.5\sigma$), is easy to implement and simple to use. A disadvantage is that it does not use information from the past and it only monitors the mean.

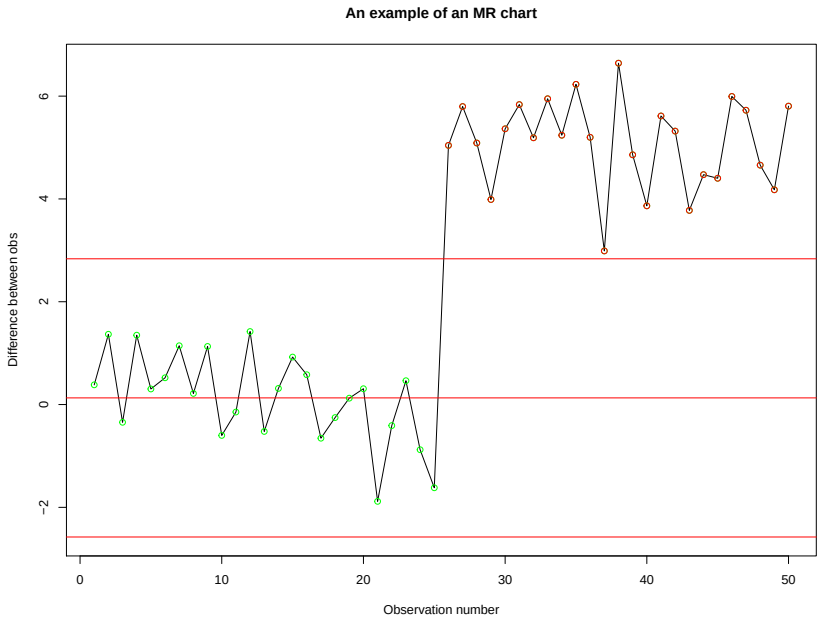


Figure 3.2: An example of a MR chart

Cumulative Sum

The second control chart that we will describe is the CUSUM chart (Page (1954)). The main disadvantage of a Shewhart chart is that it only uses the information from the last observation. This makes it difficult to detect small or gradual shifts (Hawkins and Olwell (1998)). The solution for this problem is the CUSUM chart. In this chart we monitor the values

$$C_n^+ = \max(0, C_{n-1}^+ + x_n - (\mu_0 + k))$$
$$C_n^- = \min(0, C_{n-1}^- - x_n + (\mu_0 - k)),$$

where C_n^+ is for a positive shift and C_n^- is for a negative shift. Usually k is called the reference value and is often chosen as $\frac{|\mu_0 - \mu_1|}{2}$. There is also a choice on the size of the decision interval h. Usually h is chosen as 5σ (Montgomery (2009)). We will find an out of control signal if

$$C_n^+ > h$$
$$C_n^- < h.$$

This method is good in detecting small shifts ($< 1.5\sigma$). The method is mathematically optimal and the method is effective for individual measurements. The downsides are that the desired shift to be detected must be known beforehand and the performance is sensitive to the choice of parameters h and k.

Exponential Weighted Moving Average

The last univariate method that we will discuss is the Exponential Weighted Moving Average (EWMA). This is a compromise between the Shewhart chart and the CUSUM chart. It has as advantage that it does not only rely on the last observation in the sequence. The EWMA is given by

$$z_i = \lambda x_i + (1 - \lambda) z_{i-1},$$

where $0 < \lambda < 1$ and $Z_0 = \mu_0$. If we choose $\lambda = 1$ the method only depends on the current measurement, which is a Shewhart chart. If we choose $\lambda = 0$ then it is dependent on all previous values and it is a CUSUM chart. The control limits are given by

$$\begin{aligned} \text{UCL} &= \mu_0 + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}(1-(1-\lambda)^{2i})} \\ \text{CL} &= \mu_0 \\ \text{LCL} &= \mu_0 - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}(1-(1-\lambda)^{2i})}, \end{aligned}$$

where L is the width of the control limits. The control limits are dependent on the time of the measurement, it can be seen that they converge fast for a large lambda. In Montgomery (2009) it is stated that small values of λ detect small shifts and that often $L = 3$ (the usual three-sigma limits) works reasonably well, particularly with larger values of λ. This method is good in detecting small shifts ($< 1.5\sigma$) and is less sensitive to the parameter setting than the CUSUM chart. It is insensitive to the normality assumption, because it is a weighted average. Another advantages of this method over the CUSUM is that it does not require a value k, meaning that there is less information needed on the change which might occur. An advantage over the Shewhart chart is that EWMA uses previous measurements.

Generalized Likelihood Ratio Charts

In Di Bucchianico (2019) it is shown that the CUSUM is based on the Sequential Probability Ratio Test (SPRT) and the Neyman-Pearson Lemma. This derivation shows important ideas which are also present in Generalized Likelihood Ratio Charts (GLR). The SPRT has as hypothesis

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1. \end{aligned}$$

This is tested with the likelihood ratio $\Lambda_n = \frac{f_{\theta_1}(X_1)\ldots f_{\theta_1}(X_n)}{f_{\theta_0}(X_i)\ldots f_{\theta_0}(X_n)}$, with the following threshold scheme

$$\begin{aligned} \text{continue monitoring} &: \text{ if } a < \Lambda_n < b \\ \text{accept } H_0 &: \text{ if } \Lambda_n \leq a \\ \text{accept } H_1 &: \text{ if } \Lambda_n \geq b, \end{aligned}$$

where a and b determine the type I and II errors.

As already discussed in Section 3.3.1, the SPC problem can be seen as sequential hypothesis testing. A Generlized Likelihood Ratio chart (GLR) uses a maximum likelihood estimate of the change time, instead of assuming it is known (Runger and Testik (2003)). It is based on the log-likelihood which is defined as

$$L(\theta_0, \theta_1|X) = \sum_{i=1}^{n} \log\left(\frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)}\right), \tag{3.1}$$

where f is the probability density function of the process and θ_0 and θ_1 are the parameters which we want to test in this distribution. The GLR method tries to find the starting point of a change by maximizing the likelihood data over the data history given by

$$g_n(\theta_0, \theta_1) = \max_{1\leq j\leq n} \sum_{i=j}^{n} \log\left(\frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)}\right). \tag{3.2}$$

This value g_n can be compared to a decision limit h to detect out of control behavior. Another variation on this control chart is the cuscore chart discussed in Runger and Testik (2003).

These methods can be designed to detect specific changes in a process and they do not assume a certain distribution of the process. A disadvantage is that the GLR chart is computational intensive Runger and Testik (2003). This however becomes less of a problem with the computational power of today.

Self-Starting Methods

Until now the methods required a phase I and a Phase II. In the lens case it is very difficult to determine a phase I period. Often phase I is chosen as the start of the process until enough data is collected for reliable estimates. For the ASML machines this data is often not available. If it is available, it might not be realistic to assume that this data describes a normal process. This is because the process is not always stable at the beginning of production. A solution to the lack of a clear defined phase I could be to use a self-starting method. These methods do not require a phase I and estimate the needed parameters while monitoring. Self-starting methods are described in Hawkins et al. (2003). The main ideas will be repeated here.

We start without any information of the process, therefore we have to estimate the mean and the variance. If we have a phase I with n measurements we estimate the mean with

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the sum of squared deviations with

$$w_n = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

For self-starting we start with $n = 1$ and define $\overline{x}_0 = 0$ and $w_0 = 0$. Then we use the recursion

$$\overline{x}_n = \overline{x}_{n-1} + \frac{x_n - \overline{x}_{n-1}}{n}$$

and

$$w_n = w_{n-1} + \frac{(n-1)(x_n - \overline{x}_{n-1})^2}{n}.$$

The sample variance is defined as $S_n^2 = \frac{w_n}{n-1}$. For $n \geq 3$ each new observation can be standardized by

$$T_n = \frac{x_n - \overline{x}_{n-1}}{s_{n-1}}.$$

In general the idea is to use a transformation such that the transformed data is independent and identically distributed. This is done using the following transformation

$$Q_n = \Phi^{-1}\left[F_{n-2}\left(\sqrt{\frac{n-1}{n}}T_n\right)\right],$$

where Φ^{-1} is the inverse normal cumulative distribution and F_{n-2} denotes the cumulative t distribution with $n-2$ degrees of freedom. The Q_n are standard near normal distributed.

The self-starting method similar to the Shewhart control chart is the Q chart. This chart is defined as

$$Q_n(x_n) = \Phi^{-1}\left(F_{n-2}\left[\sqrt{\frac{n-1}{n}}\frac{x_n - \overline{x}_{n-1}}{s_{n-1}}\right]\right), \text{ for } n \geq 3.$$

The control limits for the Q statistic are -3 and 3. This transformation can also be used for other charts. The Q statistic can be used in the CUSUM chart instead of the x_n variable to create a self-starting chart. The self-starting CUSUM is more robust to non-normal data then the original CUSUM Hawkins (1987). This is because the statistic T_n is distributed according to a t distribution. For the self-starting EWMA the same Q statistic is used instead of the x_n variable, see Li et al. (2010) for more details.

Summary

As we have seen there are many different charts for individual measurements. The MR chart do not take the history of the data into account. This makes it difficult to detect small shifts in the mean. The solution for this is to use the CUSUM chart or the EWMA chart which is a compromise between the Shewhart chart and the CUSUM chart. If more information is known about the change which is caused by the WELLE sticker failure a specific GLR chart could be designed for this shift. In the lens case changes are often small and occur over a longer period in time. Therefore we expect that the EWMA and CUSUM chart perform better than the Shewhart chart. It is difficult to find a phase I for the lens case, therefore self-starting methods could be a solution.

3.3.3 Multivariate Statistical Process Control

Using univariate SPC methods in a multivariate context is not always a good idea. This will be shown in the following example based on Montgomery (2009). Given two variables x and y plotted against each other in Figure 3.3. There is one point around (0,-4.5), which clearly deviates from the other points. The Shewhart control charts for x and y are given in Figures 3.4 and 3.5. There is no out of control point detected when the variables are monitored individually with a Shewhart chart. If we use the Hotelling's T^2 for multivariate data it detects this outlier. The Hotelling's T^2 will be described in the following section. The chart is given in Figure 3.6. This example shows that using univariate methods in a multivariate setting could cause false negatives. An important issue for multivariate methods is that if the amount of variables increases the covariance matrix gets more difficult to estimate. We will now discuss some multivariate SPC methods.

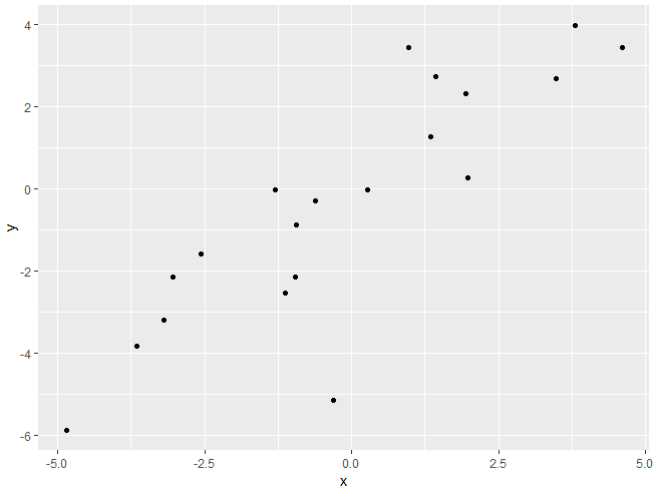


Figure 3.3: Scatterplot of the example data x and y

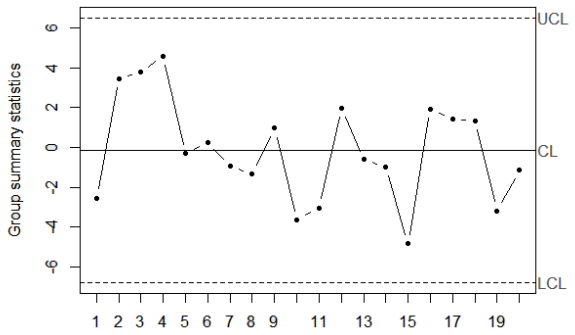


Figure 3.4: Shewhart chart for variable x

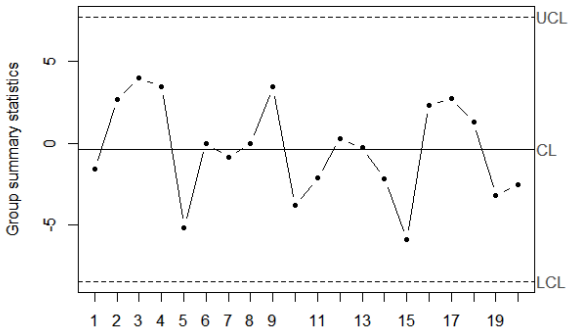


Figure 3.5: Shewhart chart for variable y

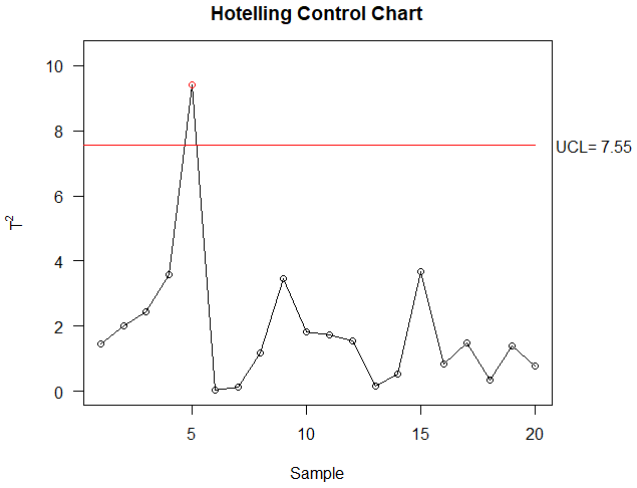


Figure 3.6: The Hotelling's T^2 chart on variables x and y

Hotelling's T^2

The first multivariate method that we will discuss is the Hotelling's T^2 method (Hotelling (1947)). This method can be used for subgrouped data and for individual observations. We will only discuss Hotelling's T^2 for individual observations, for more information on the subgrouped Hotelling's T^2 see chapter 11 in Montgomery (2009) or Mason and Young (2002). Given n samples with p quality characteristics for each sample as phase I. It is assumed that these samples are multivariate normal distributed with mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. The sample mean vector is denoted by $\overline{x}$ and the sample covariance matrix by S. The Hotelling's T^2 statistic is defined as

$$T^2 = (x - \overline{x})^T S^{-1}(x - \overline{x}).$$

The phase I control limits are

$$\begin{aligned} \text{UCL} &= \frac{(n-1)^2}{n}\beta_{\alpha,p/2,(n-p-1)/2} \\ \text{LCL} &= 0, \end{aligned}$$

where $\beta_{\alpha,p/2,(n-p-1)/2}$ follows a beta distribution with parameters α and $p/2, (n-p-1)/2$. The phase II control limits are

$$\begin{aligned} \text{UCL} &= \frac{p(n+1)(n-1)}{n^2 - np}F_{\alpha,p,n-p} \\ \text{LCL} &= 0 \end{aligned}$$

An approximation to these limits is $\text{UCL} = \frac{p(n-1)}{n-p}F_{\alpha,p,n-p}$. This is a reasonable approximation if $n > 100$. If the covariance matrix is known we can also use $\text{UCL} = \chi^2_{\alpha,p}$, this however should be used with caution if $p > 10$, then it is needed to have at least $n > 250$. If both the mean vector μ and the covariance matrix Σ are unknown we need to estimate them. This can be done by

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T,$$

where $x_i \in \mathbb{R}^p$ represents the ith measurement vector. The Hotelling's T^2 is directionally invariant, meaning that it detects a shift in the mean vector based on the magnitude of the shift and not in its direction. The control limits of the Hotelling's T^2 method can be interpreted as an ellipsoid in $\mathbb{R}^p$. If a measurement falls outside of the ellipsoid the process is out of control. This method only takes into account the last observation. This method is very similar to the Shewhart method for univariate SPC. We will also discuses the multivariate EWMA and multivariate CUSUM.

Multivariate Cumulative Sum

There are several ways of extending the univariate CUSUM to a multivariate CUSUM (MCUSUM). We will discuss the vector-valued scheme, which is obtained by replacing the scalars in the univariate CUSUM with vectors. Denote $C_t = \sqrt{L_t^t \Sigma^{-1} L_t}$, where

$$L_t = \begin{cases} 0 & \text{if } C_T \leq k \\ (L_{t-1} + X_t - \mu_0)(1 - \frac{k}{C_t}) & \text{otherwise,} \end{cases}$$

where $L_0 = 0$. Now we have to choose k, this can be chosen as $k = \frac{\sqrt{(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)}}{2}$ (Crosier (1988)). We can also choose a value for h, tables to choose h depending on the desired ARL can be found in Crosier (1988). An application of the MCUSUM is given in Harrou et al. (2015), where they choose $\text{ARL}_0 = 200$, which results in a false alarm rate of 0.05%, and the following parameter choices $k = 0.5$ and $h = 6.885$. Advantages of this method are that it uses information of the past and is good for detecting small changes. A disadvantage is that we have to choose k, which could be difficult if there is not a clear change of interest in the data.

Multivariate Exponential Weighted Moving Average

To extend the univariate EWMA chart to a multivariate EWMA chart(MEWMA) we replace the scalars with vectors. The MEWMA is given by

$$Z_i = \lambda x_i + (1-\lambda)Z_{i-1},$$

where $0 \leq \lambda \leq 1$, $Z_0 = 0$, $Z_i \in \mathbb{R}^n$ and $x_i \in \mathbb{R}^n$. This gives us a vector Z_i for each time. We want to monitor one value for each time therefore we use a method similar to Hotelling's T^2 on our new vector. The plotted quantity on the control chart is

$$T_i^2 = Z_i^T \Sigma_{Z_i}^{-1} Z_i,$$

where the covariance matrix is $\Sigma_{Z_i}^{-1} = \frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})\Sigma$. The control limit h can be chosen such that an ARL that is desired is achieved. The control limits and corresponding ARL are given in Chapter 11 in Montgomery (2009). An advantage of the MEWMA over the Hotelling's T^2 is that it uses information of the past instead of only the most recent observation. It also is good at detecting small changes. A disadvantage is that we have to choose λ and h, for which some experimenting has to be done with the data to get the desired result.

Self-Starting Methods

Also for multivariate methods it is possible to extend the methods to get self-starting methods. There exist several multivariate self-starting methods, examples can be found in Capizzi and Masarotto (2010) and Maboudou-Tchao and Hawkins (2011).

One possible method is to use the scalar accumulation chart, which uses that the sample mean and the sample covariance matrix can be updated using the following formulas

$$\begin{aligned}
\overline{x}_t &= \overline{x}_{t-1} + \frac{1}{t}d_t \\
S_t &= S_{t-1} + \frac{1}{t}\left(d_t d_t^T - \frac{t}{t-1}S_{t-1}\right),
\end{aligned}$$

where $d_t = x_t - \overline{x}_{t-1}$ is a column vector, $\overline{x}_0 = 0 \in \mathbb{R}^p$ and $S_0 = 0 \in \mathbb{R}^{p\times p}$. The following statistic is then used

$$Q_t = \Phi^{-1}\left(F_{p,t-p-1}\left[\frac{(t-1)(t-p-1)}{tp(t-2)}d_t^T S_{t-1}^{-1} d_t\right]\right),$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function and F is the cdf of a F distribution. This method is very similar to the univariate self-starting transformation.

Another possibility is to use the Self-Starting Multivariate Exponentially Weighted Moving Average Control Chart (SSMEWMA) introduced by Maboudou-Tchao and Hawkins (2011). They use the standardized recursive residuals. First the calculate the recursive residuals $r_{n,i}$ for $n > i+1$, which are the residuals from the predicted value $\hat{x}_{t,j}$ for the tth observation of

the jth variable. They use regression on the previous $t-1$ observations to get the prediction. These are standardized using

$$t_{n,i} = \frac{r_{n,i}}{\sqrt{\sum_{k=i+1}^{n-1} r_{k,i}^2/(n-i-1)}},$$

which follow a t distribution with $n-i-1$ degrees of freedom.

These $t_{n,i}$s are transformed using the probability integral approach, which is given by $u_{n,i} = \Phi^{-1}[F_{n-i-1}(t_{n,i})]$, where Φ^{-1} represents the inverse normal and F_{n-i-1} the cumulative t distribution function with $n-i-1$ degrees of freedom. The result can be gathered in the matrix U which consists of independent vectors which are multivariate normal distributed with mean vector 0 and covariance matrix I_p.

The monitoring starts from $t = p+2$. Define $z_t = (1-\lambda)z_{t-1} + \lambda u_t$ and $S_t = (1-\lambda)S_{t-1} + \lambda u_t u_t^T$. Then we define the following statistics

$$M_t = z_t^T \Sigma_{z_t}^{-1} z_t$$

and

$$C_t = \text{tr}(S_t) - \log|S_t| - p.$$

To determine if the original process is out of control we compare the statistics with control levels h_1 and h_2, which should be chosen such that the desired properties of the chart are achieved. We conclude that the system is out of control if $M_t > h_1$ or $C_t > h_2$. A different method named the CUSCORE and a short review of existing multivariate methods can be found in Capizzi and Masarotto (2010).

Summary

The multivariate charts are very similar to the univariate charts. They are needed because it could happen that two variables separately are not outliers, while together they are. When we use a large amount of univariate charts this increases the amount of false positives. Therefore multivariate charts should be used in multivariate situations. However using multivariate charts with a large amount of variables should be handled with care, because of the estimation of the covariance matrix. For multivariate charts there are self-starting methods available, but they are more complicated than the univariate self-starting methods. For the lens case there are a lot of variables, so multivariate charts could be a solution. A problem is however that the dimension of the data set is too large to be used for multivariate charts. A solution for this could be to reduce the dimension first.

3.4 Dimension Reduction

In modern applications there are many variables measured at a high frequency. For instance in neural data hundreds or thousands of neurons are recorded simultaneously (Freeman et al. (2014)). But also in monitoring the performance of ASML machines hundreds of measurements are made in a short time. These high dimensional data sets can cause problems when analyzing the data (Ferrer (2014)).

- A high number of variables makes it more difficult to detect disturbances using conventional MSPC charts.
- Numerical issues with the calculation of the covariance matrix or the inverse of the covariance matrix might be caused by a large dimension.

- Visualizing high dimensional data is difficult.

These problems can be solved by reducing the dimension. A solution could be to select a subset of the original variables. The problem however is that it is unclear in which variables a WELLE sticker failure is shown. Another option is to transform the data to a lower dimension. This can be used with principle component analysis (PCA). In this method all original data is used.

3.4.1 Principal Component Analysis

PCA is used in many different fields for example in anomaly detection in industrial networks (Iturbe et al. (2017)) or in smartphone data (Mirsky et al. (2017)). In Ferrer (2014) it is recommended to use latent variables for SPC, PCA finds these latent variables. It transforms the data into new variables, in such a way that the variables can be ordered on the amount of variation they explain in the data. For example highly correlated variables explain similar variation, these variables can be replaced with one variable that explains all that variation. The idea is that only a few PCA variables explain all the variation in the original variables.

Derivation of Principal Component Analysis

The derivation of PCA is stated in Jolliffe (2011), this will be summarized here to explain the ideas behind PCA. Given the original data as $Z \in \mathbb{R}^{n\times p}$, where n is the amount of samples and p are the different measurements. Suppose that $n > p$. First the measurements are normalized, such that hey all have the same variance and mean 0. This is done by

$$X_{ij} = \frac{Z_{ij} - \overline{Z}_j}{\sqrt{(\mathrm{Var}(Z_j))}},$$

where $\overline{Z}_j = \frac{1}{n}\sum_{i=1}^{n} Z_{ij}$ and $\mathrm{Var}(Z_j) = \frac{1}{n}\sum_{i=1}^{n}(Z_{ij} - \overline{Z}_j)^2$.

We can now calculate the covariance matrix $C \in \mathbb{R}^{p\times p}$ by

$$C = Cov(X) = \frac{1}{n-1}(X - \overline{X})^T(X - \overline{X}) = \frac{1}{n-1}X^TX,$$

where we use that the mean of each variable is 0. For the first transformed variable we want to find a linear combination such that $\alpha^T x$ has maximum variance. The linear combination is denoted by

$$\alpha_1^T x = \alpha_{11}x_1 + \alpha_{12}x_2 + ... + \alpha_{1p}x_p,$$

where α_1 is a vector with coefficients for the transformation. The next transformed variable needs to be uncorrelated with $\alpha_1^T x$ and still maximize the variance. These vectors $\alpha_i, i = 1, ..., p$ are called loadings (coefficients) and have unit length 1, which is denoted by $||\alpha|| = 1$. Formulated as an optimization problem this gives

$$\begin{aligned} \max_{\alpha} \quad & \alpha^T C\alpha \\ \text{subject to} \quad & ||\alpha|| = 1. \end{aligned}$$

This can be solved by using the Lagrange multiplier, which gives the following minimization problem

$$\min_{\alpha} \quad \alpha^T C\alpha - \lambda(\alpha^T\alpha - 1).$$

Solving this by differentiating to α and setting equal to 0 gives

$$\begin{aligned} C\alpha - \lambda\alpha &= \vec{0} \\ C\alpha &= \lambda\alpha. \end{aligned}$$

The solution to this problem can be found by using the eigenvectors of C. This means that we can get p different loadings. Solving the original maximization problem gives

$$\alpha^T C\alpha = \alpha^T \lambda\alpha = \lambda\alpha^T\alpha = \lambda.$$

Maximizing this is done by choosing the largest eigenvalue. To get the best possible set of different α's we can sort the eigenvalues from large to small and pick the corresponding eigenvectors as α. The amount of variance explained by each principal component can be calculated by using the eigenvalues as follows

$$\text{Explained variance PC } i = \frac{\lambda_i}{\sum_{j=1}^{n}\lambda_j}.$$

This transformation can be easily computed by using the eigenvalue decomposition of the covariance matrix C. Since the covariance matrix C is symmetric it can be written as $C = P\Lambda P^T$, with $PP^T = P^TP = I$ (Orthogonal) and Λ a diagonal matrix with the eigenvalues on the diagonal. The matrix P is also called loadings matrix or rotation matrix.

To transform the original data we calculate $T = XP$, where $T \in \mathbb{R}^{n\times p}$. The matrix T contains all the newly transformed variables which are called the principal components or scores. Select only m principal component such that the important features of the data are explained and the noise is filtered out. Note that if the original variables are correlated that m could be much smaller than p.

Using Principal Component Analysis

Now we assume that we have some data that we already observed and normalized to get $X \in \mathbb{R}^{n\times p}$, we also have k new samples denoted by $Z_{\text{new}} \in \mathbb{R}^{k\times p}$. Then we can use PCA to obtain from X, the matrix with loadings $\tilde{P} \in \mathbb{R}^{p\times m}$, where m is the amount of principal components we retain. Now we can transform Z to the new variables $\tilde{T} \in \mathbb{R}^{n\times m}$ by $\tilde{T} = X\tilde{P}$. This transformed data can be used for analyzing and finding interesting patterns that were not clearly visible before. Often a model is created on existing data such that it can be used for new data. The new sample Z_{new} can be normalized by using the earlier used mean and variance $\overline{Z}_j$ and $\text{Var}(Z_j)$. The transformed variables are obtained by $T_{\text{new}} = X_{\text{new}}\tilde{P}$. The only thing we did not cover was how to choose m, the amount of principal components to retain.

How many Principal Components?

There are several methods to select the amount of principal components. In Jackson (1993), Jackson (2005) and Jolliffe (2011) several methods are explained and evaluated. Some of these methods will be explained here briefly.

One of the easiest criteria to choose how many principal components is to use the cumulative percentage of the total variance. In this method a threshold is chosen to determine how much of the total variation in the original data should be explained by the principal components. Denote the cumulative threshold as

$$t_i = \frac{100}{p}\sum_{k=1}^{i}\ell_k,$$

where ℓ_k is the amount of variation explained by the kth PC. The threshold is denoted by t^*, which is in percentage, such that m can be chosen according to

$$m = \underset{i=1,\ldots,p}{\text{argmin}} \quad t_i > t^*.$$

This method could retain a lot of components that summarize noise, especially if the threshold is chosen close to 100%. When p is very large for instance $t^* = 95\%$ overestimates the amount of components (Jackson (1993)). It could also result in large m. This means that t^* should be made smaller if the goal is to retain only a few principal components.

Another widely used method is the scree plot. It plots eigenvalues λ_k against k. In this method we search for a point for which holds that the points to the left are 'steep' and the points on the right are 'not steep'. This graph should have an elbow form such that it is clear at which point the change occurs. This method often overestimates the number of components (Jackson (1993)).

Another criteria is the Kaiser-Guttman criteria. Only the principal components with $\lambda_i > 1$ are retained. This method often selects too many components except if the data has a strongly correlated structure.

The broken-stick method is based on randomly dividing a unit length stick into p segments. Then the expected length of the kth longest segment is

$$g_k = \frac{1}{p} \sum_{i=k}^{p} \frac{1}{i}.$$

To decide if the kth PC should be retained we need g_k to be smaller as the proportion of variance explained by PC k. The broken-stick method performs good in many situations according to Jackson (1993).

A method known as Bartlett's test can be used to test which amount of PCs to retain. This test has null hypothesis

$$H_{0,q} : \lambda_{q+1} = \lambda_{q+2} = ... = \lambda_p,$$

and the alternative hypotheses is that at least two of the last $(p - q)$ eigenvalues are not the same. This test is used for different values of q so we start with $H_{0,p-2}$ if this is not rejected we test $H_{0,p-3}$ until we find a q^* for which H_{0,q^*} is rejected then we choose $m = q^* + 1$. This procedure assumes that the data follows a multivariate normal distribution. A downside is that this method uses multiple tests, which could result in more false positives if not handled correctly. A correct significance level cannot be chosen because at the start of the procedure it is unknown how many tests will be done. Therefore this method could give wrong results Jackson (1993).

A final method that is used is cross-validation. However in Bro et al. (2008) it is concluded that there are many different approaches to implement cross-validation. These approaches are however not an extensions of the original idea of cross-validation. This is because the data that is left out is often used to predict itself in someway. Therefore it is concluded that cross-validation for PCA does not give meaningful results.

Based on the literature broken-stick method gives the best result, therefore we will use that method. Do note that according to Camacho and Ferrer (2014) the amount of principal components to retain could be different depending on the purpose of the PCA.

Principal Component Analysis and Statistical Process Control

We are interested in the combination of PCA with SPC to better detect changes in the data. A common approach is to first fit a PCA model and then monitor the PCA variables with Hotelling's T^2. Another method is to use the PCA as a model and see if new values give an expected result based on the model. This is done with the Q statistic also referred to as the Squared Prediction error (SPE). The Q statistic is more sensitive to faults with a smaller magnitude than the Hotelling's T^2 statistic (Harrou et al. (2015)). It is important that the PCA model should be fit on fault free data and standardized data (Harrou et al. (2015). A

disadvantage of using PCA is that when an out of control signal is given it is not clear which original variables are responsible for the signal. For the WELLE sticker case this is not a disadvantage because our main interest is not in the specific Zernike coefficient which caused the out of control signal.

Dynamic Principal Component Analysis

A downside of PCA is that it is a linear projection of the data and it is not able to consider the process dynamics (Banko et al. (2011)). A solution for this is to use Dynamic Principal Component Analysis (DPCA). This method does not only use the current measurements, but also previous measurements. Given two time series Y and U then PCA has as input $[y_t, u_t]$, for dynamic PCA the input would be $[y_t, u_t, y_{t-1}, u_{t-1}, ..., y_{t-L}, u_{t-L}]$, where L denotes the time lag. In Tsung and Apley (2002) a method is shown to select the L for feedback-controlled processes. In Shi and Tsung (2003) DPCA is used to model and diagnose feedback-controlled processes. The lenses in the ASML machines are a feedback-controlled process, because at every measurement the systems corrects for unwanted behavior. Therefore DPCA could be used for the WELLE sticker case.

Summary

PCA could be used within ASML to make it easier to visualize high dimensional data by using it as a dimension reduction method. It could be used to extract valuable information and reduce noise from high dimensional data, such that this can be used to for prediction of certain events or model certain behavior. We will use the broken stick method as the main method to select the amount of principal components. PCA and SPC can be combined in two ways, if we need a sensitive method for small shifts we should use the CUSUM or the EWMA, otherwise the Hotelling's T^2 is better. We could consider to use DPCA, because it would describe the possible process dynamics better.

3.5 Automated Process Control

In the WELLE sticker case we have to deal with a process that is continuously adjusted. These adjustments are applied within ASML on several processes. This is important to realize, because it adjusts the process that we observe. A lot of processes within ASML are controlled, because that is needed to achieve the precision that is needed within the machines. Therefore we briefly give some background on how APC works. The theory to control a process comes from the field Automated Process Control (APC) (also Engineering Process Control (EPC)), which is part of control theory. It is used to control a process using control actions. This is necessary if a production process suffers from disturbances that cannot be removed. For example the lenses suffer from lens heating and wear of parts. This is something that cannot be removed and must therefore be controlled.

3.5.1 Description

In an automated process there is a quantity that needs to be controlled, this is the value that is measured. This process has a target value (also set point) at which it should be. There is a certain disturbance, which could be modeled and based on all the previous information a control action has to be taken, such that the process remains close to the target value. This gives the following notation (van Zante-de Fokkert et al. (1999)):

Control actions X_t

Disturbance Z_t

Measurement U_t

Target value Y_t

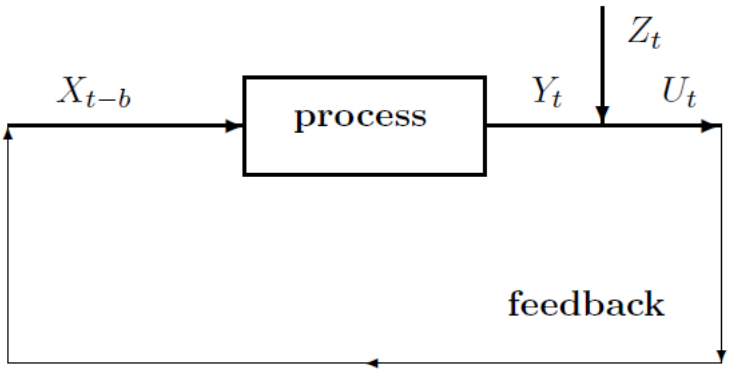


Figure 3.7: An schematic view of a control system from van Zante-de Fokkert et al. (1999)

A schematic view of the process using this notation can be seen in Figure 3.7. To successfully model a control process a disturbance model has to be chosen for Z_t. This is often done by choosing an ARMA(p,q) model. This is denoted by

$$\Phi(B)Z_t = \Theta(B)a_t,$$

where $a_t \sim N(0, \sigma_a^2)$ and B is the backshift operator such that $Ba_t = a_{t-1}$. Here

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$$

and

$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_p B^p$$

are polynomials in B. This can be used to define the control actions X_t. Often the control action is choosen by using Proportional-Integral-Derivative (PID), Proportional-Integral (PI), Integral (I) controllers or exponentially weighted moving average control Shi and Tsung (2003). The measurements of an APC controlled process are correlated, because of the adjustments (Di Bucchianico and van Gellecum (2007)). There is much more to discuss in this field, but for the case of the WELLE sticker replacement the control actions are already modeled.

The most important observations are that the systems is adjusted and that the measurements of an APC controlled process are correlated. An interesting idea could come from the observation that the calibration (Calibrating Lens Model) of the system is executed manually. It could be that SPC can be used to detect if it is needed to calibrate the system.

3.6 Combining Statistical Process Control and Automated Process Control

For the WELLE sticker case we have to combine SPC and APC. We do not have a choice in how the control actions are done. For completeness we however do give some options to combine SPC and APC (Di Bucchianico and van Gellecum (2007)). These are

SPC as deadband Use SPC on the process, when it signals out of control use APC.

Run-to-run process control This happens in process in which the control actions can only be executed in between a run. SPC acts as a supervisor to indicate the need of a control action in between runs.

Statistical monitoring of feedback control APC regulates the process, but the quality characteristic of the process is monitored by SPC. To protect against issues the controller cannot solve.

Joint monitoring scheme Let APC control the process and monitor both the control actions and the measurements of the process together.

In the lens case the APC system is already determined. In this case it is also needed to continuously control the process because of the lens heating issue. This means that we have to monitor the control actions and the measurements simultaneously in a joint monitoring scheme. Monitoring only the measurements in an APC controlled systems has some disadvantages. The measurements are correlated because of the adjustments (Di Bucchianico and van Gellecum (2007)). This autocrellation on the measurements could cause many traditional charts to perform poorly (Tsung and Tsui (2003)). A solution is proposed by Tsung and Apley (2002) and is to use dynamic PCA. To monitor a multivariate time series we need to estimate the covariance matrix. This is however not always possible because of linear dependencies. A solution for this is to use the generalized inverse.

Selecting the L in dynamic PCA is important because if L is selected smaller than necessary the dynamic relation of the process is not fully captured. This could lead to ineffective monitoring. Selecting a L larger than necessary could lead to repeated structure in the input and output. Large L, implies better detection of small shifts, but slower detection of large shifts (Tsung and Apley (2002)).

Another issue with monitoring the measurements without the corrections is the Window of Opportunity problem (Tsung and Apley (2002)). This means that a starting problem shows only for a brief period in the measurement before the system corrects the problem. At this point the problem might not be that catastrophic, for instance if only a very small part of the WELLE sticker starts to let loose. This small issue might be corrected by the system, but the system is producing sub optimal results. This makes it more difficult to detect when state of the sticker reaches a point at which it cannot be corrected anymore. So to stress the importance of this point, the measurements would not show a clear sign of a breaking part, while the corrections show a certain shift or drift. Therefore it is strongly advised to use both the measurements and the corrections.

A final point at which machine learning can come in, is suggested by Tsung (2000) and Shi and Tsung (2003). After finding an out of control behavior is is possible to diagnose the system by creating a fault database. This database is used to classify new signals with machine learning algorithms. This way an issue could be diagnosed based on previous experience. If an issues is very similar then the same solutions as for the previous time such an issue was solved could be used.

Summary

It is very important to use the measurements and the corrections at the same time such that failures can be detected early. Using all those variables means for the WELLE sticker case with SyCo that we have a lot of variables. Therefore PCA is very important such that the dimension can be reduced. These newly created variables can then be used for monitoring and detecting failures.

3.7 Orthogonal Polynomials

To better understand Zernike polynomials we first define orthogonal polynomials. These have all kinds of uses within mathematics. For example within statistics orthogonal polynomials are used to fit a regression model (Bingham and Fry (2010)). This is used in the case that we first fit a first order regression model $y = \hat{\beta}_0 p_0(x) + \hat{\beta}_1 p_1(x) + \varepsilon$, where P_0 and P_1 are orthogonal polynomials. Later we might be interested in a second order model $y = \hat{\beta}_0 p_0(x) + \hat{\beta}_1 p_1(x) + \hat{\beta}_2 p_2(x) + \varepsilon$. Than we have that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the same for both models. Meaning that we only have to estimate $\hat{\beta}_3$ for the new model.

To define orthogonal polynomials we first define an inner product in $\mathbb{R}$ (Temme (2011)). Define some subset V and measure $\sigma(x)$. Use the linear space $\mathcal{P}$ of polynomials of the real parameter x with real coefficients. Let $f, g \in \mathcal{P}$, then

$$(f, g) = \int_V f(x)g(x)d\sigma(x)$$

is the inner product. With the following properties $(f, g) = (g, f)$ and $(\alpha f + \beta g, h) = \alpha(f, h) + \beta(g, h)$, with $\alpha, \beta \in \mathbb{R}$ and $f, g, h \in \mathcal{P}$. We have that

$$(f, f) = \int_V f^2(x)d\sigma(x) \geq 0.$$

Furthermore we have that $||f|| = \sqrt{(f, f)}$.

Constructing a set of orthogonal polynomials can be done using the Gram-Schmidt orthogonalization method. We start with $f_0(x) = 1, f_1(x) = x, f_2(x) = x^2, ..., f_n(x) = x^n$ and the defined inner product. We will construct a set of polynomials $p_0, p_1, p_2, ..., p_n$. Then

$$p_0 = \frac{f_0}{||f_0||}$$

and

$$p_i(x) = \frac{f_i - \sum_{j=0}^{i-1}(f_i, p_j)p_j}{||f_i - \sum_{j=0}^{i-1}(f_i, p_j)p_j||}$$

such that $||p_i|| = 1$ and $(p_j, p_k) = 0$, $\forall j \neq k$.

3.7.1 Zernike Polynomials

Zernike polynomials are a set of orthogonal polynomials on the unit disc, which are often used in optics because they resemble problems found in optics. Each Zernike polynomial describes the aberrations in each lens on the unit disk. The Zernike polynomials are defined on the unit circle in polar coordinates. They are given by

$$Z_n^m(\rho, \theta) = \begin{cases} N_n^m R_n^m(\rho)\ \sin(m\theta) & \text{if } m < 0 \\ N_n^m R_n^m(\rho)\ \cos(m\theta) & \text{if } m \geq 0, \end{cases}$$

where the function R_n^m is called the radial function. This is given by

$$R_n^m = \begin{cases} \sum_{\ell=0}^{(n-|m|)/2} \frac{(-1)^\ell (n-\ell)!}{\ell!((n+m)/2-\ell)!((n-m)/2-\ell)!}\rho^{n-2\ell} & \text{for } n-m \text{ even} \\ 0 & \text{for } n-m \text{ odd} \end{cases}.$$

The normalization constant is defined as

$$N_n^m = \sqrt{\frac{2(n+1)}{1+\delta_{m,0}}}.$$

From R_n^m it can be seen that if $n-m$ is odd gives that the polynomials are zero.

As example we will calculate Z_0^0. We take $n=m=0$. This gives that

$$Z_0^0(\rho,\theta) = N_0^0 R_0^0(\rho),$$

$$R_n^m = \frac{(-1)^0(0)!}{0!((0)/2-0)!(0)/2!}\rho^0 = 1,$$

and

$$N_n^m = 1.$$

Combining this gives us that

$$Z_0^0(\rho,\theta) = 1.$$

Note that when $m=0$, we have that the $\cos(m\theta)=0$ which means the Zernike polynomial Z_n^0 is only depended on ρ.

We will show the orthogonal property of these polynomials on the unit disk. For $m' \neq m$ and $n' \neq n$ we have that

$$\int_0^1 \int_0^{2\pi} Z_n^m(\rho,\theta) Z_{n'}^{m'}(\rho,\theta)\rho d\theta d\rho = N_n^m N_{n'}^{m'} \int_0^1 R_{n'}^{m'}(\rho) R_n^m(\rho)\rho d\rho \int_0^{2\pi} f(\theta,m,m')d\theta = 0,$$

where $f(\theta,m,m')$ is a combination of cos and sin depending on m and m'. There are three possible combinations.

$$\begin{aligned}\int_0^{2\pi} \sin(m\theta)\sin(m'\theta)d\theta &= \int_0^{2\pi} \frac{1}{2}(\cos((m-m')\theta) - \cos((m+m')\theta)d\theta \\ &= \frac{1}{2}\left[\frac{\sin((m-m')\theta)}{m-m'} - \frac{\sin((m+m')\theta)}{m+m'}\right]_0^{2\pi} = 0\end{aligned}$$

$$\begin{aligned}\int_0^{2\pi} \cos(m\theta)\cos(m'\theta)d\theta &= \int_0^{2\pi} \frac{1}{2}(\cos((m-m')\theta) + \cos((m+m')\theta)d\theta \\ &= \frac{1}{2}\left[\frac{\sin((m-m')\theta)}{m-m'} + \frac{\sin((m+m')\theta)}{m+m'}\right]_0^{2\pi} = 0\end{aligned}$$

$$\begin{aligned}\int_0^{2\pi} \cos(m\theta)\sin(m'\theta)d\theta &= \int_0^{2\pi} \frac{1}{2}(\sin((m+m')\theta) + \sin((m-m')\theta)d\theta \\ &= \frac{1}{2}\left[-\frac{\cos((m-m')\theta)}{m-m'} - \frac{\cos((m+m')\theta)}{m+m'}\right]_0^{2\pi} = 0\end{aligned}$$

So for these combinations we have that the polynomials are orthogonal. It is also possible that $m=m'$ and $n \neq n'$ than we have two cases.

$$\int_0^{2\pi} \cos(m\theta)^2 d\theta = \pi$$

$$\int_0^{2\pi} \sin(m\theta)^2 d\theta = \pi$$

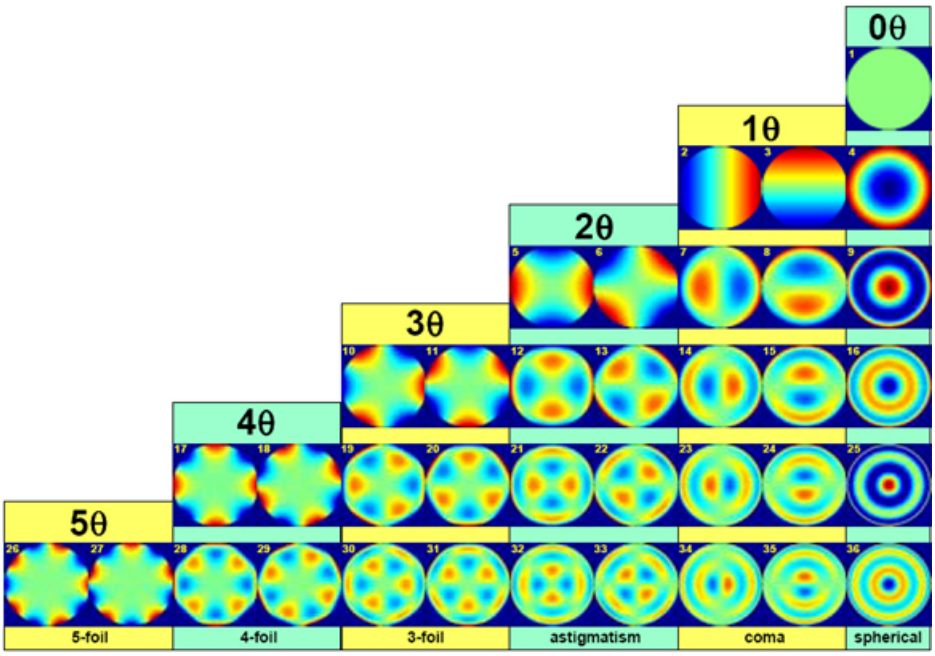


Figure 3.8: The first 36 Zernike polynomials (source: ASML)

Then we have that the radial part is

$$\int_0^1 R_n^{|m|}(\rho) R_{n'}^{|m|}(\rho) \rho d\rho = \frac{\delta_{n,n'}}{2n+2},$$

where $\delta_{n,n'} = \begin{cases} 1 & n = n' \\ 0 & n \neq n' \end{cases}$. So we have that the Zernike polynomials are orthogonal.

There are different ways of indexing these polynomials. The traditional way is to use two indexes m and n as we have already seen. It is often more convenient to use a single index. There are three methods we will show these are the OSA/ANSI indices, the Fringe/Zemax indices and the Noll's sequential indices.

OSA/ANSI standard indices $j = \frac{n(n+2)+m}{2}$.

The Fringe/Zemax indices $j = (1 + \frac{n+|m|}{2})^2 - 2m + \frac{1-sgn(m)}{2}$.

The Noll's sequential indices $j = \frac{n(n+1)}{2} + |m| + \begin{cases} 0 & m > 0 \wedge n = \{0,1\} (\text{mod} 4) \\ 0 & m < 0 \wedge n = \{2,3\} (\text{mod} 4) \\ 1 & m \geq 0 \wedge n = \{2,3\} (\text{mod} 4) \\ 1 & m \leq 0 \wedge n = \{0,1\} (\text{mod} 4) \end{cases}$

Within ASML the Fringe index is used.

In Figure 3.8 the first 36 Zernikes can be seen. Note that the first the fourth, the ninth, the sixteenth, the 25th and the 36th all have circular forms. For these Zernike polynomials it is the cases that $m = 0$. it can be seen that the higher the Zernike order the more specific the surface is that is described. ASML does not use the first Zernike polynomial, because they cannot measure that one. A linear combination of these Zernike coefficients can describe any wavefront. This is denoted by

$$w(x, y, \beta) = \sum_{i=0}^{\infty} \beta_i Z_i(x, y) + \varepsilon,$$

where $Z_i(x, y)$ is the ith Zernike polynomial and we assume that $\varepsilon \sim N(0, \sigma^2)$. In practice the ∞ is replaced by 64, 36 or 25.

Chapter 4

Data Description

4.1 Origin of the Data

For each machine within ASML one year of SyCo data is available. This data is collected for all machines with a WELLE sticker replacement. In a previous study data was collected from January 2017 until January 2018. This gives us for most machines data starting at January 2017 until January 2019. There are 53 machines for which the WELLE sticker is replaced in the last 2 years. There is no data available for replacements that happened more than 2 years ago. Since SyCo is done voluntarily by the customers. ASML advises to do this every day once, but this advised is not always followed. This causes large gaps in the data. The time intervals between measurements are also different because the owner has to decide when to execute SyCo. Besides the 53 machines there are 263 machines which should be healthy.

There is also data available about the machines, for instance the type, the owner and the delivery data. For each machine we can also find actions for the lens. This concerns the dates of the parts that are replaced. There is no information available about the reason a part was replaced and if it solved the issue. For each machine the log files of 1 year until the present are available. In these files the dates of lens calibrations could be found by searching for the code EMZQ, with event code TM-1103. These calibrations are downloaded for all 53 machines for one year.

4.2 Data Description

A description for the measured data will be given, this data is in long format and has the following variables:

Machine (EQUIPMENT_NR) The machine for which this measurement is. Most machines have a 4 digit number, but there are also machines which have letters in the name.

Date (SAMPLE_DT) The time at which the measurement is made.

Zernike (Zernike) The Zernike polynomial that is measured. This is one of Z2 till Z64.

Coefficient (Variable_ID) The coefficient of the linear model. These are offset, tilt, curvature and 3rd which are denoted by {0,1,2,3}.

Drift (Drift_type) The drift type, either Residual drift or Total drift.

Value (VALUE) The value of the corresponding measurement.

4.3 Zernikes Explained

Each Zernike polynomial is measured in the field of the lens, this is done on 13×5 points. There are 4 coefficients reported, these 4 coefficients describe a linear model, which can be used to approximate the original measurements. An example is given in Figure 4.1, where each line represents the measurements at one time of one Zernike polynomial. The colour of the lines denotes different measurements in time. There are 20 lines representing the change in 20 days. This gives an idea of how the 4 coefficients translate to the model and what kind of aberrations

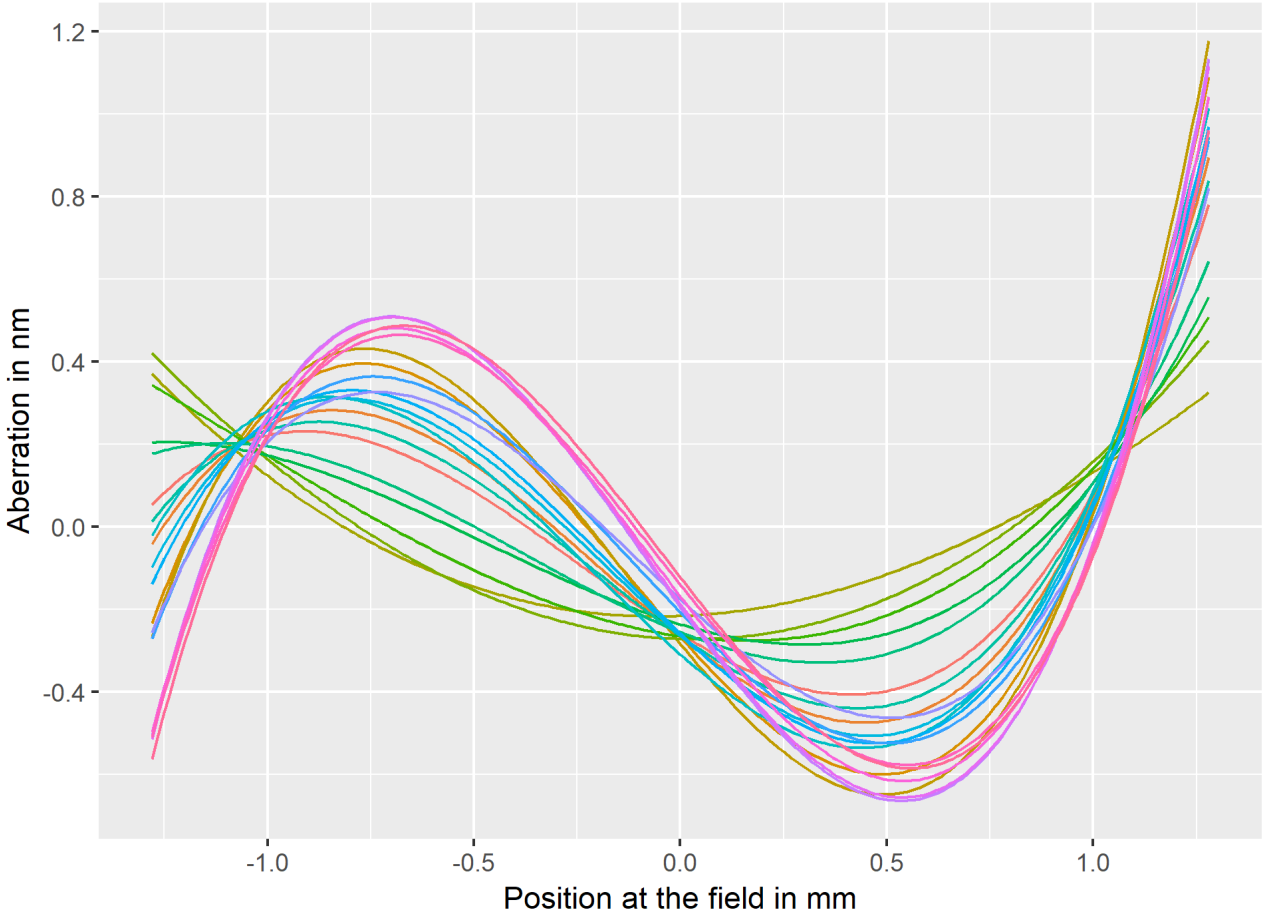


Figure 4.1: Example of the meaning of the linear model

there are in the complete lens field. It also shows that there is change over time in these models.

4.4 Data Visualization

To get an idea of the data we show for machine 18648 the residual drift and the total drift. While analyzing this machine it turns out that from the $63 \times 4 \times 2 = 504$ possible variables there are only 234 variables measured, the other variables are 0 for all entries. We omit these 0 variables from our plot. The result for residual drift can be seen in Figure 4.2 and for total drift in Figure 4.3.

These images are very chaotic, because of the many variables. The most interesting to see is the clear difference before and after the replacement in both figures. There is also clear deviating behavior in at least one of the coefficients in most plots before the replacement. In Table 4.1 the Zernike polynomial that attain the minimum and maximum value before the replacement are given. It can be seen that Zernike polynomials Z2, Z3, Z6, Z7, Z9, Z10 and

(a) Offset

(b) Tilt

(c) Curvature

(d) 3th order

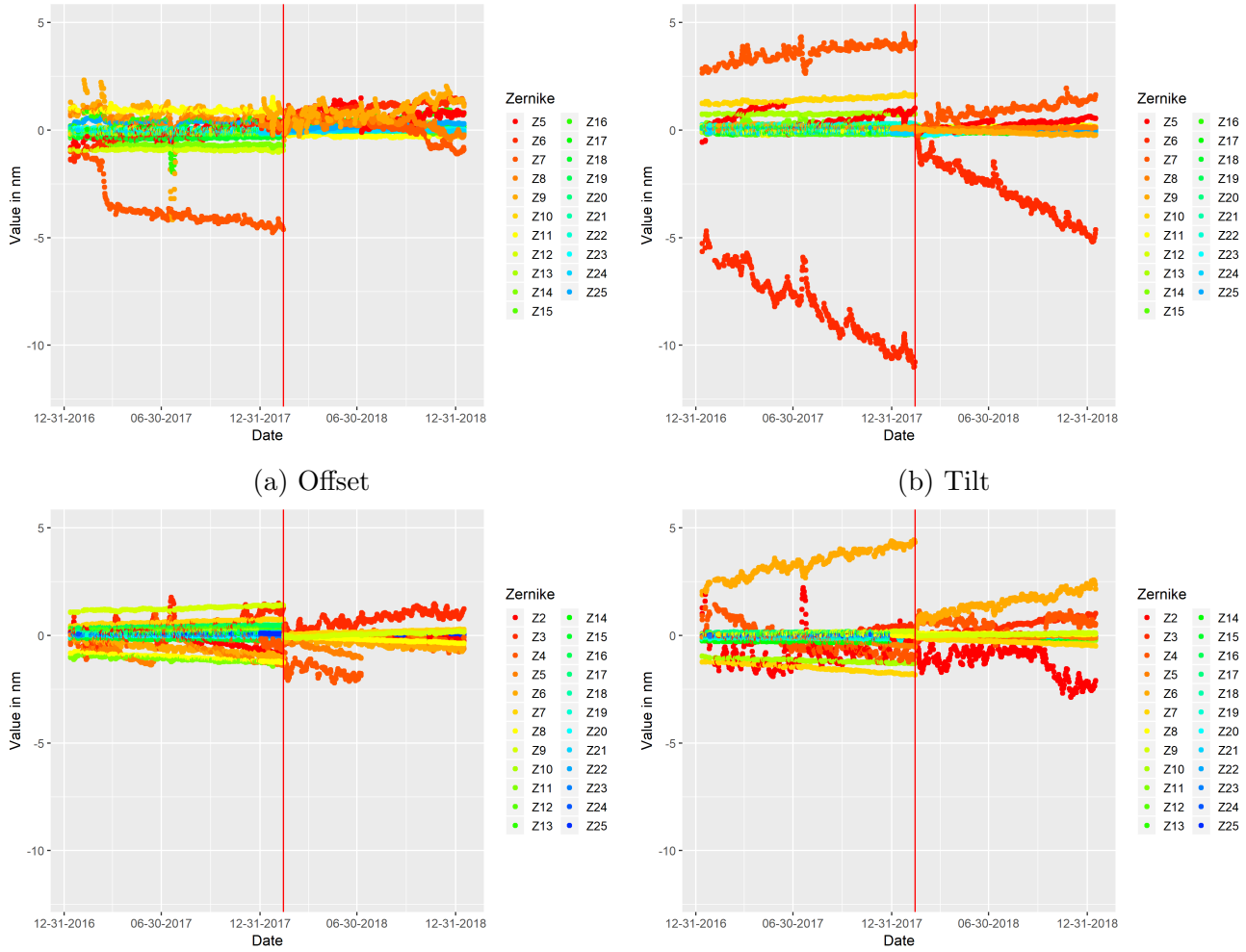
Figure 4.2: Residual drift for machine 18648, the red line indicates the WELLE sticker replacement

Z11 have the minimum or maximum value. Especially Z6 is 3 times the minimum and 2 times the maximum.

Drift	Coefficient	Min	Max
Residual	Offset	Z9	Z11
Residual	Tilt	Z6	Z10
Residual	Curvature	Z6	Z9
Residual	3th order	Z10	Z6
Total	Offset	Z7	Z9
Total	Tilt	Z6	Z7
Total	Curvature	Z11	Z3
Total	3th order	Z2	Z6

Table 4.1: The Zernike coefficients that attain the minimum and maximum value before the WELLE sticker replacement

When looking at the mean and variance of each individual coefficient we see that the mean and variance decrease as the Zernike polynomial number increases. This is something that we expect because the larger the number of the Zernike polynomial the more subtle the pattern



(a) Offset

(b) Tilt

(c) Curvature

(d) 3th order

Figure 4.3: Total drift for machine 18648, the red line indicates the WELLE sticker replacement

that the Zernike polynomial describes. Something that we are interested in is the correlation between variables. In Figure 4.4 the correlation between offsets for residual drift can be seen. There are some strong correlations for instance (Z3,Z4 and Z5), (Z27 and Z28) and (Z29 and Z30). Similar plots can be made for other types. It is very difficult to spot a clear pattern between all the possible plots. The most important observation is that there exist correlations between the variables, which could indicate that we can reduce the dimension of this data.

4.5 Data Preprocessing

For the analysis of the data it is needed to transform the data from wide to long format. In the wide format we remove all the zero columns, because the zero columns are not measured. This is also needed to normalize the data, because a column with zero variance cannot be normalized. For each value we subtract the mean of the column and divide the value by the standard deviation. This gives that each column has mean 0 and variance 1. The normalized data is better comparable, which is needed because of the differences in residual drift and total drift.

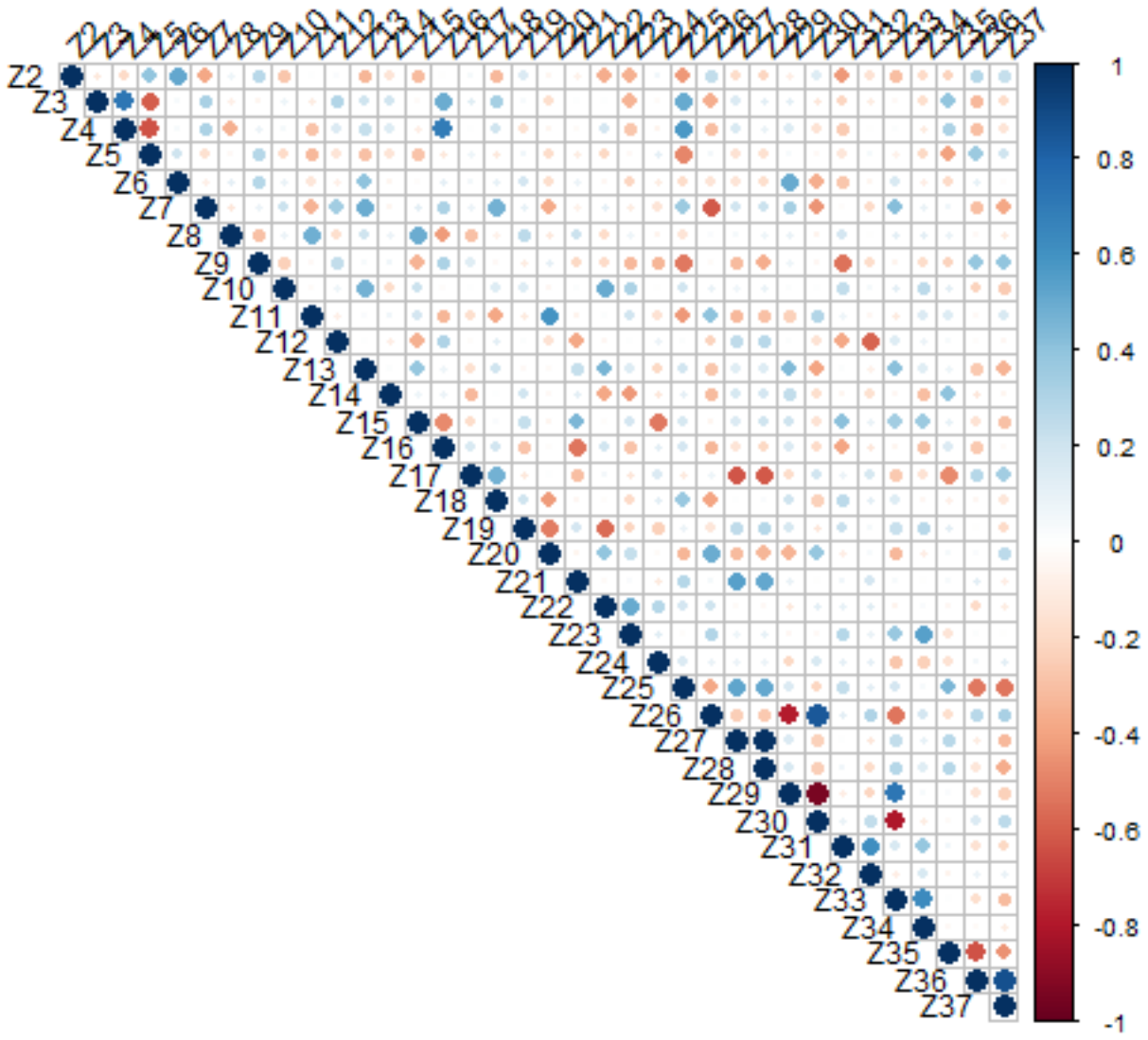


Figure 4.4: Correlation between offsets for residual drift for machine 18648

Chapter 5

Methodology

5.1 Approach

In this thesis we want to detect performance issues caused by the WELLE sticker. We expect that this is expressed in gradual changes in the SyCo data. These are changes such as drift or small sudden jumps. A large sudden jump would mean that the machine breaks completely. If a large sudden jump occurs, we expect to see warning signs before this jump. If this is not the case then the failure cannot be predicted. We focus on slow changes, because the SyCo data is only measured once a day. This means that we cannot detect sudden failures that occur in less than a day.

5.1.1 Changes

Detecting these gradual changes in an environment where adjustments are made is difficult. This is because of the window of opportunity in the measurements. An example of this phenomena can be seen in Figure 5.1. In the measurements of this system there is only a short moment in time in which a change can be detected, while in the corrections there is a clear difference after a change happened in the measurements. In practice there is often noise in the

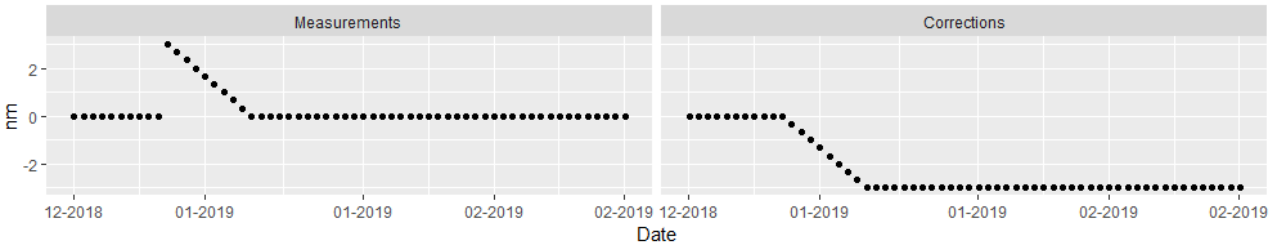


Figure 5.1: An example of a corrected system without variance

data, which makes it more difficult to see such a jump in the measurements. A situation with noise can be seen in Figure 5.2. In this example it is more difficult to detect a jump in the measurements, while in the corrections it is clear after a while that something has changed. This example shows that it is important to also take the corrections into account when the system is being controlled. For the SyCo data this means that the total drift and the residual drift are used together. Therefore we see the data at each measurement point as a vector of the following form:

$$[Z2\_0_R, Z2\_1_R, ..., Z64\_2_R, Z64\_3_R, Z2\_0_T, Z2\_1_T, ..., Z64\_2_T, Z64\_3_T],$$

where R stands for residual drift and T for total drift. If any of these variables is not measured (is zero for every measurement) for a machine we omit the variable from this vector. This gives

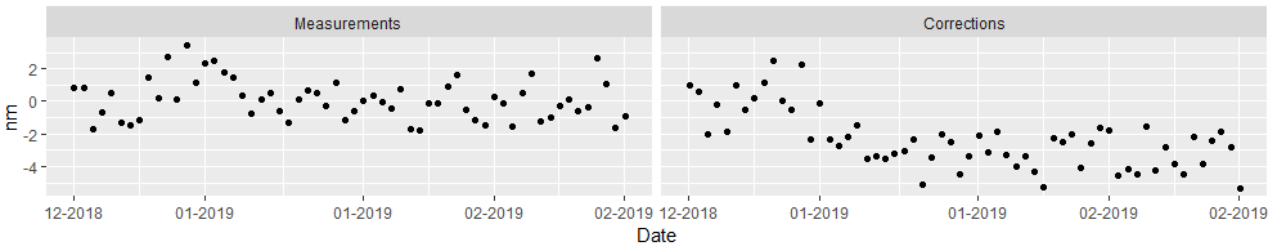


Figure 5.2: An example of a corrected system with variance

us two cases: machines for which we have a 234 variables and machines for which we have 498 variables. This is a large number and therefore we want to reduce the dimension of this data.

5.1.2 Dimension Reduction

Detecting performance issues in data with these dimensions can give problems such as

1. Estimating a large covariance matrix
2. Noise introduced by the large amount of variables
3. Monitoring them separately could cause many false alarms

Therefore it is important to reduce the dimension of the data. Another reason to reduce the dimension could be to make it easier to visualize behavior of the lens.

Reducing the dimension could be done by selecting only the variables in which we believe a performance issue with the WELLE sticker is shown. This however is difficult to determine and could result in selecting the wrong variables. Important information for detecting failures could be missed if we use this method. A better way of reducing the dimension is to transform all the data to new variables which represent the original variables. Therefore we use principal component analysis.

The principal components depend on the data that is selected. We have the following options:

Healthy data Model healthy behavior, such that abnormal behavior is highlighted.

All data Model general behavior, such that we can see the effect of a failure.

For the first option it is needed to have only healthy data. In the lens case it is very difficult to understand if a machine is healthy or not. An assumption could be to use only data from a few months after a replacement. This however does not take other failures or existing drift into account. For the second option we expect that a lot of variance in the machines comes from the WELLE sticker replacement. This would mean the failure is captured in the first few principal components. Only those components have to be used to detect issues with the WELLE sticker.

We choose to use all the data to fit a PCA model. To determine the amount of PCs that needs to be retained we use the broken stick method, because this method works the best according to Jackson (1993). We also can decide on how many and what type of machines are used to fit the PCA model. We expect that the type of the machine and which Zernike coefficients are measured have an effect on how the lenses behave. Therefore we choose to sort the machines in different classes. This should reduce the variance caused by differences between machines. We want to create as few models as possible, but we also want the model to focus on failures instead of type to type differences. This gives us 9 different classes for which we have cases with a WELLE sticker replacement. For each class we fit a PCA model and we use all data from the machines with a replacement.

Class	Type	Zernike set	Amount of Machines	Replacement in Data
0	NXT:1950I	1	6	4
1	NXT:1950I	2	26	15
2	NXT:1950I	3	1 (machine 18543)	0
3	NXT:1965CI	2	4	4
4	NXT:1960BI	1	6	6
5	NXT:1960BI	2	4	2
6	NXT:1970CI	1	4	4
7	NXT:1970CI	2	1 (machine 15681)	1
8	NXT:1970CI NO FLEXWAVE	2	1 (machine 27861)	1

Table 5.1: The 8 classes based on type and Zernike coefficients

5.1.3 Classes in the Data

To reduce the variance caused by machine types we create classes of machines. These classes consist of machines which have the same type and the same Zernike coefficients measured. The different types of machines have different lens types. This could have a large influence on the measurements. In theory we have Zernike coefficients for Zernike polynomial 2 till 64 and then 4 coefficients for each to describe the full field of the lens. This means that there are potentially 252 variables for total drift and 252 variables for residual drift. In practice some coefficients are not measured. For total drift the following Zernike coefficients are not measured for all machines: Z2_0, Z2_1, Z3_0, Z3_1, Z4_0 and Z4_1. We define three Zernike classes:

1. The Six Zernike coefficients mentioned above are not measured. Meaning that there are 498 coefficients.

2. 270 Zernike coefficients are zero, these are the six mentioned above, Z26 till Z64 for total drift and Z38 till Z64 for residual drift. This gives a remaining 234 variables.

3. 504 Zernike coefficients are zero, meaning that nothing is measured.

This gives us 9 classes which can be seen in Table 5.1. It can be seen that only machine 18543 has nothing measured or logged. Most WELLE sticker replacements occur for machine type NXT:1950I. This might be caused by the amount of machines of this type. This is also a type that is longer in use than the other types. In the last column of this table, the amount of machines for which we have data containing the WELLE sticker replacement is given.

When analyzing these class we will ignore classes 2, 7 and 8 because these have only one machine. The main focus will be on class 1 because this classes is well represented.

5.1.4 Detecting Changes

Using our transformed data we want to detect changes, for this purpose we can use machine learning techniques or statistical process control techniques.

The focus will be on SPC techniques because, those are specifically designed to detect sudden changes or trends in time series data. While machine learning often ignores time in the data. Furthermore there are no labels which forces us to use semi-supervised or unsupervised machine learning. For semi-supervised machine learning we need to make assumptions about which data is healthy and which is not and unsupervised machine learning is difficult to evaluate.

A downside of traditional SPC methods is that they require a phase I, which is difficult to determine. This is because the complexity of the machines. Therefore we will use self-starting SPC methods, these do not require a phase I. We will use univariate methods for this project. The use of multivariate methods could be an improvement.

We will investigate three different methods, these are the Shewhart, EWMA and the CUSUM. We expect that the EWMA gives the best results because it can detect drift and it is not needed to know what kind of changes you want to detect. Shewhart only detects sudden jumps. We expect that these are present at the replacement of the WELLE sticker, but not in the remaining data. CUSUM requires an idea of the size of the change. As long as we do not know this we can only guess and hope the result is good.

5.1.5 Diagnosing

If we find an out of control signal based on SPC, than we want to know why we found that signal. Some explanations could be a calibration of the system, a failure or a false positive. If it is a failure we also would like to know if a similar issues already occurred, such that we can apply the same solution. To do this first a database of existing signals must be created. A machine learning algorithm can decide if our new signal is similar to an existing signal. If it is similar we know how to solve the issue, otherwise we can add the new issue to our database of signals.

5.1.6 The Model

The complete model has as input: the class of the machine, SyCo total drift and SyCo residual drift. Based on the class of the machine we decide which PCA model we have to use to transform the data. The output will be the date at which a signal is found. If the model is used online, than the output will be that something is wrong as soon as that is detected. Furthermore we want as output, why the signal is given. If a similar situation has occurred than the output will tell us which similar situation. If it is a new problem than the output will be that it is a new problem.

5.1.7 Verification

To see if the model works we need a way of measuring how good it works on available data. Also if we want to compare different models we need a way of comparing the result. The only information we have is the date of a WELLE sticker replacement and a few dates of a calibration of the lenses. The method that we use should give not too many signals. Therefore ratio of signals per measurement is important. We are interested in signals that we can explain. All the signals before a calibration, explain why the calibration is needed. All the signals after a calibration explain the effect of a signal. Similar for the WELLE sticker replacement. To evaluate how good a model works an expert should look at the signals that are found and try to explain them. If a signal cannot be explained than it could be an indication for a failure.

5.2 Tools

RStudio (RStudio Team (2015)) is used, which is an IDE for R. R is a programming language used for statistical computing and visualization.

The dfphase1 package is used for analyzing the changes in the PCA data for individual machines (Capizzi and Masarotto (2018)).

Chapter 6

Analysis Lens Aberrations

In this chapter we will present the results of the suggested approach. First we will discuss some decision that are already made in ASML and their effect. We will give an example of a PCA fit for one machine. We will investigate fitting PCA on all class 1 data and methods to analyze this. Then we will see how SPC can be applied on that machine. As last we present the general results for all the classes.

6.1 Analyzing the Zernike Fit

In Section 2.3 it is explained how we get the coefficients of the linear model for each Zernike polynomial. These are measured on a grid of 5×13, where first the mean is taken over the 5 values. Does it make a difference if we omit the taking the mean step? In Appendix A.1 the general fitting procedure and the procedure by first taking the mean is analyzed. In the SyCo situation we have that for each group we have similar amount of measurements. This means that there is no difference in the mean of the estimate, but the variance of the model is different. So using the mean procedure gives the wrong variance for the overall model. This also results in different variances of the model estimators. For ASML the most important is that the estimations are correct. Therefore they need to keep the groups on the grid the same size. If the variance is used for model evaluation or confidence intervals for the estimations than this will give a wrong result.

6.2 Transformation to Orthogonal Polynomials

The SyCo data is reported in four linear model coefficients. This model is fitted with regular polynomials meaning $y = \beta_0 \cdot 1 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3 + \varepsilon$ The polynomials $1, x, x^2$ and x^3 are not orthogonal. In Appendix A.2 a transformation is calculated such that the coefficients can be transformed to coefficients fitted with orthogonal polynomials. The result is that for each set of 4 Zernike polynomial coefficients ($\beta = [\beta_0, \beta_1, \beta_2, \beta_3]^T$) we have to multiply with

$$\beta' = \begin{bmatrix} 3.605551 & 0 & 2.268671 & 0 \\ 0 & 2.860036 & 0 & 3.213537 \\ 0 & 0 & 2.010962 & 0 \\ 0 & 0 & 0 & 1.367278 \end{bmatrix} \cdot \beta.$$

So now we have found a matrix that can be used to transform the data such that the coefficients are independent. It can be seen that β_2 and β_3 only change in scale. This transformation could be used as an improvement in the analyze of the variables. For our goal this is however not important, because we already transform the variables. The variables β_2 and β_3, are only

scaled. For PCA we have to normalize these variables. This would mean that the transformation to orthogonal polynomials is undone to use it for PCA. Therefore we do not use this transformation.

6.3 Principal Component Analysis

In this section we analyze the result of PCA, where we fit a model on a single machine and on a classes of machines. We will analyze the machine 18648, which is a machine of class 1. Our main goal is to create a few new variables in which the WELLE sticker replacement is clearly visible.

6.3.1 Analyzing Individual Machines

We hope to see a clear difference between data before and data after the WELLE sticker replacement in the machine. We use the data of the machine in the following order

$$[Z2\_0_R, Z2\_1_R, ..., Z64\_2_R, Z64\_3_R, Z2\_0_T, Z2\_1_T, ..., Z64\_2_T, Z64\_3_T],$$

where each variable is a vector with all measurements in time, R stands for residual drift and T for total drift. All the coefficients which are not measured should be removed. The columns are normalized such that they are all equally important in the PCA.

For class 1 machines this means that there are 234 variables to analyze simultaneously. We fit a model on only this machine and we select the amount of components based on the broken stick method as discussed in Section 3.4.1.

When a WELLE sticker replacement happens we expect the Zernike coefficients to change. This change should introduce variation in the data, which should be captured by at least one of the selected principal components.

Machine 18648 had a WELLE sticker replacement on 02-14-2018. A plot of all the data of this machine is given in Section 4.4. In all the different plots there is a clear difference before and after the WELLE sticker replacement. The behavior after the replacement show signs of drift, especially in Z5 for the total drift. In the residual drift before the replacement there is in the tilt a clear jump around 6-1-2017 in Z6. There is a clear trend in some Zernike coefficients before and after the WELLE sticker replacement. There also is an interesting part in August 2017 in the total drift, where the behavior is suddenly different for a few days and then jumps back to what it used to be.

It is difficult to accurately describe what happens in this data because of the many variables. To visualize the data better we fit one PCA model on the residual drift and total drift set for the complete time. The broken stick method is visualized in Figure 6.1. Only the first 20 PCs are shown, the amount of variance that is explained in each PC is shown. The red line represents g_k, for $k = 1, ..., 20$. It can be seen that the broken stick method tells us to retain six principal components. By applying PCA we went from 234 variables to only 6 new variables. These six principal components explain 88,66 % of the variance in the data and can be seen in Figure 6.2. The vertical red line represents the WELLE sticker replacement. PC 1 shows an upward trend both before and after the replacement. At the time of the replacement there is a large shift in value of PC 1. PC 2 shows a faster upward trend both before and after the replacement and a shift in value at the WELLE sticker replacement. PC 3 shows a small shift at the WELLE sticker replacement and an upward trend after the replacement. Before the replacement the behavior seems to change, there are a few shifts in the level. There is a notable peak around August 2017. PC 4 starts with a downward trend then a small peak round August 2017. Then there is a slow upward trend. At replacement there is a small shift. After the

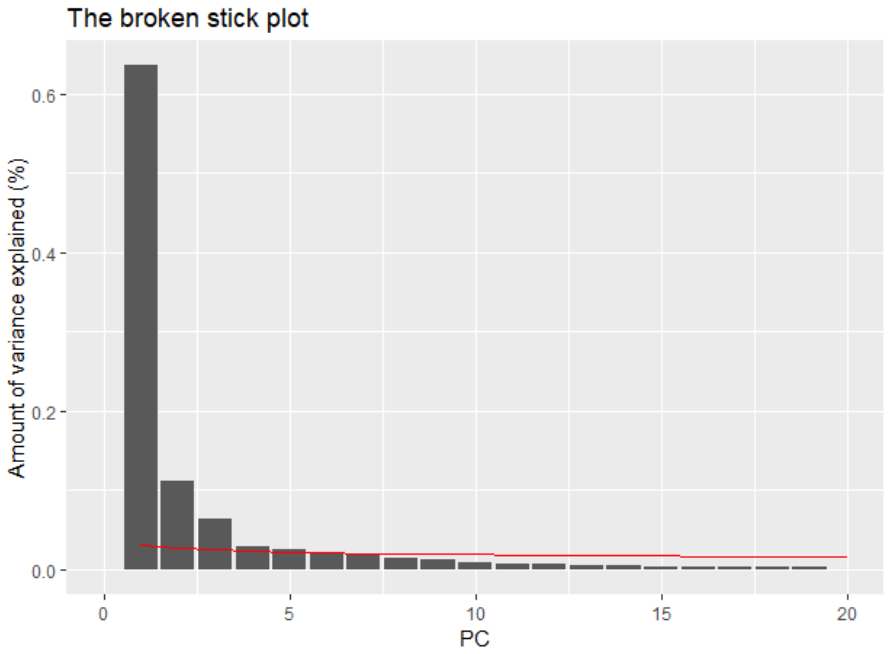


Figure 6.1: The broken stick method for machine 18648

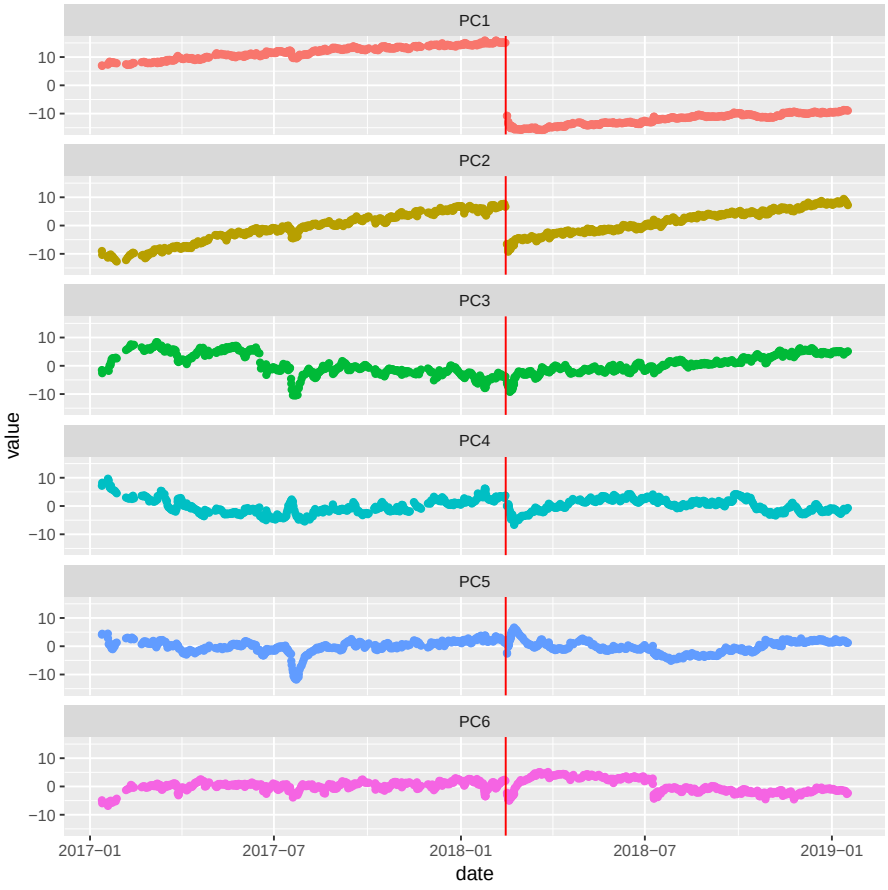


Figure 6.2: The principal components for machine 18648

replacement there is a small moment of change and then there seems to be somewhat stable behavior. PC 5 has a notable peak around August 2017. Before the replacement there seems to be stable behavior. At the replacement there is a small peak and after the replacement there is a parabolic behavior. PC 6 starts with an upwards trend. Then the level remains constant. At the replacement there is a small peak and at the end the process seems to be stable. Around July 2018 there seems to be a change.

To better visualize the changes we can use the mphase1 method from the R package dfphase1 (Capizzi and Masarotto (2018)). This method has as input how much change points should be detected. We choose 10 change points and the result is given in 6.3. The biggest change is found at the moment of the WELLE sticker replacement. Then this method mostly detects changes that are caused by the trend in the data.

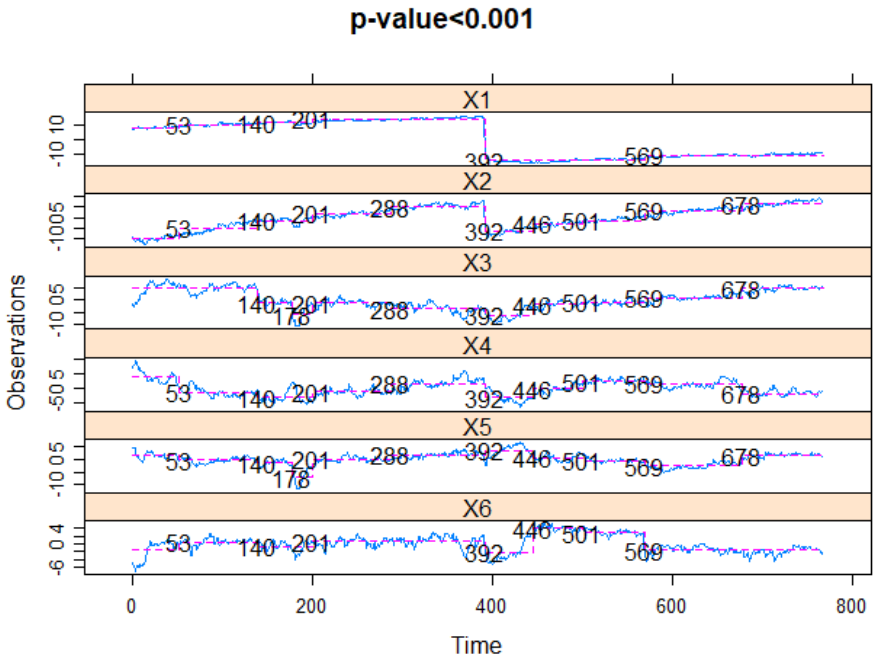


Figure 6.3: The change point detection method mphas1 for machine 18648

It finds in all PCs signs of a change at the WELLE sticker replacement (observation 392). A clear difference in the behavior before and after the replacement can be seen in PC 1 and PC 2, because the level is different. This shows that it is possible to visualize the WELLE sticker replacement. Another observation is that in all principal components something happens at, 06-17-2017 (observation 140), 07-18-2017 (observation 201) and 07-09-2018 (observation 569).

A method to handle the auto correlation and the dynamics of the data is dynamic PCA Shi and Tsung (2003). We apply dynamic PCA by also adding the previous measurements to the PCA. The principle components can be seen in Figure 6.4. There is an additional principal component. It can be seen that the values on the y-axis increase, but the behavior of the principal components remains very similar. After the WELLE sticker replacement there is one value in between the large shift in PC 1 and PC 2. This is because the first part of the data is data after the WELLE sticker replacement and the second part of the data is before the WELLE sticker replacement. Based on visual inspection adding a time lag does not reveal more information of the machine.

To summarize we have seen that applying PCA gives us less variables. The new variables lose their interpretation, because we do not know which original variables are responsible for which behavior in the PCs. But the general behavior of the system is highlighted in only a few variables. Using PCA on an individual machine could be used when the interest is in the

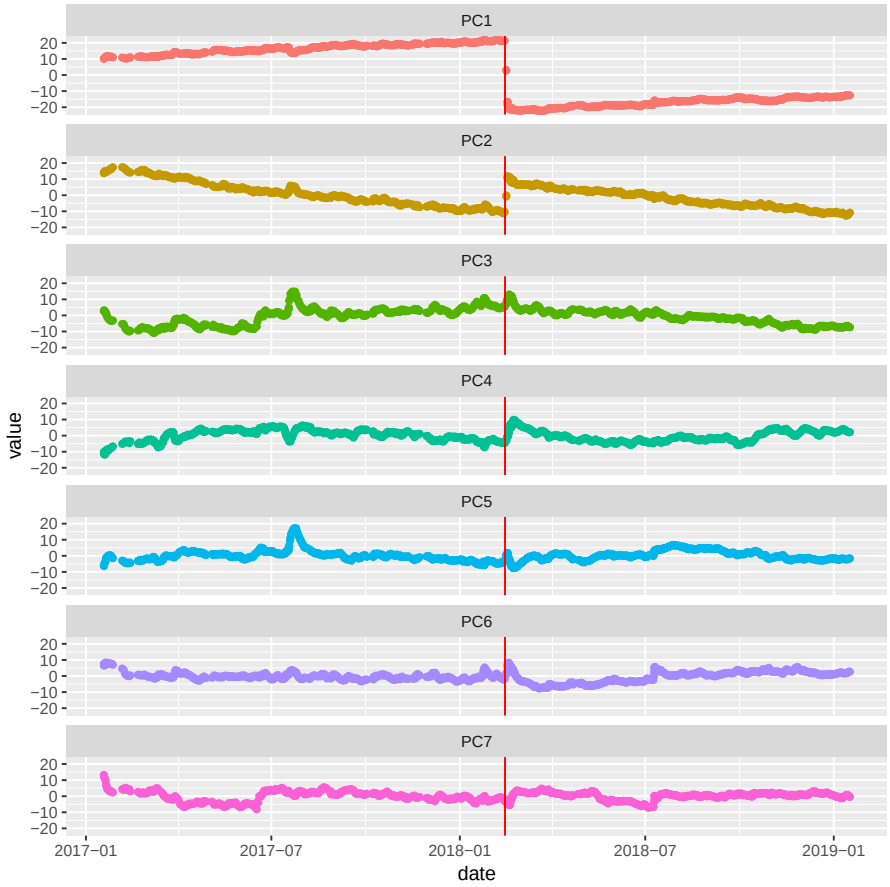


Figure 6.4: The PCs with lag 2 for machine 18648

behavior of an individual machine. For detecting abnormal behavior for multiple machines we however want one model for each class. Therefore we will now analyze PCA for multiple machines.

6.3.2 Analyze Multiple Machines

We fit one PCA model on a class of machines. Our example machine is 18648 and this is a class 1 machine. The same data representation as for individual machines is used. This gives a total of 15 principal components based on the broken stick method. Together they explain 80,9% of the variance in the data. We first visualize the first two principal components for all class 1 machines. This can be seen in Figure 6.5.

Note that the machines have a different amount of measurement in different time frames, especially in the ”12 weeks before swap” and ”after lens swap” plots. Also some machines have large gaps around the replacement which means that not all machines have data in the plots ”3-0 weeks before swap” and ”3-0 weeks after swap”. The most interesting behavior is the change of points before and after the WELLE sticker replacement. It can be seen that in 0-3, 3-6 and 6-9 weeks after the replacement the points seems to be in one cluster. It can also be seen that there seems to be almost no change in the behavior in the 12 weeks before the replacement. Some machines change a lot after the replacement. While some seems to remain stationary the whole time, this would mean that there are different WELLE sticker failure fingerprints. For our machine 18648 we can see that there is a clear jump after the WELLE sticker replacement. There is also a clear drift after the replacement. We will show the time evolution of machine 18648 fitted with the new model. We hope that the same information as before is also captured in this new PCA model.

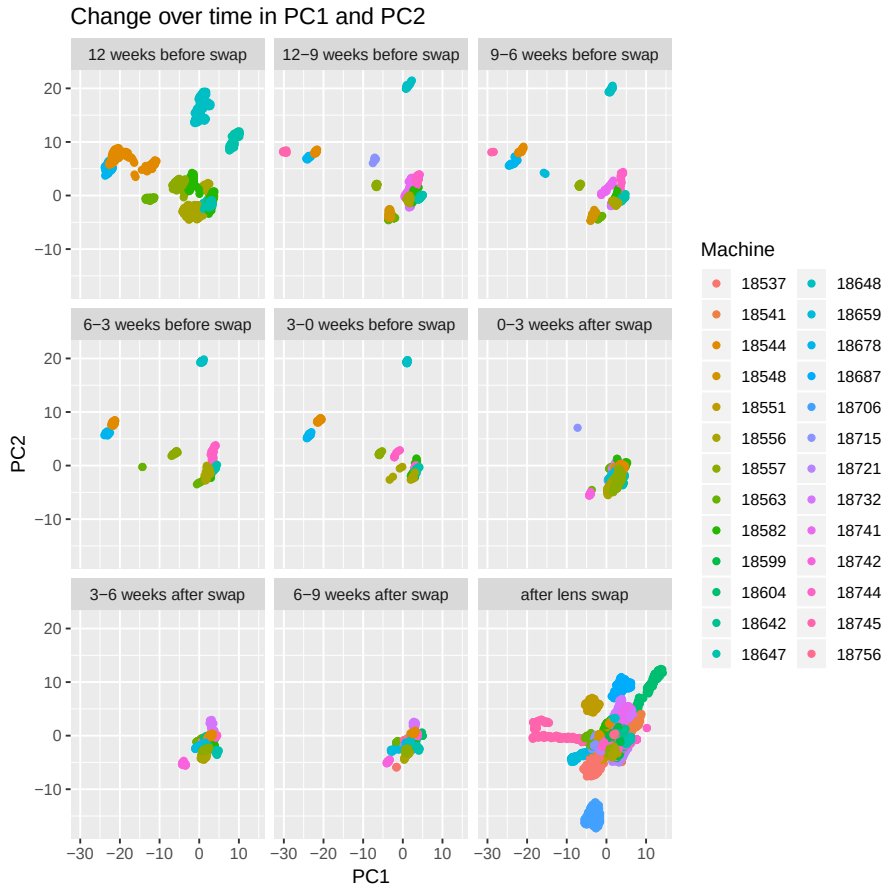
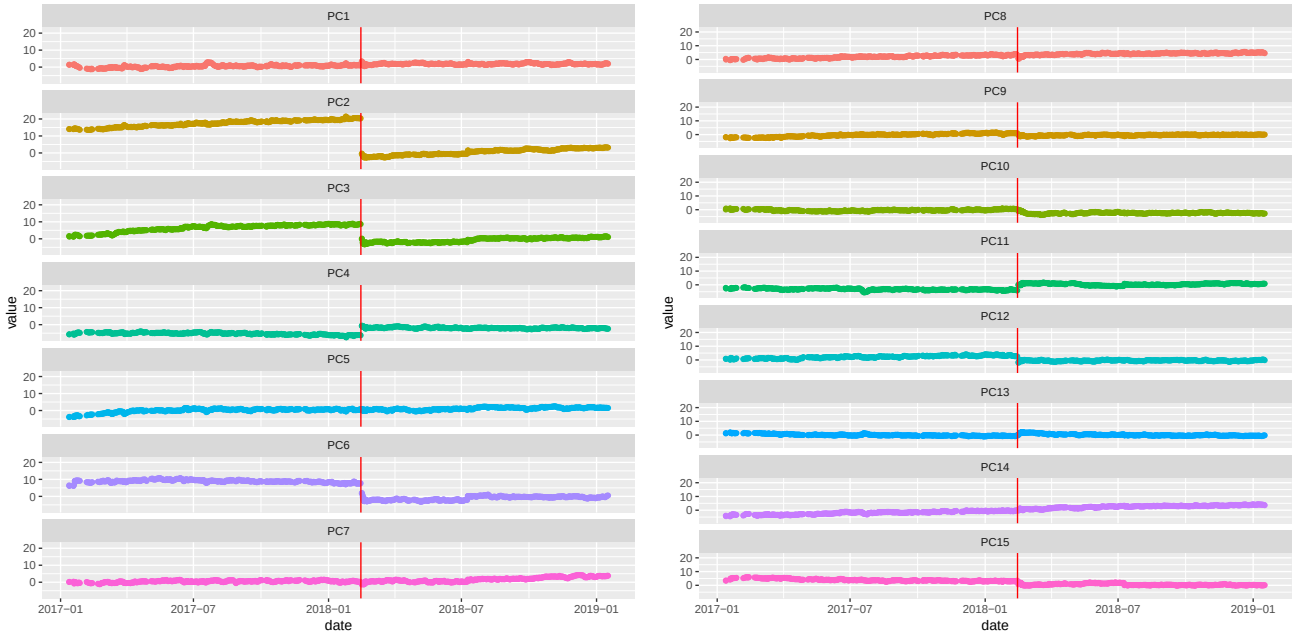


Figure 6.5: PC1 and PC2 fit of all class 1 machines over time



(a) PC 1 till PC 6

(b) PC 7 till PC 15

Figure 6.6: Machine 18648 fitted with the class 1 PCA model

The behavior of the 15 PCs in time for machine 18648 are given in Figure 6.6. A drift is clearly visible in this machine in PC 2, PC 3, PC 8 and PC 14. In the individual machine there was a clear peak at 07-18-2017, this is clearly visible in PC 1, PC 3, PC 11 and PC 13. So the

interesting patterns we found before are also in this new model.

To investigate differences between machines in each principal component we could use a box plot. A box plot is method to visualize different groups based on the quantiles of the data. In our case the machines are the different groups. We give as example the box plot for PC 1, where we make a split in the data between before and after the replacement, such that the differences are clearly visible. A box plot could help us identifying which machines behave similar and to derive the meaning of the principal components. The box plot is given in Figure 6.7.

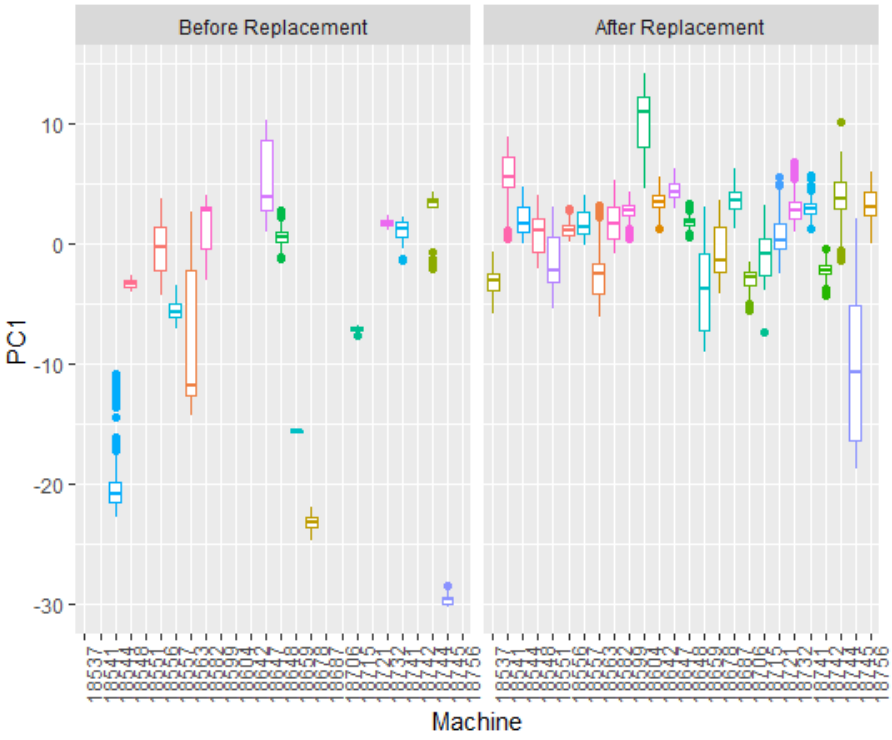


Figure 6.7: Box plot of PC1 for machines of class 1

There are large differences in mean and variance between machines for PC 1 in the before plot. Machine 18544 shows a lot of outliers. Machines 18563 and 18647 have a larger variance than the others. Machines 18648, 18732 and 18741 are the most similar in before. In the after plot there seems to be much more similarities. Most machines have a mean between -5 and 5. The machine that shows the most deviation is 18745. This PC seems to explain the difference before and after the replacement, this can be seen because of the difference between before and after.

The principal components are difficult to interpret, we suspect that some principal components explain the differences between machines, some explain general behavior and some explain the WELLE sticker replacement. There seems to be different PCs that explain the WELLE sticker failure in different machines. This means that there is not one single fingerprint for a WELLE sticker failure. The PCA fit on a class of machines does not remove information. We will now try to detect changes in this data.

6.4 Detecting Change

After applying PCA we have a data set with for class 1 only 15 variables. It is easier to detect changes in these 15 variables than in the original variables. We start with an intuitive example of a machine learning approach. After that we investigate the self-starting Shewhart, EWMA and CUSUM. We are especially interested in

1. explain changes that we find in general, with calibration of the lenses.
2. See how many signals different methods give before the replacement.

6.4.1 Machine Learning

We start with an example of how we could detect changes given our new PCA data. We have already seen how PC 1 and PC 2 evolve over time in Figure 6.5. This example is a machine learning approach in which we only use the first 2 PCs. For simplicity we use a classifier in the shape of a rectangle. We now make the assumption that all data 3 weeks after the replacement is healthy. We create a rectangle such that all healthy data is inside the rectangle. Now we can classify other points based on this rectangle.

Inside the region Classified as healthy

Outside the region Classified as unhealthy

The rectangle is drawn in Figure 6.8. It can be seen that following this idea that in the first

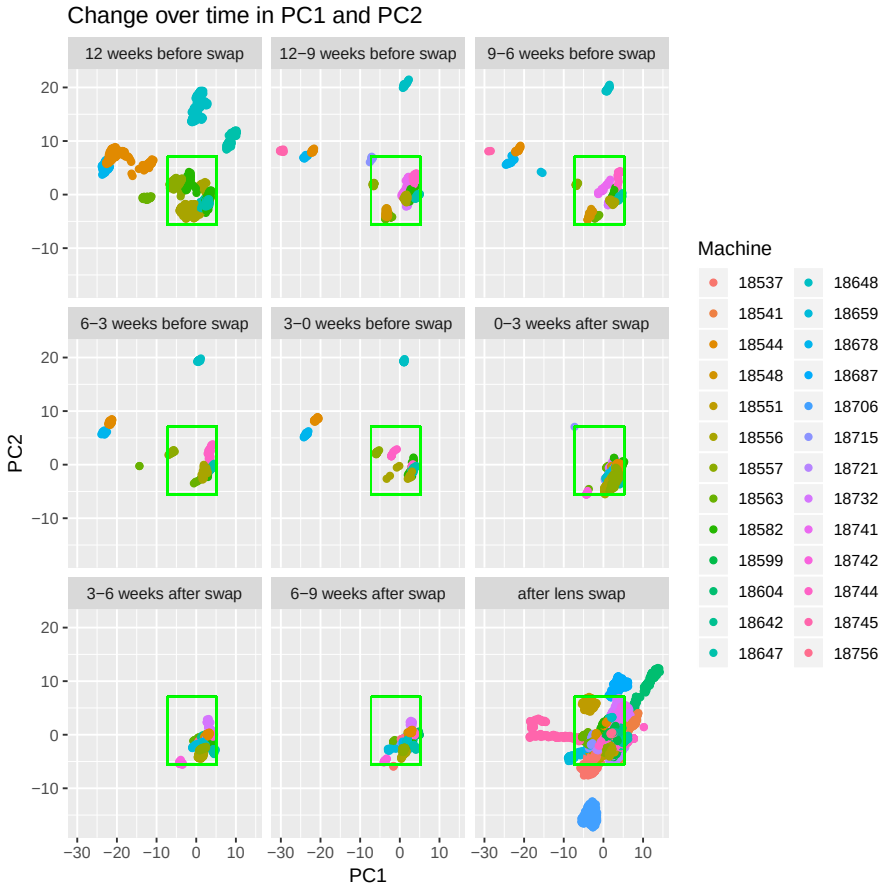


Figure 6.8: An example of a detecting mechanism for PC 1 and PC 2

9 weeks after the replacement the machines would still be classified healthy, but after a while we see a lot of measurements outside of the healthy box. A few machines has data classified as healthy before the replacement. So for those machines this method would not work. There are some machines which make a clear jump caused by the replacement and for those machines this method could work.

To verify we could use machines without a replacement, these are shown in Figure 6.9. Data from machine 18594 remains completely in the green rectangle, the other 3 machines

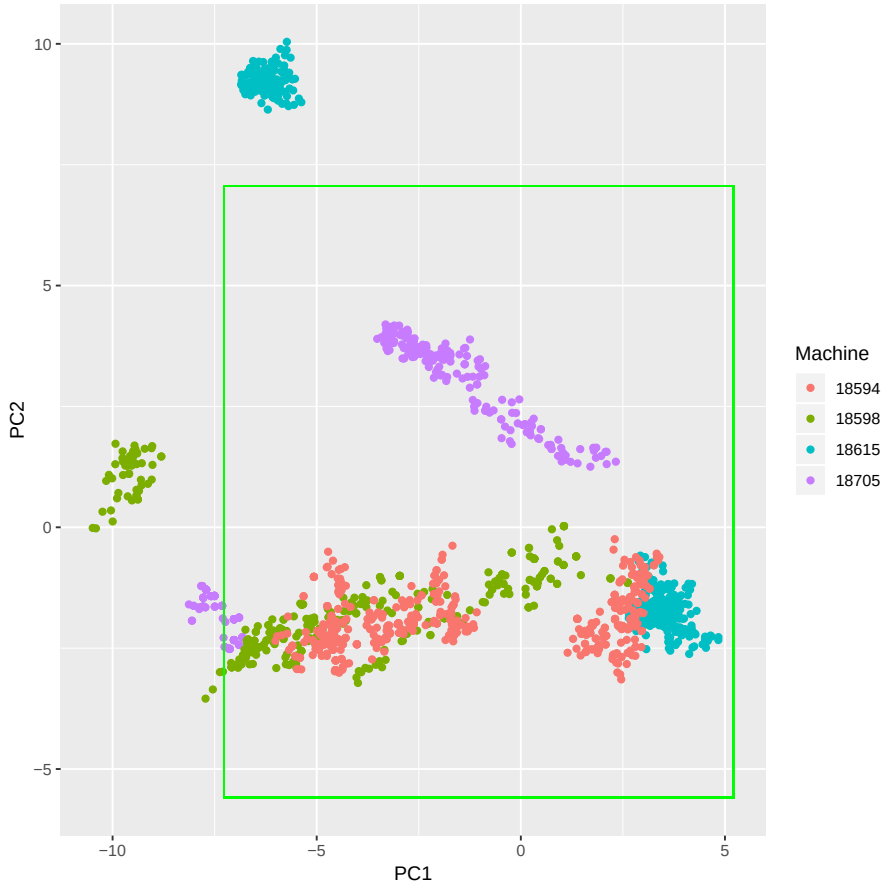


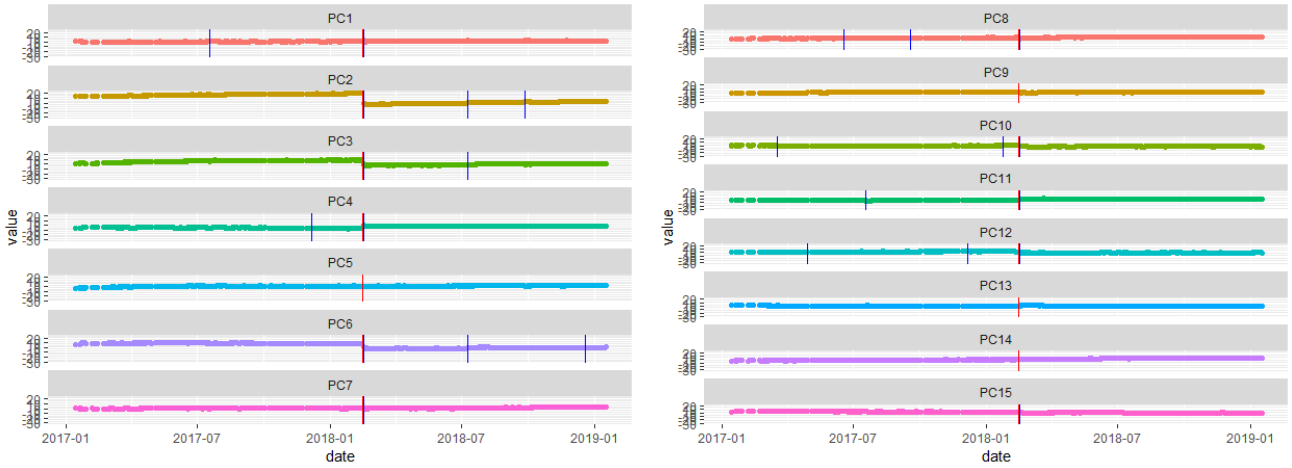
Figure 6.9: Healthy machines of class 1

have moments in which the machine is outside of the rectangle. Machines 18598 and 18615 start outside of the rectangle, but at some point they jump back in the rectangle. This could be caused by a calibration of the system or a part that is replaced. This does suggests that machines inside the region are healthy. While outside the rectangle the machines are still operational, but a calibration could improve the situation.

Note that this procedure is far from optimal and is used to show the ideas of a machine learning approach. There are many aspects that could be improved, examples are:

- Investigate different assumptions, label a long period as healthy or add machines without a replacement and label those as healthy.
- Also label the period before the replacement as unhealthy. Then it is possible to use a better fitting procedure in which for example the accuracy of the model is optimized.
- Use a training and test set, to prevent overfitting.
- Allow other types of regions, instead of only a rectangle.
- Use all dimensions instead of only the first 2.

Something that this model does not take into account is time. If measurements are slowly moving towards the boundary of the region and a new measurement suddenly is outside the region than it is detected that something is wrong. However often there are already clear signs that a change is occurring. Therefore we want to investigate SPC methods, because these methods could also detect these kind of trends in time.



(a) PC 1 till PC 6

(b) PC 7 till PC 15

Figure 6.10: Machine 18648 with self-starting Shewhart

6.4.2 Statistical Process Control

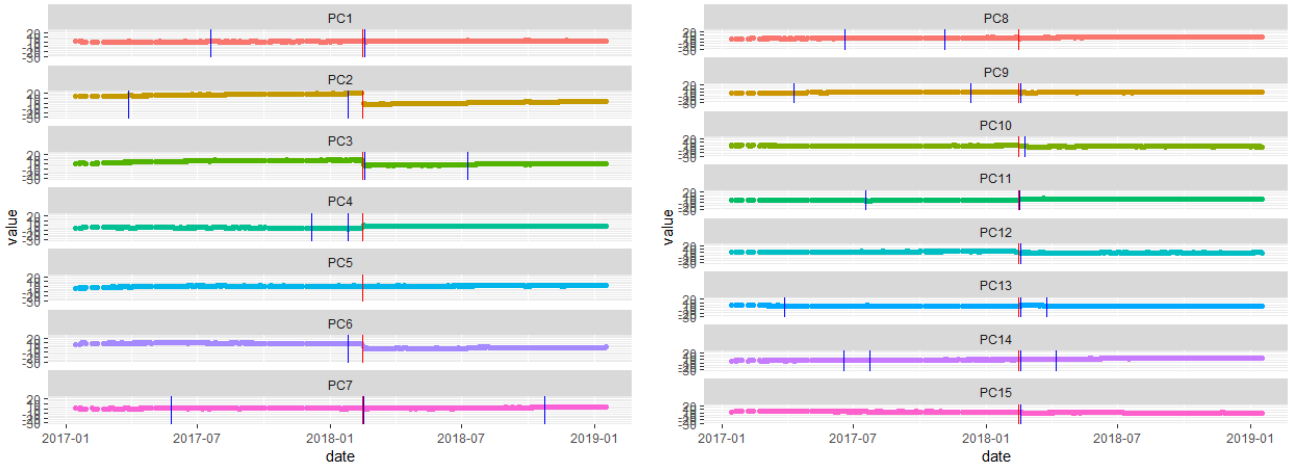
We start with applying three methods to the machine 18648. We start with the self-starting Shewhart method. We expect that this method does not find many signals, because it only detects large sudden changes.

For the Shewhart method we use the standard 3 sigma limits and after a signal is detected we restart the method. The result is given in Figure 6.10, where the vertical red line represents the replacement and the vertical blue line is the moment a signal is detected. In PC 1, PC 4, PC 8, PC 10, PC 11 and PC 12 there is a signal before the replacement. In PC 4 and PC 12 a signal is found at the same time. Another interesting finding is that PC 5, PC 9, PC 13 and PC 14 do not find any signal, while the others at least find a signal at the replacement. In PC 7 and PC 15 only the replacement is detected. In PC 2, PC 3 and PC 6 a signal is only found after the replacement. So this method for machine 18648 detects different changes. It detects the WELLE sticker replacement. It detects 9 signals before the replacement. This is a reasonable amount, unfortunately we cannot verify if these are caused by calibrations. It also detects 5 signals after the replacement. So the Shewhart method, for which we suspected that it would not detect that much, detects a reasonable amount of signals.

For the EWMA chart we use $\lambda = 0.2$ and $L = 1.5$, after a signal we restart the method. The result for machine 18648 is given in Figure 6.11. In PC 1, PC 2, PC 4, PC 6, PC 7, PC 8, PC 9, PC 11, PC 13 and PC 14 a signal is found before the replacement. In PC 2 and PC 13 and PC 2 and PC 4 a signal at the same time is found. In PC 5 no signals are found. In PC 3, PC 5, PC 10, PC 12 and PC 15 no signals before the replacement are found. The EWMA detects 15 changes before the replacement and 4 signals after the replacement. This is more than the Shewhart method, some changes are found by both methods. But for example the changes in PC 10 and PC 12 are not found by the EWMA, but are found by the Shewhart.

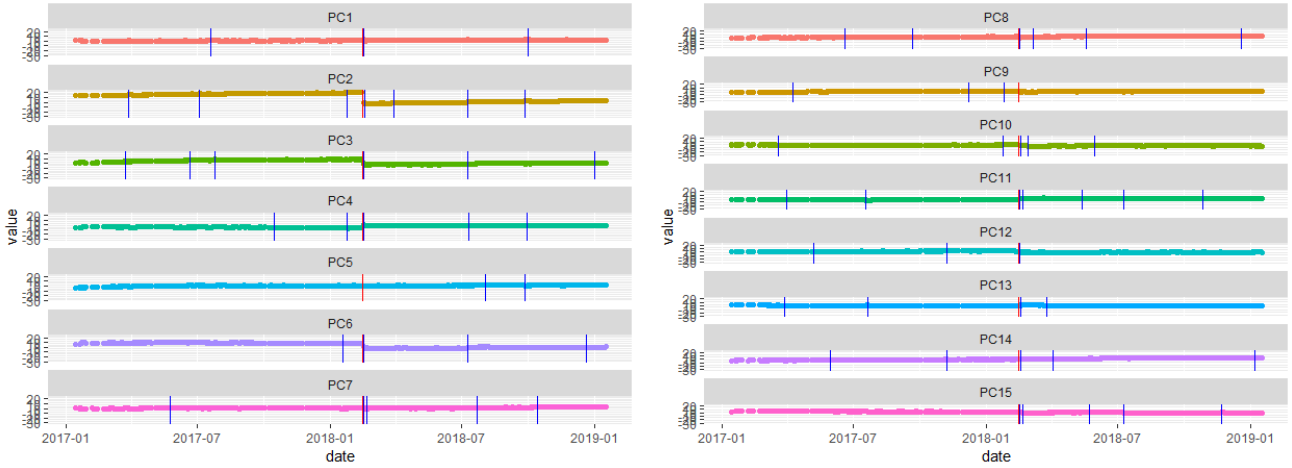
For the CUSUM chart we try to find a shift from $\mu_0 = 0$ to $\mu_1 = 3$, where $k = \frac{|\mu_1 - \mu_0|}{2}$ and the limit is 5, again we restart the method after a change is detected. The result can be seen in Figure 6.12. There is no signal detected in PC 5 and PC 15 before the replacement. There are multiple signals before the replacement in most PCs. In PC 5 and PC 9 the replacement is not detected. There are more changes detected in this CUSUM compared with the other methods.

It can be seen that with the parameters we choose there are quite some differences between the three methods. All methods however detect changes in the PCs before the replacement. The three methods all give a signal at 2017-07-19 in PC 1, we already detect this moment earlier with the visual inspection. The Shewhart chart signals less than the EWMA chart,



(a) PC 1 till PC 6 (b) PC 7 till PC 15

Figure 6.11: Machine 18648 with self-starting EWMA



(a) PC 1 till PC 6 (b) PC 7 till PC 15

Figure 6.12: Machine 18648 with self-starting CUSUM

which signals less than the CUSUM chart.

The result of these methods can also help us with understanding the PCs, for instance in PC 5 no signal is found for two methods and only for the CUSUM two signals after the replacement is found. This means that for machine 18648 PC 5 does not explain the replacement or any other events which would cause a change.

To further investigate the meaning of the signals for machine 18648 we use the dates of a calibration to explain the signals. These dates are only available from 02-03-2018 until 02-01-2019. We are interested in signals before and after the calibrations. Therefore we sort each signal in

- 2 to 1 weeks before the calibration
- 1 week before the calibration
- 1 week after the calibration
- 3 days before or after the replacement
- unexplained

Note that each signal can be assigned to only one group and that we first check for calibrations and then for the replacement. It is almost always the case that a calibrations happen around a replacement, this could mean that the replacement explains almost no signals in our case.

We also want to investigate the effect of different parameters for our three methods. Therefore we use different parameters for our methods. For Shewhart we investigate the following control limits:

1. $(-3\sigma, 3\sigma)$
2. $(-4\sigma, 4\sigma)$
3. $(-5\sigma, 5\sigma)$

For EWMA we investigate the following control limits

1. $L = 1.2$
2. $L = 1.5$
3. $L = 1.7$

For CUSUM we investigate the following shifts

1. $\mu_0 = 0$ and $\mu_1 = 2$
2. $\mu_0 = 0$ and $\mu_1 = 1.5$

These different methods are analyzed for machine 18648 in the Table 6.1. We used data starting at 02-03-2018 until 01-16-2019 and the replacement happened on 02-14-2018. There were 10 calibrations in this time. We use all 15 principal components individually. It can be seen that

Method	Parameter	2 weeks before calibration	1 week before calibration	1 week after calibration	3 days before swap	3 days after swap	Other signals	Total signals
Shewhart	1	0.00	0.60	0.35	0.00	0.00	0.05	20
Shewhart	2	0.00	1.00	0.00	0.00	0.00	0.00	7
Shewhart	3	0.00	0.00	0.00	0.00	0.00	0.00	0
EWMA	1	0.02	0.08	0.43	0.00	0.00	0.47	49
EWMA	2	0.00	0.08	0.58	0.00	0.00	0.33	12
EWMA	3	0.00	0.00	0.33	0.00	0.00	0.67	3
CUSUM	1	0.00	0.00	0.00	0.00	0.00	1.00	1
CUSUM	2	0.00	0.00	0.60	0.00	0.00	0.40	30

Table 6.1: The explanation for signals from different SPC methods for machine 18648

by changing the parameters of the methods we can determine how many signals are found. Because the methods with higher limits find less signals. It can be seen that the Shewhart method seems to detect changes that are either caused by a calibration or changes that require the need of a calibration. EWMA and CUSUM mostly detect a signal after a calibration, but also seem to detect some changes that cannot be explained. In this data it is expected that most signals are explainable, because most of the data is after the replacement.

We now use the different methods on 4 months of data before the replacement to see if these methods can detect changes which could explain the need of a replacement. We are interested in the ratio $\frac{\#\text{ Signals}}{\#\text{ Measurements}}$. This ratio tells us how often a signal is found. We want to have a

Method	Parameter	signal to measurement ratio
Shewhart	1	0.0148
Shewhart	2	0.0025
Shewhart	3	0.0000
EWMA	1	0.1010
EWMA	2	0.0271
EWMA	3	0.0074
CUSUM	1	0.0025
CUSUM	2	0.0567

Table 6.2: The results for different SPC methods on 4 months of data before the replacement for machine 18648

method with a low ratio, but it should not be zero. In Table 6.2 the result for machine 18648 can be seen. The Shewhart with control limits (-3,3) do not find a signal for this machine. The other Shewhart methods, the CUSUM 1 and the EWMA 3 have a relative low ratio. This shows that for this machine SPC can detect signals in 4 months before the replacement. How useful these signals are is difficult to determine, because there is no information available to explain these signals.

One of the results is very interesting, since it is not what we expected. This is the fact that the the Shewhart method finds changes. We expected that this method would only find large sudden changes and that these are not in the data. It turns out that these are in the data. Even more interesting is that it seems to occur before or after calibrations. This means that based on machine 18648 we can conclude that the Shewhart method could be used to detect the need of an calibration. Other methods found different signals that were not explained by the calibrations.

These results are only based on machine 18648 of class 1. In the next sections we will investigate this more general for all classes except 2, 7 and 8, because they only have one machine with a WELLE sticker replacement. We start with class 1 because it is the most represented class.

6.5 Class 1

We start with fitting an individual PCA model on all machines, to see if there are differences between the amount of principal components. For each machine the amount of PCs, the total variance explained and if the replacement is in the data is given in Appendix B.2 in Table B.3. It can be seen that in all cases the amount of PCs is less than 9, which means that in all cases the amount of variables is drastically reduced.

For most machines it holds that the replacement is in the first PC, but for some machines there is some other large event in the first PC.

It is interesting to see that for most machines with a swap in the data it holds that they retain more PCs and have a lower variance explained than the machines without a swap in the data. This is something that is expected because the replacement adds more variance to the data.

For most machines there also seems to be dates at which behavior changes. For two machines other parts are replaced in the data. These are machine 18678 which had at 10-8-2018 the variable attenuator replaced. For machine 18556 the variable attenuator was replaced at 12-10-2018. Both machines did not show this clearly in the data

As already mentioned for class one we find 15 principal components, which explain 80.85 %

of the variance.

The results of different SPC methods applied to only the time span for which we have calibration dates is given in Table 6.3. For the Shewhart 1 we can see that there are quite

Method	Parameter	2 weeks before calibration (%)	1 week before calibration (%)	1 week after calibration (%)	3 days before swap (%)	3 days after swap (%)	Other signals (%)	Total signals
Shewhart	1	0.09	0.24	0.19	0.00	0.00	0.48	348
Shewhart	2	0.07	0.48	0.16	0.00	0.00	0.29	92
Shewhart	3	0.05	0.51	0.23	0.00	0.00	0.21	43
EWMA	1	0.10	0.13	0.18	0.00	0.00	0.59	958
EWMA	2	0.11	0.13	0.22	0.00	0.00	0.54	385
EWMA	3	0.09	0.16	0.30	0.00	0.00	0.46	203
CUSUM	1	0.09	0.20	0.36	0.00	0.00	0.35	120
CUSUM	2	0.11	0.13	0.21	0.00	0.00	0.55	517

Table 6.3: The explanation for signals from different SPC methods for class 1 machines

some signals unexplained, while for the other Shewhart methods most signals are explained. The EWMA methods find the most unexplained signals. The CUSUM 1 method finds signals that can be explained. So these results are similar to those of machine 18648, in the sense that Shewhart signals are explained by calibrations. The results of SPC approach applied to 4 months before the replacement are given in Table 6.4. Here we see that in order of least signals

Method	Parameter	class 1
Shewhart	1	0.0271
Shewhart	2	0.0047
Shewhart	3	0.0022
EWMA	1	0.0968
EWMA	2	0.0390
EWMA	3	0.0209
CUSUM	1	0.0121
CUSUM	2	0.0516

Table 6.4: The ratio of signals before a replacement for class 1

the first 5 are: Shewhart 3, Shewhart 2, CUSUM 1, EWMA 3 and Shewhart 1. In comparison with only machine 18648 we can see that Shewhart 3 does find some signals in other machines. However ratio of 0.0022 is too little to reliable detect changes.

For class 1 we see that different SPC methods can detect interesting signals. Were we can use Shewhart to detect the need of a calibration.

6.6 Comparison between Classes

In Tables B.1,B.3,B.4,B.6,B.8 and B.10 the results of the individual PCA fits for all the different classes are given. The amount of principal components seem to be similar for most classes. Only class 5 deviates from the rest, because it has three machines with 6 PCs and 1 machine with 5 PCs, while the rest has also machines with more principal components. The highest amount of principal components is 10 in class 6. In classes 0, 1, 3 and 4 there are machines with 9 PCs.

There is one machine that clearly deviates from all the other machines. This is a machine in class 0, which has only 1 principal component which explains 52.39 % of the variance. Furthermore in class 4 there is a machine with only 3 PCs and in class 1 there are two machines with 4 PCs.

The result of one PCA fit for all machines for different classes can be seen in Table 6.5. The

Class	Amount of PCs	Variance explained	Amount of machines
All	20	73.36	50
0	13	87.24	6
1	15	80.85	26
3	11	86.31	4
4	13	87.25	6
5	8	86.78	4
6	11	90.33	4

Table 6.5: The multiple machines PCA for different classes

difference in the amount of PCs seems to relate to the amount of machines which are used. The less machines the less principal components are needed. The only exception is class 5, because it has 4 machines while it only needs 8 PCs. Note that class 5 was also lower than the others in the individual PC fit. What we see here is that there are still differences between machines although we split the machines on different classes. It can also be seen that if we use all machines that there are 20 principal components. So adding more machines adds more variances, which is expected. So the choice of splitting the machines in classes does reduce the type to type variation in the PCA model.

The results for the SPC models on data for which we have calibration dates is given in Tables B.2,B.5, B.7,B.9 and B.11.

Class 4, 5 and 6 have a high percentage of unexplained signals for all methods. While class 0 and 3 have a low percentage of unexplained signals. Class 1 seems to be in between these two groups with unexplained signals

class 5 and 6 seem to have more signals before a calibration. In all these classes there are not signals found 3 days before and after the replacement. This is caused by the fact that there is often no data around the replacement for a few days and if there is data than there is also a calibration which than explains the signal. For class 4, 5 and 6 it is needed to further investigate why these signals are not explained. It could be that because these machines have a different type, that they have another sort of calibration that would explain these signals. For classes 0, 1 and 3 it seems that the Shewhart method finds for all 3 classes explainable signals. The EWMA and CUSUM seem to find signals at different points. This could be explained by the fact that these methods also detect trends. Which could mean that they detect even earlier the need of a calibration.

The ratio of signals found in 4 months before the replacement for all classes can be seen in Table 6.6. It can be seen that all ratios are larger than 0. A ratio of 0.005 gives one signal each 200 measurement. In 4 months we expect at least one signal. This means that we hope to find a signal around every 100 measurements. So we want a ratio around $\frac{1}{100} = 0.01$. This gives us that the Shewhart 1, EWMA 3 and the CUSUM 1 are the best fits to get the desired amount of signals.

6.7 Summary

In this chapter we have seen that applying PCA gives us a transformation that represent the data well in a low dimension. Applying PCA on individual machines gives us a tool to

Method	Parameter	class 0	class 1	class 3	class 4	class 5	class 6
Shewhart	1	0.0140	0.0271	0.0292	0.0241	0.0124	0.0167
Shewhart	2	0.0051	0.0047	0.0060	0.0075	0.0035	0.0053
Shewhart	3	0.0042	0.0022	0.0007	0.0015	0.0014	0.0044
EWMA	1	0.0819	0.0968	0.0817	0.0866	0.0546	0.0925
EWMA	2	0.0340	0.0390	0.0397	0.0467	0.0214	0.0396
EWMA	3	0.0229	0.0209	0.0232	0.0248	0.0124	0.0247
CUSUM	1	0.0140	0.0121	0.0112	0.0173	0.0104	0.0203
CUSUM	2	0.0407	0.0516	0.0450	0.0520	0.0263	0.0608

Table 6.6: The ratio of signals before a replacement for all classes

analyze the lenses of the machine in a easier way. Self-starting SPC methods applied on the transformed data are able to detect changes in the data. Some of these changes can be explained by a calibration. The Shewhart method find larger sudden jumps in the data. Most of these jumps are found before a calibration is used. The CUSUM and EWMA detect changes after a calibration has occurred. When detecting signals before the replacement we saw that most methods except the Shewhart 3 find enough signals. It is unclear if these changes are able to explain the need of a replacement, because of the missing calibration dates.

Chapter 7

Summary and Conclusion

7.1 Summary and Conclusions

The main goal of this thesis is to find out if it is possible to timely detect a failure in the ASML machines. For this purpose we used the WELLE sticker failure case. SyCo total drift (Corrections) and SyCo residual drift (Measurements) are used as data, which are measured in terms of Zernike polynomials. The lenses are continuously being controlled, which could make it very difficult to detect a failure based on the measurements. Therefore both the measurements and the corrections are used, combing these two sets gives a maximum of 498 different variables, which could cause problems in most methods. Therefore the data needs to be transformed to a lower dimension, while retaining the information and filtering out the noise. A desirable outcome is a transformation that can be applied to all machines. This turns out to be difficult, because of the different machine types and the different amount of Zernike coefficients. To reduce the type to type variance, 9 classes of machines are created. Three of these classes consist of 1 machine and are omitted from the analysis. We used PCA to transform the original data and determined the amount of PCs with the broken stick method. PCA can be used on individual machines to analyze those machines in depth. This shows that the machines have changing behavior over time, caused by the control actions, calibrations, lens drift and replacements. When PCA is used on the classes it turns out that machines within a class have different behavior. This causes that if there are more machines used in one class, that in that class more principal components need to be retained. For class 1 this resulted in retaining 15 principal components instead of the 234 original variables. It is also the case that the WELLE sticker failure shows different fingerprints, this can be seen because for different machines the WELLE sticker replacement was clearly visible in different PCs.

The failures of interest are failures that gradually change the machine, such as wear or drift. These gradual changes are visible in the new variables in shifts or in a trend. The method that we use needs to be able to detect these changes over time. We briefly gave a machine learning approach, but concluded that it was not appropriate because it did not use time. The idea of this approach is to assign some region as healthy. As soon as the measurements of a machine drifts outside of this region it is classified as not healthy. This method is used because of the clear cluster that was visible when plotting PC 1 and PC 2 of class 1 machines. In these plots there is also a clear set point after the replacement. The values of the machine remain at this point for some time, but after a while they drift away from this point.

It is possible to detect this drift earlier, if we also take time into account. The statistical process control methods are able to detect change when it starts drifting away from the current state. Therefore this method can detect these issues earlier. An interesting result of the Shewhart method was that it was able to detect changes in the data. In class 0, 1 and 3 most of the signals of the Shewhart method can be explained by calibration, either a signals is found before

a calibration (The need of a calibration) or after a calibration (The effect of a calibration). The EWMA and the CUSUM have less explained signals, it could be that these methods detect the need of a calibration earlier than two weeks. These detected signals before a calibration, tells us that something changed in the machine. This change could be a sign of a beginning failure. We also tried to detect signals 4 months before a WELLE sticker replacement. It turns out that the Shewhart, EWMA and the CUSUM are all capable of detecting changes before the replacement. The amount of signals can be tuned to be more or less by changing the parameters. This shows that something happens in the period of 4 months of the WELLE sticker replacement, which could explain the failure.

We can summarize these results as follows:

1. Creating machines classes based on type and Zernikes measured reduces type to type variance for PCA.
2. PCA transforms the data and retrains at most 15 new variables.
3. PCA on SyCo data shows different clusters before and after the replacement.
4. Machines seem to be different even if they have the same type.
5. All machines show clear signs of changing behavior over time.
6. The machines are adjusted continuously. The largest changes are explained by calibration and replacements.
7. Univariate self-starting SPC methods find signals. This seems to detect the need of a calibration, but there are still more events that change the system.

The following key lessons are learned:

1. Control actions contain useful information about degradation and should thus be used in monitoring
2. SPC is suited for change in time (gradual changes), SPC is targeted at detecting change
3. Machine learning is more flexible, but cannot intrinsically deal with temporal aspects
4. Prediction procedures should be tuned to classes of machines

In Figure 7.1 we can see which requirements are met. We can see that we were able to distinguish between different behavior using PCA. We can detect abnormal behavior using self-starting SPC. We also found a transformation for the sensor data, which highlights the different behaviors. We partially succeeded in predicting failures, because it is uncertain whether the signals are found because of a failure or a scheduled event. We also partially succeeded in finding a general method, because it is unclear if this method works for all classes. We did not succeed in explaining why abnormal behavior is found. We also did not succeed in getting a good understanding of the performance of the algorithm. This is because there are too many signals which are unexplained.

7.2 Discussion

In this section we will discuss some of the choices which are made and some of the weaknesses of this approach.

Requirements	Achieved?	Comment
Distinguish between normal and abnormal behavior	✓	PCA makes it easier to distinguish between behavior
Detect abnormal behavior in an automated way	✓	SPC on the PCA transformed data is able to detect changes
Predict failures in an automated way	✓ ✗	Uncertain whether the signals predict a failure
Explain the abnormal behavior of the machines	✗	Abnormal behavior is not. Use ML, on a library of previous seen issues and their solutions, to classify new changes.
A general approach for multiple machines	✓ ✗	There is a need for classes based on the differences between machines. It is however not clear how good the method works for the different classes.
A transformation for the sensor data, which highlights different behavior	✓	PCA transforms the data such that this is visible
A good understanding of the performance of the algorithms	✗	There is no clear performance measure of this algorithm, since it is not clear what to detect

Figure 7.1: Summary of the requirements

Verifying the Result

It is difficult to verify if the signals that are found are false positives or signals of interest. This is because for most machines the calibration dates are missing before the replacement. It could also be that there are events in the machine that we missed which cause large changes. For example the expected life time of a WELLE sticker is around 2 years. This would mean that there should be more machines which had a WELLE sticker replacement. It could be that the replacement was part of a maintenance, but this data was not available. Events like these could help us explain more signals in the data, such that only the signals of failures remain. We choose to not label the data, because this would be heavily based on assumptions. This would however turn the problem into a classification problem. These are easy to verify. This is not done because if the labeling is not done correctly the results are not representative. An idea that has been presented in the approach, but which is not elaborated in this thesis, is the idea of detecting signals with SPC, and classifying the result with machine learning approaches. This method might find similarities between signals. This could be used to explain unexplained signals. Also if a new signal is very similar to an already occurred failure, the same solution could be used to solve the issue.

Class Creation

Another discussion point is the creation of the classes. Within each machine type there could be variation caused by different updates or different parts. It might be better to create even more classes, based on software updates, or certain versions of parts. Such that there is even less type to type variation in the classes. To understand if this improves the detection method, there is first the need for a better way to compare the results.

PCA Data

There is also the choice of using all machine data instead of healthy data to fit the PCA model. The choice for all data is made because it is unclear when the machine is healthy. A solution could be to use the following assumption: ”data after a replacement is healthy.”. Again it is

needed to have a way of comparing these two choices, because we want to know which data set gives a better result.

Effect of Applying the Model

It is unclear what the effects are if this method would be applied in the machines. Although it seems that we are capable of detecting the need of a calibration earlier than it is applied now, we do not know what the impact would be when doing this every time a signal is given. This is because if a change is detected we do not know how severe this change is. It could be that the change that is detected can be handled by the control actions without disturbing the production process. An unneeded calibration would stop the production, which would result in earning less money. At this point the method does not tell us how important a signal is. This is needed to be sure that the machine is not stopped unnecessarily. A hybrid approach where machine learning is used to detect what kind of signal (the severity) is detected could solve this issue.

7.3 Recommendations

There are some recommendations that could help to improve research for similar projects or help with further data analysis studies.

Follow-up Research

First of all it is important to continue with this research. We saw that these methods are able to detect changes, but it is needed to further investigate these methods before they can be used in practice. More on these improvements will be discussed in Section 7.4. If this method is used in practice, it could save a lot of money, because there are less unscheduled downs caused by a WELLE sticker failure. It also could convince customers to use SyCo daily, because it could save them money if a failure is detected early. This would result in more consistent measurements, which could improve the method even more.

Data storage

A list of possible improvements for data storage is given below.

Amount of data At this point most data is stored for only 1 year. This is not enough when dealing with parts that have an expected life time of 2 years. For the WELLE sticker case there is now more SyCo data available than 1 year for some machines. This is different depending on the cases. In our case the additional SyCo data is stored on a server from Diagnostics, but it is not accessible for other employees. This could also happen in other departments. Meaning that at this point a lot of data is already stored, but not in an uniform system.

Uniform storage system The dates of the replacements of parts are stored in different systems. An expert found a different list of machines with WELLE sticker replacements than those that are available on W2IN. It would be better to have one place where all data is available.

Machine data It is needed to have a good overview of everything that happened to the machine. So in the WELLE sticker case it is needed to have an overview of when calibrations are executed, when a part is changed, when the software is updated and all other kind of events that could explain changes in data.

Failure registration The failures are not registered in one complete list and there is no information about the reason of replacement. This issues should be addressed such that the data can be better explained. Also if there is more information available about the failure then this could be used for similar failures in the future.

Involve Experts

Experts need to be involved in these kind of projects. One of the desires of the customer support is to work with as few as possible experts when diagnosing the machines. This project is however about designing a method to detect failures. Involving an expert would improve the effectiveness of the process. This is because an expert

- can give a complete overview of all the involved process, such that nothing important is missed.
- understands which events cause certain changes in the data.
- might have access to different kind of data or has more knowledge of other interesting data.
- has a network of other experts which could help to gather even more information.

7.4 Future Research

For the future research it is needed to involve experts, such that a better understanding of the machine can be achieved. This better understanding is needed to be able to explain more signals. It is also needed to find out how severe a signal is, possibly by using different parameters for the SPC methods or by using machine learning methods and previous seen signals. It is also needed to develop a way of comparing the result. This could be done by asking an expert to label data or by labeling data based on an individual PCA. The kind of change which cause a failure could be analyzed, such that the CUSUM can be designed to better detect this specific change.

It could be investigated what the effect is of a calibration. For instance is it possible to detect from the data if a calibration improved the situation? These calibrations which did not improve the situation could indicate a failure. It could be an idea to restart the SPC methods after a calibration to see if there is a signal before the next calibration.

For this project we only used SyCo data, which is measured every day once. It could be that LoCo data (or other lens data) is better suited to detect changes. An advantage would be that the corrections are more precise than the SyCo corrections. Also failures that occur in a short time frame (one day or two days) could be predicted with LoCo. An issue with LoCo and the corrections is that the time stamps are not the same. So some corrections based on the LoCo data have a different time stamp than the LoCo time stamp.

It could be investigated what the effect is of using different training data for the PCA model. This could be important because in the literature it is often advised to only use healthy data. It could also be investigated what happens if machines are added without a WELLE sticker replacement. Another option is to explore dynamic PCA. It is suggested in the literature to use this for control systems.

In this research only univariate methods are used on all the individual PCs. It could be determined if there are PCs which are more important for detecting failures or calibrations. It could be investigated how multivariate methods behave on the PCs or on a set of PCs.

A combination of machine learning and SPC methods could be investigated. Where SPC is used to detect changes and machine learning to classify this change with for example a neural

network. Instead of using SPC methods on the PCA model it can also be used on standardized recursive residuals. Also the machine learning approach as presented in this thesis can be investigated. Where improvements could be made in

- Investigate different assumptions, label a long period as healthy or add machines without a replacement and label those as healthy.
- Also label the period before the replacement as unhealthy. Then it is possible to use a better fitting procedure in which for example the accuracy of the model is optimized.
- Use a training and test set, to prevent overfitting.
- Allow other types of regions, instead of only a rectangle.
- Use all dimensions instead of only the first 2.

But do note that although this method could improve the current detection system, it does not use time. Therefore it does only detect a trend in the data as soon as it hits the boundary of the unhealthy region, while the SPC approach is able to detect this trend earlier.

Bibliography

Z. Banko, L. Dobos, and J. Abonyi. Dynamic principal component analysis in multivariate time-series segmentation. *Conservation, Information, Evolution-towards a sustainable engineering and economy*, 1(1):11–24, 2011.

N. Bingham and J. Fry. *Regression: Linear Models in Statistics*. Springer Science & Business Media, 2010.

M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

R. Bro, K. Kjeldahl, A.K. Smilde, and H. Kiers. Cross-validation of component models: A critical look at current methods. *Analytical and bioanalytical chemistry*, 390:1241–51, 04 2008.

J. Camacho and A. Ferrer. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: practical aspects. *Chemometrics and Intelligent Laboratory Systems*, 131:37–50, 2014.

G. Capizzi and G. Masarotto. Self-starting CUSCORE control charts for individual multivariate observations. *Journal of Quality Technology*, 42(2):136–151, 2010.

G. Capizzi and G. Masarotto. *Phase I Distribution-Free Analysis with the R Package dfphase1*, pages 3–19. 01 2018. doi: 10.1007/978-3-319-75295-2_1.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

J. Chen and A. K. Gupta. *Parametric Statistical Change Point Analysis: with Applications to Genetics, Medicine, and Finance*. Springer Science & Business Media, 2011.

M. C. Chuah and F. Fu. ECG anomaly detection via time series analysis. In *International Symposium on Parallel and Distributed Processing and Applications*, pages 123–135. Springer, 2007.

R. B. Crosier. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30(3):291–303, 1988.

A. Di Bucchianico. Lecture notes 3tu course applied statistics, 01 2019.

A. Di Bucchianico and E.R. Van Den Heuvel. Shewhart's idea of predictability and modern statistics. In S. Knoth and W. Schmid, editors, *Frontiers in Statistical Quality Control 11*, pages 237–248, Germany, 2015. Springer.

A. Di Bucchianico and J.F.B. van Gellecum. Tools for achieving a successful SPC-EPC integration, 2007.

digitalphotographylive. Understanding chromatic & spherical aberration of lenses, 2012. URL http://digitalphotographylive.com/chromatic-spherical-aberration/. [Online; accessed January 2, 2019].

A. Ferrer. Latent structures-based multivariate statistical process control: A paradigm shift. *Quality Engineering*, 26(1):72–91, 2014.

J. Freeman, N. Vladimirov, T. Kawashima, Y. Mu, N. Sofroniew, D. Bennett, J. Rosen, C. Yang, L. Looger, and M. Ahrens. Mapping brain activity at scale with cluster computing. *Nature methods*, 11(9):941, 2014.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

M. Frisén. Methods and evaluations for surveillance in industry, business, finance, and public health. *Quality and Reliability Engineering International*, 27(5):611–621, 2011.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44, 2014.

M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, 11(4):1–31, 04 2016.

F. Harrou, F. Kadri, S. Chaabane, C. Tahon, and Y. Sun. Improved principal component analysis for anomaly detection: Application to an emergency department. *Computers & Industrial Engineering*, 88:63 – 77, 2015.

D. M. Hawkins. Self-starting CUSUM charts for location and scale. *The statistician*, pages 299–316, 1987.

D. M. Hawkins and D. H. Olwell. *Cumulative Sum Charts and Charting for Quality Improvement*. Springer Science & Business Media, 1998.

D. M. Hawkins, P. Qiu, and C. W. Kang. The changepoint model for statistical process control. *Journal of Quality Technology*, 35(4):355–366, 2003.

H. Hotelling. Multivariate quality control. *Techniques of statistical analysis*, 1947.

I. Hwang, S. Kim, Y. Kim, and C. E. Seah. A survey of fault detection, isolation, and reconfiguration methods. *IEEE transactions on control systems technology*, 18(3):636–653, 2010.

M. Iturbe, I. Garitano, U. Zurutuza, and R. Uribeetxeberria. Towards large-scale, heterogeneous anomaly detection systems in industrial networks: A survey of current trends. *Security and Communication Networks*, 2017, 2017.

D. A. Jackson. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214, 1993.

J. E. Jackson. *A User's Guide to Principal Components*, volume 587. John Wiley & Sons, 2005.

N. A. James and D. S. Matteson. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *arXiv preprint arXiv:1309.3295*, 2013.

I. Jolliffe. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

Z. Li, J. Zhang, and Z. Wang. Self-starting control chart for simultaneously monitoring process mean and variance. *International Journal of Production Research*, 48(15):4537–4553, 2010.

E. M. Maboudou-Tchao and D. M. Hawkins. Self-starting multivariate control charts for location and scale. *Journal of Quality Technology*, 43(2):113–126, 2011.

R. L. Mason and J. C. Young. *Multivariate Statistical Process Control with Industrial Applications*, volume 9. Siam, 2002.

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

K. G Mehrotra, C. K. Mohan, and H. Huang. *Anomaly Detection Principles and Algorithms*. Springer, 2017.

N. Meskin and K. Khorasani. *Fault Detection and Isolation: Multi-Vehicle Unmanned Systems*. Springer Science & Business Media, 2011.

Y. Mirsky, A. Shabtai, B. Shapira, Y. Elovici, and L. Rokach. Anomaly detection for smartphone data streams. *Pervasive and Mobile Computing*, 35:83–107, 2017.

D.C. Montgomery. *Introduction to Statistical Quality Control*. Wiley, sixth edition, 2009.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015. URL `http://www.rstudio.com/`.

G. C. Runger and M. C. Testik. Control charts for monitoring fault signatures: Cuscore versus GLR. *Quality and Reliability Engineering International*, 19(4):387–396, 2003.

W. A. Shewhart. The application of statistics as an aid in maintaining quality of a manufactured product. *Journal of the American Statistical Association*, 20(152):546–548, 1925.

D. Shi and F. Tsung. Modelling and diagnosis of feedback-controlled processes using dynamic PCA and neural networks. *International Journal of Production Research*, 41(2):365–379, 2003.

A. Tartakovsky, I. Nikiforov, and M. Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman and Hall/CRC, 2014.

N. Temme. *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. John Wiley & Sons, 2011.

F. Tsung. Statistical monitoring and diagnosis of automatic controlled processes using dynamic PCA. *International Journal of Production Research*, 38(3):625–637, 2000.

F. Tsung and D. W. Apley. The dynamic T^2 chart for monitoring feedback-controlled processes. *Iie Transactions*, 34(12):1043–1053, 2002.

F. Tsung and K. L. Tsui. A mean-shift pattern study on integration of SPC and APC for process monitoring. *IIE transactions*, 35(3):231–242, 2003.

O. van Dalen. Statistical monitoring of wind turbines. Bachelor's thesis, Eindhoven University of Technology, 2018.

M. van Stijn. Change detection in system parameters of lithography machines. Master's thesis, Eindhoven University of Technology, 2018.

J.I. van Zante-de Fokkert, A. Di Bucchianico, J. M. Wijnen, and J. Praagman. Run lengths of control charts for correlated output of feedback processes. *Report Eurandom*, 99020, 1999.

I. Žliobaitė, M. Pechenizkiy, and J. Gama. An overview of concept drift applications. In *Big data analysis: new algorithms for a new society*, pages 91–114. Springer, 2016.

Appendix A

Mathematical Results

A.1 The effect of taking the mean in a polynomial fit

Given n measurement points x_i and m_i measurements at each point $i = 1, 2, ..., n$ denoted by y_{ij}, where $1 \leq i \leq n$ and $1 \leq j \leq m_i$. We have two models

Normal model estimates $\hat{\beta}$ of a linear model $y = \beta_0 + \beta_1 \cdot x + ... + \beta_p \cdot x^p + \varepsilon$

Mean model estimates of $\hat{\beta}'$ of a linear model fitted on n measurements where each measurement is $\overline{y_i} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$, with linear model $y = \beta'_0 + \beta'_1 \cdot x + ... + \beta'_p \cdot x^p + \varepsilon$

We have two questions in this case

- Is there a difference in the estimates of both models?
- What is the difference in variance of the estimates?

We start with analyzing the normal model. We have n measurement points x_i and m_i measurements at each point $i = 1, 2, ..., n$ denoted by y_{ij}, where $1 \leq i \leq n$ and $1 \leq j \leq m_i$. Then the sum of squares is

$$\begin{aligned}
SS_{\text{Normal}} &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij} - (\beta_0 + \beta_1 \cdot x_i + ... + \beta_p \cdot x_i^p))^2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij} - (\sum_{\ell=1}^{p} \beta_\ell \cdot x_i^\ell))^2 \\
\frac{dSS_{\text{Normal}}}{d\beta_k} &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} -2x_i^k (y_{ij} - (\sum_{\ell=1}^{p} \beta_\ell \cdot x_i^\ell)) \\
&= \sum_{i=1}^{n} -2x_i^k (m_i \overline{y_i} - \sum_{j=1}^{m_i} (\sum_{\ell=1}^{p} \beta_\ell \cdot x_i^\ell)) \\
&= \sum_{i=1}^{n} -2x_i^k m_i (\overline{y_i} - (\sum_{\ell=1}^{p} \beta_\ell \cdot x_i^\ell))
\end{aligned}$$

Now for the mean model where $\overline{y_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$. Then the sum of squares is

$$SS_{\text{Mean model}} = \sum_{i=1}^{n}(\overline{y_i} - (\beta'_0 + \beta'_1 \cdot x_i + ... + \beta'_p \cdot x_i^p))^2 = \sum_{i=1}^{n}(\overline{y_i} - (\sum_{\ell=1}^{p} \beta'_\ell \cdot x_i^\ell))^2$$
$$\frac{dSS_{\text{Mean model}}}{d\beta'_k} = \sum_{i=1}^{n} -2x_i^k(\overline{y_i} - \sum_{\ell=1}^{p} \beta'_\ell x_i^\ell).$$

We have $m_i = m$ for all i is a sufficient condition for $\beta = \beta'$, It can be seen that if we want that $\beta = \beta'$ we need to have $m_i = m$ for all i.

This is because when setting the derivative equal to 0, an equal factor m can be divided out, while an unequal factor m_i cannot be divided out. The difference between the derivatives is a factor m_i for each measurement group. We could multiply the sum of square of each subgroup with m_i in the mean model to get the same estimate. This gives a third model the adjusted mean model with the following sum of squares

$$SS_{\text{Adjusted mean model}} = \sum_{i=1}^{n} m_i(\overline{y_i} - (\beta_0 + \beta_1 \cdot x_i + ... + \beta_p \cdot x_i^p))^2 = \sum_{i=1}^{n} m_i(\overline{y_i} - (\sum_{\ell=1}^{p} \beta_\ell \cdot x_i^\ell))^2$$
$$\frac{dSS_{\text{Adjusted mean model}}}{d\beta_k} = \sum_{i=1}^{n} -2m_i x_i^k(\overline{y_i} - \sum_{\ell=1}^{p} \beta_\ell x_i^\ell).$$

So we can adjust the mean model estimation by multiplying with the amount of measurements in each group to get the same estimates as the normal model.

The variance of a linear model is

$$\hat{\sigma}^2 = \frac{1}{n-p} SSE,$$

where SSE is the sum of squares error. We can calculate the SSE for each of the three situations

$$SSE_{\text{Normal}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - (\sum_{\ell=1}^{p} \hat{\beta}_\ell \cdot x_i^\ell))^2$$
$$SSE_{\text{Mean model}} = \sum_{i=1}^{n}(\overline{y_i} - (\sum_{\ell=1}^{p} \hat{\beta}'_\ell \cdot x_i^\ell))^2$$
$$SSE_{\text{Adjusted mean model}} = \sum_{i=1}^{n} m_i(\overline{y_i} - (\sum_{\ell=1}^{p} \hat{\beta}_\ell \cdot x_i^\ell))^2$$

This gives us

$$\hat{\sigma}^2_{\text{Normal}} = \frac{1}{\sum_{i=1}^{n} m_i - p}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - (\sum_{\ell=1}^{p} \hat{\beta}_\ell \cdot x_i^\ell))^2$$
$$\hat{\sigma}^2_{\text{Mean model}} = \frac{1}{n-p}\sum_{i=1}^{n}(\overline{y_i} - (\sum_{\ell=1}^{p} \hat{\beta}'_\ell \cdot x_i^\ell))^2$$
$$\hat{\sigma}^2_{\text{Adjusted mean model}} = \frac{1}{n-p}\sum_{i=1}^{n} m_i(\overline{y_i} - (\sum_{\ell=1}^{p} \hat{\beta}_\ell \cdot x_i^\ell))^2$$

It can be seen that the SSEs are not the same and also the variances are not the same.

A.2 Orthogonal polynomial fit

Given the coefficients of a linear model from some set X. So we have $\{\beta_0, \beta_1, \beta_2, \beta_3\}$ for the model $y = \beta_0 \cdot 1 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3 + \varepsilon$. We want to transform these coefficients to $\{\beta'_0, \beta'_1, \beta'_2, \beta'_3\}$ for the model $y = \beta'_0 \cdot p_0(x) + \beta'_1 \cdot p_1(x) + \beta'_2 \cdot p_2(x) + \beta'_3 \cdot p_3(x) + \varepsilon$. Here $p_0(x), p_1(x), p_2(x)$ and $p_3(x)$ are orthogonal polynomials.

To define a set up to third order orthogonal polynomials we need a set at which the orthogonal polynomials are designed. The grid of measuring is $X = \{-12.72;\ -10.60;\ -8.48;\ -6.36;\ -4.24;\ -2.12;\ 0.00;\ 2.12;\ 4.24;\ 6.36;\ 8.48; 10.60;\ 12.72\}$ in mm. Therefore we want to define the orthogonal polynomials on these coordinates. Note that the model that is used has unit nm, nm/cm , nm/cm^2, nm/cm^3, this means that the points are $X = \{-1.272;\ -1.060;\ -0.848;\ -0.636;\ -0.424;\ -0.212;\ 0.00;\ 0.212;\ 0.424;\ 0.636; 0.848;\ 1.060;\ 1.272\}$ in cm. Gram Schmidt orthogonalization will be used to calculate the set of orthogonal polynomials. We will first define the inner product by $(p, q) = \sum_{x \in X} p(x) \cdot q(x)$. First calculate the following inner products $(x, 1), (x^2, 1), (x^2, x), (x^3, 1), (x^3, x), (x^3, x^2), (1, 1), (x, x),\ (x^2, x^2)$. Note that whenever $n + m$ is odd we have that $(x^n, x^m) = 0$. Also $(x^2, 1) = (x, x),\ (x^3, x) = (x^2, x^2)$. This gives that

$$
\begin{aligned}
(x, 1) = (x^2, x) = (x^3, 1) = (x^3, x^2) &= 0 \\
(1, 1) &= \sum_{x \in X} 1 = 13 \\
(x^2, 1) &= \sum_{x \in X} x^2 = 2 \cdot (2.12^2 + 4.24^2 + 6.35^2 + 8.48^2 + 10.6^2 + 12.72^2) \\
&= 8.179808 \\
(x, x) &= 8.179808 \\
(x^3, x) &= \sum_{x \in X} x^4 = 2 \cdot (2.12^4 + 4.24^4 + 6.35^4 + 8.48^4 + 10.6^4 + 12.72^4) \\
&= 9.190832 \\
(x^2, x^2) &= 9.190832.
\end{aligned}
$$

The polynomials are than calculated by

$$p_i(x) = x^i - \sum_{j=0}^{i-1} \frac{(x^i, p_j(x))p_j(x)}{(p_j(x), p_j(x))}$$

and then

$$\tilde{p_i}(x) = \frac{p_i(x)}{||p_i(x)||}$$

It is also possible to calculate the orthogonal polynomials differently using

$$p_i(x) = x^i - \sum_{j=0}^{i-1} (x^i, \tilde{p_j}(x))\tilde{p_j}(x)$$

and

$$\tilde{p_i}(x) = \frac{p_i(x)}{||p_i(x)||}.$$

This is the same, because

$$\begin{aligned} p_i(x) &= x^i - \sum_{j=0}^{i-1} \left(x^i, \frac{p_j(x)}{||p_j(x)||} \right) \frac{p_j(x)}{||p_j(x)||} \\ &= x^i - \sum_{j=0}^{i-1} \frac{(x^i, p_j(x))p_j(x)}{||p_j(x)|| \cdot ||p_j(x)||} \\ &= x^i - \sum_{j=0}^{i-1} \frac{(x^i, p_j(x))p_j(x)}{(p_j(x), p_j(x))}. \end{aligned}$$

So this is the same as before where we first calculated the $p_i(x)$ and then normalized them.

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= x - \frac{(x, p_0(x))}{(p_0(x), p_0(x))} \cdot p_0(x) = x - \frac{(x,1)}{(1,1)} \cdot 1 = x \\ p_2(x) &= x^2 - \frac{(x^2, p_0(x))}{(p_0(x), p_0(x))} \cdot p_0(x) - \frac{(x^2, p_1(x))}{(p_1(x), p_1(x))} \cdot p_1(x) \\ &= x^2 - \frac{(x^2,1)}{(1,1)} \cdot 1 - \frac{(x^2,x)}{(x,x)} \cdot x = x^2 - \frac{8.179808}{13} \\ &= x^2 - 0.629216 \\ p_3(x) &= x^3 - \frac{(x^3, p_0(x))}{(p_0(x), p_0(x))} \cdot p_0(x) - \frac{(x^3, p_1(x))}{(p_1(x), p_1(x))} \cdot p_1(x) - \frac{(x^3, p_2(x))}{(p_2(x), p_2(x))} \cdot p_2(x) \\ &= x^3 - \frac{(x^3,1)}{(1,1)} \cdot 1 - \frac{(x^3,x)}{(x,x)} \cdot x - \frac{(x^3, x^2 - 0.629216)}{(x^2 - 0.629216, x^2 - 0.629216)} \cdot (x^2 - 0.629216) \\ &= x^3 - \frac{9.190832}{8.179808} \cdot x = x^3 - 1.1236 \cdot x. \end{aligned}$$

These polynomials can be normalized. Therefore we calculate

$$\begin{aligned} (p_0, p_0) &= (1,1) = 13 \\ (p_1, p_1) &= (x,x) = 8.179808 \\ (p_2, p_2) &= (x^2 - 0.629216, x^2 - 0.629216) = 4.043966 \\ (p_3, p_3) &= (x^3 - 1.1236 \cdot x, x^3 - 1.1236 \cdot x) = 1.869449 \\ ||p_0|| &= \sqrt{(p_0, p_1)} = 3.605551 \\ ||p_1|| &= \sqrt{(p_1, p_1)} = 2.860036 \\ ||p_2|| &= \sqrt{(p_2, p_2)} = 2.010962 \\ ||p_3|| &= \sqrt{(p_3, p_3)} = 1.367278 \end{aligned}$$

Dividing the polynomials with the normalization factor gives

$$\begin{aligned} \tilde{p_0}(x) &= 0.2773501 = \alpha_{00} \\ \tilde{p_1}(x) &= 0.3496459x = \alpha_{11}x \\ \tilde{p_2}(x) &= 0.4972746x^2 - 0.3128931 = \alpha_{22}x^2 + \alpha_{02} \\ \tilde{p_3}(x) &= 0.7313801x^3 - 0.8217787x = \alpha_{33}x^3 + \alpha_{13} \end{aligned}$$

We can check if we did it correct by calculating the inner products between the polynomials.

$$(\tilde{p_0}, \tilde{p_1}) = (\alpha_{00}, \alpha_{11}x) = \sum_{x \in X} \alpha_{00}\alpha_{11}x = \alpha_{00}\alpha_{11}(x, 1) = 0$$

$$(\tilde{p_0}, \tilde{p_2}) = (\alpha_{00}, \alpha_{22}x^2 + \alpha_{02}) = \sum_{x \in X} \alpha_{00}\alpha_{22}x^2 + \alpha_{00}\alpha_{02} = \alpha_{00}(\alpha_{22}(x^2, 1) + \alpha_{02}(1, 1))$$

$$= \alpha_{00}(0.4972746 \cdot 8.179808 - 0.3128931 \cdot 13) = 0$$

$$(\tilde{p_0}, \tilde{p_3}) = (\alpha_{00}, \alpha_{33}x^3 + \alpha_{13}) = \sum_{x \in X} \alpha_{00}\alpha_{33}x^3 + \alpha_{00}\alpha_{13} \cdot x = \alpha_{00}\alpha_{33}(x^3, 1) + \alpha_{00}\alpha_{13} \cdot (x, 1) = 0$$

$$(\tilde{p_1}, \tilde{p_2}) = (\alpha_{11}x, \alpha_{22}x^2 + \alpha_{02}) = \sum_{x \in X} \alpha_{11}\alpha_{22}x^3 + \alpha_{11}\alpha_{02}x = \alpha_{11}\alpha_{22}(x^3, 1) + \alpha_{11}\alpha_{02}(x, 1) = 0$$

$$(\tilde{p_1}, \tilde{p_3}) = (\alpha_{11}x, \alpha_{33}x^3 + \alpha_{13}) = \sum_{x \in X} \alpha_{11}\alpha_{33}x^4 + \alpha_{11}\alpha_{13}x^2 = \alpha_{11}(\alpha_{33}(x^4, 1) + \alpha_{13}(x^2, 1))$$

$$= \alpha_{11}(0.7313801 \cdot 9.190832 - 0.8217787 \cdot 8.179808) = 0$$

$$(\tilde{p_2}, \tilde{p_3}) = (\alpha_{22}x^2 + \alpha_{02}, \alpha_{33}x^3 + \alpha_{13}x) = \sum_{x \in X} \alpha_{22}\alpha_{33}x^5 + \alpha_{13}\alpha_{22}x^3 + \alpha_{02}\alpha_{33}x^3 + \alpha_{13}\alpha_{02}x$$

$$= \alpha_{22}\alpha_{33}(x^5, 1) + (\alpha_{13}\alpha_{22} + \alpha_{02}\alpha_{33})(x^3, 1) + \alpha_{13}\alpha_{02}(x, 1) = 0$$

Given the original coefficients from the linear model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ we can calculate a transformation to get the new model $y = \beta'_0 \cdot p_0(x) + \beta'_1 \cdot p_z(x) + \beta'_2 \cdot p_2(x) + \beta'_3 \cdot p_3(x)$. Filling in $p_0(x) = \alpha_{00}, p_1(x) = \alpha_{11}x, p_2(x) = \alpha_{02} + \alpha_{22}x^2 p_3(x) = \alpha_{13}x + \alpha_{33}x^3$ gives

$$\begin{aligned} y &= \beta'_0 \cdot \alpha_{00} + \beta'_1 \cdot \alpha_{11}x + \beta'_2 \cdot (\alpha_{02} + \alpha_{22}x^2) + \beta'_3(\cdot\alpha_{13}x + \alpha_{33}x^3) \\ &= (\beta'_0 \cdot \alpha_{00} + \beta'_2 \cdot \alpha_{02}) + (\beta'_1 \cdot \alpha_{11} + \beta'_3 \cdot \alpha_{13})x + \beta'_2 \cdot \alpha_{22}x^2 + \beta'_3 \cdot \alpha_{33}x^3. \end{aligned}$$

This gives that

$$\begin{aligned} \beta_0 &= \beta'_0 \cdot \alpha_{00} + \beta'_2 \cdot \alpha_{02} \\ \beta_1 &= \beta'_1 \cdot \alpha_{11} + \beta'_3 \cdot \alpha_{13} \\ \beta_2 &= \beta'_2 \cdot \alpha_{22} \\ \beta_3 &= \beta'_3 \cdot \alpha_{33} \end{aligned}$$

In matrix form this is

$$\beta = \begin{bmatrix} \alpha_{00} & 0 & \alpha_{02} & 0 \\ 0 & \alpha_{11} & 0 & \alpha_{13} \\ 0 & 0 & \alpha_{22} & 0 \\ 0 & 0 & 0 & \alpha_{33} \end{bmatrix} \cdot \beta'.$$

Calculating the inverse gives

$$\beta' = \begin{bmatrix} \frac{1}{\alpha_{00}} & 0 & -\frac{\alpha_{02}}{\alpha_{22}\alpha_{00}} & 0 \\ 0 & \frac{1}{\alpha_{11}} & 0 & -\frac{\alpha_{13}}{\alpha_{33}\alpha_{11}} \\ 0 & 0 & \frac{1}{\alpha_{22}} & 0 \\ 0 & 0 & 0 & \frac{1}{\alpha_{33}} \end{bmatrix} \cdot \beta.$$

Filling in the coefficients gives

$$\beta' = \begin{bmatrix} 3.605551 & 0 & 2.268671 & 0 \\ 0 & 2.860036 & 0 & 3.213537 \\ 0 & 0 & 2.010962 & 0 \\ 0 & 0 & 0 & 1.367278 \end{bmatrix} \cdot \beta.$$

Appendix B

Classes Results

B.1 Class 0

Machine	amount of PCs	Variance explained	Swap in data
18564	9	84.76	Yes
18586	6	89.13	Yes
18590	9	91.38	Yes
18646	9	89.74	No
18670	1	52.39	No
18740	7	91.71	Yes

Table B.1: Individual PCA fit for Class 0 machines

.

Method	Parameter	2 weeks before calibration (%)	1 week before calibration (%)	1 week after calibration (%)	3 days before swap (%)	3 days after swap (%)	Other signals (%)	Total signals
Shewhart	1	0.01	0.24	0.71	0.00	0.00	0.04	68
Shewhart	2	0.00	0.27	0.73	0.00	0.00	0.00	37
Shewhart	3	0.00	0.26	0.74	0.00	0.00	0.00	27
EWMA	1	0.06	0.12	0.36	0.00	0.00	0.46	219
EWMA	2	0.07	0.18	0.49	0.00	0.00	0.26	92
EWMA	3	0.06	0.20	0.60	0.00	0.00	0.14	65
CUSUM	1	0.02	0.10	0.82	0.00	0.00	0.06	49
CUSUM	2	0.07	0.19	0.47	0.00	0.00	0.27	120

Table B.2: Explained signals for class 0

B.2 Class 1

Machine	amount of PCs	Variance explained	Swap in data
18537	8	86.33	No
18541	8	89.59	No
18544	6	91.75	Yes
18548	6	85.69	Yes
18551	7	89.53	No
18556	7	87.41	Yes
18557	5	88.66	Yes
18563	6	90.91	Yes
18582	8	91.82	Yes
18599	9	77.20	No
18604	5	87.60	No
18642	9	80.49	No
18647	6	90.47	Yes
18648	6	88.66	Yes
18659	5	88.86	Yes
18678	4	89.64	Yes
18687	7	83.30	No
18706	8	79.09	No
18715	6	83.65	Yes
18721	9	84.60	No
18732	8	87.54	Yes
18741	7	83.05	Yes
18742	9	86.42	No
18744	8	86.22	Yes
18745	4	89.10	Yes
18756	7	85.64	No

Table B.3: Individual PCA fit for Class 1 machines

B.3 Class 3

Machine	amount of PCs	Variance explained	Swap in data
20210	6	81.34	Yes
20211	8	87.76	Yes
20212	6	85.20	Yes
20881	9	81.33	Yes

Table B.4: Individual PCA fit for Class 3 machines

Method	Parameter	2 weeks before calibration (%)	1 week before calibration (%)	1 week after calibration (%)	3 days before swap (%)	3 days after swap (%)	Other signals (%)	Total signals
Shewhart	1	0.04	0.39	0.53	0.00	0.00	0.05	57
Shewhart	2	0.08	0.54	0.38	0.00	0.00	0.00	13
Shewhart	3	0.00	1.00	0.00	0.00	0.00	0.00	3
EWMA	1	0.04	0.34	0.50	0.00	0.00	0.12	120
EWMA	2	0.03	0.26	0.59	0.00	0.00	0.12	58
EWMA	3	0.06	0.22	0.66	0.00	0.00	0.06	32
CUSUM	1	0.06	0.06	0.62	0.00	0.00	0.25	16
CUSUM	2	0.06	0.20	0.58	0.00	0.00	0.15	65

Table B.5: Explained signals for class 3

B.4 Class 4

Machine	amount of PCs	Variance explained	Swap in data
15395	8	84.06	Yes
15409	8	90.72	Yes
15412	3	85.87	Yes
15423	8	90.21	Yes
15434	9	91.16	Yes
15445	7	86.32	Yes

Table B.6: Individual PCA fit for Class 4 machines

Method	Parameter	2 weeks before calibration (%)	1 week before calibration (%)	1 week after calibration (%)	3 days before swap (%)	3 days after swap (%)	Other signals (%)	Total signals
Shewhart	1	0.00	0.11	0.25	0.00	0.00	0.65	57
Shewhart	2	0.00	0.08	0.16	0.00	0.00	0.76	25
Shewhart	3	0.00	0.00	0.33	0.00	0.00	0.67	3
EWMA	1	0.04	0.10	0.22	0.00	0.00	0.63	145
EWMA	2	0.00	0.01	0.34	0.00	0.00	0.64	76
EWMA	3	0.00	0.00	0.32	0.00	0.00	0.68	38
CUSUM	1	0.00	0.00	0.30	0.00	0.00	0.70	27
CUSUM	2	0.01	0.02	0.30	0.00	0.00	0.67	90

Table B.7: Explained signals for class 4

B.5 Class 5

Machine	amount of PCs	Variance explained	Swap in data
15402	6	85.10	Yes
15431	6	88.15	Yes
15440	6	87.25	No
15443	5	86.04	No

Table B.8: Individual PCA fit for Class 5 machines

Method	Parameter	2 weeks before calibration (%)	1 week before calibration (%)	1 week after calibration (%)	3 days before swap (%)	3 days after swap (%)	Other signals (%)	Total signals
Shewhart	1	0.10	0.15	0.10	0.00	0.00	0.65	20
Shewhart	2	0.00	0.29	0.00	0.00	0.00	0.71	7
Shewhart	3	0.00	0.33	0.00	0.00	0.00	0.67	3
EWMA	1	0.12	0.10	0.18	0.00	0.00	0.60	82
EWMA	2	0.03	0.10	0.16	0.00	0.00	0.71	31
EWMA	3	0.00	0.10	0.20	0.00	0.00	0.70	20
CUSUM	1	0.00	0.12	0.12	0.00	0.00	0.76	17
CUSUM	2	0.05	0.15	0.12	0.00	0.00	0.68	40

Table B.9: Explained signals for class 5

B.6 Class 6

Machine	amount of PCs	Variance explained	Swap in data
15670	7	91.63	Yes
15719	10	94.73	Yes
20837	6	88.39	Yes
28160	7	93.69	Yes

Table B.10: Individual PCA fit for Class 6 machines

Method	Parameter	2 weeks before calibration (%)	1 week before calibration (%)	1 week after calibration (%)	3 days before swap (%)	3 days after swap (%)	Other signals (%)	Total signals
Shewhart	1	0.00	0.26	0.05	0.00	0.00	0.70	43
Shewhart	2	0.00	0.35	0.00	0.00	0.00	0.65	23
Shewhart	3	0.00	0.60	0.00	0.00	0.00	0.40	10
EWMA	1	0.04	0.13	0.07	0.00	0.00	0.76	124
EWMA	2	0.02	0.19	0.12	0.00	0.00	0.67	52
EWMA	3	0.00	0.30	0.11	0.00	0.00	0.59	37
CUSUM	1	0.00	0.35	0.00	0.00	0.00	0.65	26
CUSUM	2	0.04	0.13	0.07	0.00	0.00	0.76	83

Table B.11: Explained signals for class 6