Czech Technical University
Faculty of Electrical Engineering
Department of Computer Science

MASTER'S THESIS

Out-of-control state detection for manufacturing processes at Škoda Auto

Bc. Nguyen Diem Huong

Supervisor: Ing. Macaš Martin, Ph.D

Study Programme: Open Informatics
Field of Study: Data Science

May 2022

CTU
CZECH TECHNICAL
UNIVERSITY
IN PRAGUE

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Nguyen Diem Huong** Personal ID number: **456150**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Open Informatics**

Specialisation: **Data Science**

II. Master’s thesis details

Master’s thesis title in English:

Out-of-control state detection for manufacturing processes at Skoda Auto

Master’s thesis title in Czech:

Detekce stavu mimo kontrolu pro výrobní proces ve Škoda Auto

Guidelines:

Manufacturing process can get to an out-of-control state because of assignable causes that act on the process. A crucial task of statistical process control is an early detection of such out-of-control states. The goals of the thesis are: 1) Perform a survey of statistical process control methods for a given specific case of manufacturing process (multivariate, violated normality, autocorrelation of variables, process non-stationarity) 2) Implement the selected methods of detection of out-of-control state and evaluate them quantitatively on the provided anonymized data from the real manufacturing process at Škoda Auto and on generated synthetic data. 3) Propose, implement, and perform a preliminary testing of machine learning methods. Identify the most crucial issues and challenges. 4) Implement a software for statistical process control, which can import measurement data from DFQ file (Q-DAS format), preprocess them, apply selected methods of assignable cause detection, and present the results visually

Bibliography / sources:

1. Montgomery, Douglas C. Introduction to statistical quality control. John Wiley & Sons, 2020.
2. Sukchotrat, Thuntee, Seoung Bum Kim, and Fugee Tsung. "One-class classification-based control charts for multivariate process monitoring." IIE transactions 42.2 (2009): 107-120
3. Eva, Jarošová, and Noskievičová Darja. Pokročilejší metody statistické regulace procesu. Grada Publishing as, 2015.

Name and workplace of master’s thesis supervisor:

Ing. Martin Macaš, Ph.D. Cognitive Neurosciences CIIRC

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **22.06.2021** Deadline for master's thesis submission: **20.05.2022**

Assignment valid until: **19.02.2023**

Ing. Martin Macaš, Ph.D. Supervisor’s signature	Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student’s signature

Acknowledgement

I am very much obliged to my colleagues who have worked with me on this project, namely Ing. Macaš Martin, Ph.D. for being my supervisor.

My deepest gratitude also belongs to everyone that has supported me throughout my studies.

Author statement

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, May 2022 ....

Abstrakt

Pokrok v technologii přišpěl v posledním desetiletí ke zvýšení komplexity výrobních procesů ve výrobním průmyslu. Tyto výrobní procesy pak generují data se složitější strukturou, což společně v kombinaci s produkcí velkého množství dat zapříčinilo pokles efektivity klasických metod regulace procesu. Metody strojového učení nejsou v tomto oboru příliš používané, ale právě tyto metody ukazují největší potenciál řešit problémy, se kterými se dnešní regulace procesu musí často zabývat. Mezi ně se řadí například vysoká vícerozměrnost, nelinearita, multimodalita a vzájemné korelace mezi proměnnými. Tato práce se věnuje výzkumu statistické regulace procesů a využití klasických metod oproti vybrané metodě strojového učení, one-class support vector machine. Výsledky v sérii experimentů ukázaly, že se v tomto komplexním prostředí metoda one-class support vector machine orientuje lépe než metody klasické a v drtivě většině případů získává lepší výsledky. Metody byly také implementovány do aplikace pro statistické řízení procesů, která importuje data z měření, předzpracuje je a aplikuje tyto metody detekce působení zvláštních příčin na proces výroby a výsledky detekce prezentuje uživateli.

Klíčová slova: detekce stavu mimo kontrolu, strojové učení, statistická regulace procesu

Abstract

The progress of technology in the last decade contributed to the growth of complexity among processes in the manufacturing industry. These processes then generate data with increasingly complex data structure, which, in combination with the data production of today's world, resulted in the dwindling effectivity of the classical statistical process control methods. Although underused, machine learning-based methods have the potential to handle intricate dynamic processes with multivariate, nonlinear, and multimodal data that are also mutually correlated. This thesis focuses on the research of statistical process control and the comparison of its classical methods with a machine learning method, one-class support vector machine. The results of a series of conducted experiments show that the machine learning method adapts better to the complex environment and outperforms the classical methods. Finally, the thesis discusses the implementation of statistical process control software, which imports measurement data from a DFQ file (Q-DAS format), preprocesses them, applies selected methods of assignable cause detection, and presents the results visually to the user.

Keywords: out-of-control state detection, machine learning, statistical process control

Contents

List of Figures

List of Tables

Acronyms

CTQ critical-to-quality. 8, 11–13

FNR false negative rate. 30, 34, 35, 44

FPR false positive rate. 30, 34, 35

LCL lower control limit. 19, 20, 31, 35, 51

LSL lower-specification limit. 8, 20, 51

OOC out-of-control. 1, 9, 11, 12, 19, 31, 47, 60

OSVM one-class support vector machine. 17, 18, 21, 22, 31, 34, 35, 37, 40, 44, 46, 47, 52, 53, 57, 59, 60

SPC statistical process control. 1–3, 9–18, 25, 34, 35, 37, 46, 47, 59, 60

SVM support vector machine. 21, 22, 34

TQM total quality management. 9

UCL upper control limit. 19, 20, 31, 35, 51

USL upper specification limit. 8, 20, 51

Chapter 1

Introduction

The out-of-control (OOC) state detection is one of the crucial parts of quality control in manufacturing processes. As the name suggests, it detects faults in processes that create products that do not conform to the company's quality standard. The OOC state detection is performed by employing statistical process control (SPC), an approach of quality management that is commonly used in industries that operate with measurable parameter characteristics.

In Fig. 1.1, we can see a flowchart depicting the operation of SPC from input to output. According to an individual sample plan, a product (or a part of a product) is taken to have its parameters measured. If these parameters do not conform to the company's quality standards, the cause of this nonconformity needs to be identified and thoroughly investigated. After the cause of the problem has been found, an appropriate action (called a **control action**) needs to be deployed to remove it.

These assignable causes are not an innate part of the process, but they will inevitably and maybe continuously appear at the beginning of every SPC application. They should always be identified and removed as soon as possible. The reason being that these causes create an unpredictable process that has high variability, further affecting the production by creating products that considerably vary piece by piece. If correctly removed, assignable causes appear less over time until they are completely removed.

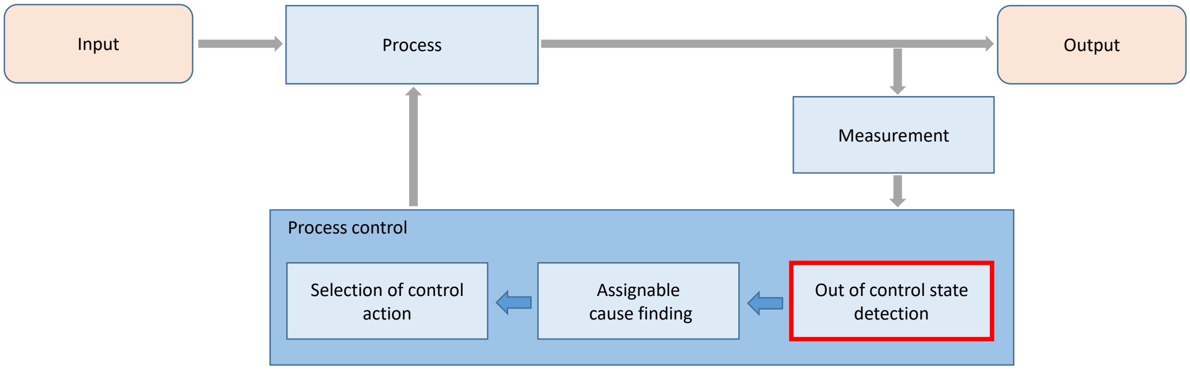


Figure 1.1: Statistical process control flowchart, adopted from [1]

Although the beginning of the industrial revolution marks the 18th century and processes suitable for SPC might be found even before that, the foundations of SPC were laid only almost a hundred years ago in the 1920s. The start of SPC is known to be pioneered by W. A. Shewhart and his Shewhart Control Chart, which popularized statistical methods in the manufacturing industry. Ever since then, SPC has been studied and extensively practiced, creating various methods to combat the rising complexity of both data and processes over time. Despite that, the intriguing systems of today's technology seem to create a very complex environment where the classical SPC methods fail to perform well.

The vast progress in technology in the manufacturing industry contributed to the considerable improvement of the quality of the machines used in production and the possibility of almost unlimited data collection. Both increased machine effectivity and the increased data pool lead to the rise of very complex dynamic processes. The data instances from such processes commonly possess properties that are intricate to work with. The classical SPC methods commonly underperform in these situations and often arrive at misleading conclusions. Examples of such data properties that are complicated for SPC to operate with are multimodality, nonnormality, heavy correlation, multivariety for a large number of variables, et cetera.

While there is a lot of research related to the typical univariate and multivariate SPC methods, the research focusing on today's data properties is still at an early stage. Although there are propositions for a new generation of machine learning-based SPC, many companies (if not most) still use the basic methods developed in the last century or settle for an empirical estimation based on experience. Such an approach ignores the underlying complexity of the present-day state detection task as many processes are interlinked in a complicated manner, creating complex patterns that are too difficult for a human or an older statistical process control tool to follow.

Although there is no formal summary on how to use machine learning in SPC correctly, some attempts were more than successful. To list a few examples from well-known companies, Intel uses predictive maintenance based on anomaly detection to predict IoT sensor breakdowns. This has reportedly saved them hundreds of million [2]. From the automotive industry, BMW uses artificial intelligence for quality assurance, boosting the efficiency of several production teams [3]. Machine learning is believed to have a significant potential in the process control field, although it is currently underused. This thesis focuses on the research of SPC in practice and encourages the use of machine learning in the field by evaluating and benchmarking chosen classic SPC methods against a selected machine learning algorithm.

Goal

This thesis is a part of a project with Škoda Auto, and as a result, it was developed after many discussions and teleconferences with various domain experts from the field of mechanical engineering. The goal included completing the following tasks:

- Conduct extensive theoretical research on SPC, commonly and historically used methods and approaches.
- Research machine learning in SPC.
- Do data analysis of the data set provided by Škoda Auto and, consequently, research problems related to the discoveries in data.
- Implement commonly used SPC methods and a selected machine learning algorithm. Evaluate and benchmark them on real data.
- Generate synthetic data from the real data, evaluate, and benchmark the methods on them.
- Implement an application that visualizes the results so that they can be used for analytical purposes even by non-experts.

Due to the fact that the data are real data from the manufacturing factory of Škoda Auto, this thesis, unfortunately, cannot reveal any internal information and domain expertise that was important for the creation of the output. Furthermore, the data itself and the code base of the application also cannot be revealed due to them containing confidential information. Visualization and graphs have been anonymized.

Structure

Since out-of-state detection and SPC is quite a specific topic to cover, the introduction provides the reader with the basic idea of SPC, out-of-state detection and touches on topics that will be discussed in more detail in the thesis. The following chapter Quality Management (Chap. 2) dives into important terminologies and definitions of management of quality, of which SPC is also part of.

Chapter Out-of-control State Detection (Chap. 3) is partly theoretical, partly practical. It contains related works, formal definitions of the classical SPC methods that are still commonly used, and introduces the selected machine learning algorithm One-class SVM. However, it also includes the analysis of the real data (only what can be revealed).

The rest of the thesis is purely practical. The whole chapter Experiments (Chap. 4) is devoted to the experiments conducted on synthetically generated data, which offers a better overview of the performance. The implementation of the application CIRQUE is described in the following chapter Application CIRQUE (Chap. 5), while also discussing the visualization theory behind the application design decisions. Ultimately, chapter Conclusion (Chap. 6) provides a complete overview of the research result and offers ideas of what can be done in the future.

Chapter 2

Quality management

Even before the rise of computers and automation in modern technology, ensuring the quality of products or services was crucial for most industries. From companies dealing with manufactured commodities to companies that provide commonly used services, the act of overseeing and managing activities to ensure a certain level of excellence is an essential concept for all of them. Examples of industries that make extensive use of quality management are, e.g., the car industry or textile industry, where the term quality could be defined varyingly. For cars, quality can be defined by parameters such as the diameter of body parts, depth of holes, and for a garment, it could be its durability or reliability. In services, these quality-improving methods can be applied to public transport or healthcare. Generally, quality management is relevant to most matters related to an organization or a company, including manufacturing, process development, finance, accounting, marketing, and even logistics. To genuinely improve, the processes of quality management undoubtedly require the efforts of many people within and even outside the organization to better the product and the general company culture.

Principles of quality management were already introduced as early as in the 19^{th} century. Frederick W. Taylor proposed principles in 1875 that he called “scientific management”, establishing dividing work into smaller assignments to simplify the manufacturing and assembling process [4]. As the mass production industries were developing at that time, his work managed to improve the productivity and the quality of manufactured goods considerably. Around this time, companies also started to impose the concept of work standards (standard time to finish the work, specified number of units per period, et cetera.), which was beneficial for the company’s productivity, but not very effective quality-wise [5].

In the 1920s, the rise of statistics and statistical theories used in businesses became apparent. Ford published *My Life and Work* focused on improving process efficiency, Fisher finished his papers on designed experiments, and finally, in the year 1924, W. A. Shewhart introduces the control chart concept [5]. This time in history is frequently dubbed as the formal beginning of statistical quality (process) control, and later on, Harold F. Dodge and Harry G. Romig add to it by developing statistical acceptance sampling. Around the year 1935, the statistical quality control methods were already extensively used at Western Electric, although these methods had not been exactly acknowledged by the public at that time [6].

The usage of statistical quality control was then greatly increased during World War II, where the importance of quality control was more than apparent. During this time, the manufacturing industries realized how important statistical methods were and started implementing them into their quality management. By the 1950s, these implemented statistical methods were applied widely in the chemical industry in the United States. However, the expansion of these methods into other industries in the US was relatively slow.

On the other side of the world, Japan encountered an after-war industrial-economic crisis. At that time, most Japanese netizens were considered predominantly illiterate, and the Japanese products were, more often than not, very low quality. After recognizing this issue, certain companies, such as Toyota, started heavily implementing quality management and quality control into their production processes. Japan produced better products at a lower price with a more efficient production system. By the 1960s, Japan became one of the most competent export countries, surpassing its Western competitors in rapid speed [7] [8].

Around the late 1970s, the discovery of Japan's improvement inspired the rest of the world to look into statistically designed experiments to reproduce Japan's success. The use of statistics in quality management has also considerably helped domestic businesses in the USA, which at that time suffered a substantial loss. The beginning of the extensive use of statistical methods for quality assurance was motivated by the fierce foreign competition and took a big part in the reemerging of the domestic industries.

The rest of this chapter is dedicated to the basic theoretical knowledge needed for the following chapters of the thesis. This chapter discusses several definitions of quality and what they actually portray in practice, introduces the quality engineering terminology frequently used throughout the subsequent parts of the thesis, and briefly presents other quality control methods besides statistical process control for overview. The last section is dedicated solely to statistical process control.

2.1 Defining quality

The whole concept of quality management comes from a business philosophy that lasting success can be obtained only by assuring the customer's satisfaction. A customer, an individual shopper, or even a corporate, then perceives quality as a factor that helps them make a choice. These choices are primarily based on the shopper's subjective values, such as which material is used, how appealing it is, or even the company's money-back guarantee. Conceptually, quality can be understood as the presence of wanted, desirable features that assist the consumer in final decisions.

In quality management theory, these subjective values can be put into more objective terms that adequately summarize the structure of the definition of this term. Quality has several differentiated components that are called **dimensions of quality** [9]. There are eight components that sufficiently represent quality in industries and businesses:

- **Performance** describes whether the product does its intended job. The customer typically compares specific functions of many different products to decide which one serves the best purpose, e.g., *When comparing visualization programs Power BI and Tableu for data analysis, Tableu's computing performance is better for a bigger volume of data.*
- **Reliability** can be understood as a probability of the product failing. Although many products can be repaired, a product is still called *unreliable* if it fails at its job too frequently (and before its effective service life ends). Healthcare appliances are typically heavily dependent on this particular dimension of quality. *A car that breaks down every month and needs to be repaired is unreliable.*
- **Durability** describes the effective service life of a product. It can be understood as how many uses until deprecation or until it is no longer economical to repair it anymore. The durability of a product is a desirable factor in the automobile industry and major appliances.
- **Serviceability** is associated with ease of repair – be it fixing a product or correcting a mistake in the service business. Serviceability involves the speed of repair, the competence of personnel, as well as the economic aspect of it. Every company handles its serviceability differently, which for customers typically directly reflects the quality of the product or service.
- **Aesthetics** responds to how visually appealing the product is. It involves the shape, color, sensory features such as tactile or smell characteristics of the product. *A popular color choice for fast cars is red due to its psychological subtext. Red is usually appealing for people who are more likely to drive fast.*
- **Features** describes how much more the customer gets, besides the basic functions that the product or service provides. Frequently, products and services that provide bonus features are considered higher quality than those who do not.
- **Perceived quality** corresponds to the company's reputation and the reputation of the product itself. Perceived quality is directly affected by the way the company handles failure scenarios. This dimension of quality is essential to customer loyalty.
- **Conformance to standards** is related to the likeness of the original design. A high-quality product should fulfill the conditions that the designer has set. Many manufactured products are created to be a piece with a much more intricate body. Not meeting these requirements can result in serious quality problems.

Montgomery [5] pointed out that although these eight dimensions are sufficient for businesses and industries, the service industry offers three more dimensions. He added another three dimensions to describe service and transactional business organizations:

- **Responsiveness** is how responsive the service providers were to the customer's request. It involves both the response time and the attitude.
- **Professionalism** responds to how qualified and competent the service provider is.
- **Attentiveness** is "how the customer feels about their concerns being addressed".

This multifaceted aspect of quality gives us many ways how to define it. There are two basic definitions of quality: the more traditional aspect and one that talks about variability dubbed as the “modern” definition.

Definition 1 Quality is fitness for use.

This is the traditional definition of quality, inspired by the attitude that products and services must simply fulfill the user’s demands. Every product and service has two aspects – the predetermined intentional grade/level of quality and how the product/service complies or *conforms* with the design specifications.

All cars are vehicles used for transportation. However, there are specific differences in sizes, performances, displays, etc., that are intentional design choices. Each car has a predetermined grade or level of quality that is intentional and predetermined by design. On the other hand, how well the product conforms to design specifications relies on the motivation of employees to achieve a good product, chosen procedures, etc. These aspects are technically called quality of design and quality of conformance.

Definition 2 Quality is inversely proportional to variability.

This is the "modern" definition of quality that implies that when variability decreases, the quality of the product increases. By variability at this time, we mean *undesirable* deflection from a target value.

The second definition is often called the modern definition of quality because of its all-in-one nature. Low variability means lower manufacturing costs and, later on, fewer repairs and rework. That translates into less wasted time, effort, and money. Such a definition of quality improvement directly relates to the following definition:

Definition 3 Quality improvement is the reduction of variability in processes and products.

The **variability** of the product or service comes in many forms. It may be the competence of the staff, equipment used in the making, and even the chosen materials of the design. Although it is impossible to make every product identically unit to unit, the companies try to minimize the variability to the point where it is economically doable. If the variability in the product is small, it does not impact the customer as much. However, if the variability between units is large, it impacts the user and is deemed unacceptable.

As we eliminate variability in processes and products, we also eliminate *waste*. The efficacy of this particular definition is shown in the service industry, where there is typically no parameter to measure. In such a situation, *waste* may equal to an error of an employee. Such a mistake can be costly as it takes time and effort to correct it.

2.2 Quality engineering terminology

The previous section contains a general, broad definition of the term quality. Every customer-consumed product also has a set of characteristics that the customer or the user thinks of as quality.

Such characteristics are called quality characteristics or also critical-to-quality (CTQ) characteristics. They collectively represent what "quality" is from the customer's point of view, and they can directly or indirectly relate to the dimensions of quality mentioned previously.

Quality characteristics have several different types that further specify them:

- **Physical** such as length, weight, voltage, or viscosity.
- **Sensory** such as taste, color, and appearance.
- **Time orientation** such as reliability, durability, and serviceability.

Quality engineering is a group of various activities used to retain the quality of a product in the company. Using these actions, the company tries to assure that CTQ characteristics of the product stay at a *required level* and the *variability* of the manufactured good is minimal.

Variability is a statistical term that can also be described only through statistical means. Statistical methods typically classify data on CTQ characteristics as **attributes** or **variables** based on the continuity of the data.

Data that can be classified as an attribute is discrete, and it could be some form of count as a number of emergency patients that had to wait more than 15 minutes before being treated. Variables, on the other hand, are continuous and typically measurements such as length.

Quality engineering activities often refer to certain specifications that must be met for the quality of the product to be high quality. Previously, it was mentioned that the company tries to assure that CTQ characteristics of the product stays at a *required level.* That required level in quality engineering terms is called the **nominal level** or also **target value**. It is the *desired value* of such characteristic of the product.

Because for some manufactured products, it is essentially impossible to stay on the nominal or target value at all times, these values are frequently bounded by a range. This range is usually decided upon by experts of respectable fields as they choose an acceptable variability that the company can allow while not compromising the quality of the product too much. The largest acceptable value for CTQ characteristics to have is called the **upper specification limit (USL)**. The smallest acceptable value for the characteristic to have is named the **lower-specification limit (LSL)**.

As was already mentioned, not simply the range but also the whole set of specifications are constructed by design engineers as a result of the engineering design process. The design configuration of the target value is firstly devised by following the engineering science principles. After that, a prototype is constructed and then tested. However, compared to what one may imagine, testing is generally not done in a statistically based experimental design manner. It is also done without the presence of the manufacturing processes that will be making the final product. The design engineers then set the rest of the specifications, releasing the product to manufacturing. This is called the **over-the-wall** approach.

The over-the-wall methodology does have certain setbacks. The general result from this approach yields many **nonconforming products**, meaning it creates products that do not satisfy at least one of the predetermined specifications. This type of problem, where the approach fails to create conforming products, is called a **nonconformity**. However, these products do not have to be automatically deemed inadequate and unfit for use [5].

A serum contains a nonconforming level of an active (effective) ingredient. Although the percentage of the active ingredient is below its lower specification limit, the user can still reap its effects, but it takes longer for the results to show on the skin.

A detected nonconformity in a product that is severe to the point of unacceptably affecting the quality of it is called a **defect**. If the product contains one or more defects, it is called **defective**.

It is, however, not uncommon for some of these characteristics to only have a single upper specification limit or only a lower-specification limit. A typical example of such a product is a bumper on a car. The bumper has a nominal value and lower specification limit, but it no longer possess the upper specification limit.

2.3 Quality control methodologies

The reader has so far encountered two critical terms that are commonly confused or mistakenly used interchangeably – **quality management** and **quality control**. The term *quality management* is an umbrella term for any approach used to control, manage and regulate quality within a company, while *quality control* is an approach of solely controlling the quality and is a part of quality management. Quality management could be anything from a business methodology for improving services, total regulation of every company department, or solely using statistically based approaches to find conforming products. The total regulation is, however, also a quality control methodology, as well as SPC, our primary tool for the OOC state detection of this thesis.

Before we turn our attention solely to SPC, let us briefly introduce some other possible methods of quality control to complete the reader's general overview.

Certain quality management approaches focus on improving their products or services through the whole company's management and operations. Total quality management (TQM) draws its base idea from the concept of a company-wide effort to collectively improve the quality of the final product. TQM means to assemble all departments to work on their management, operations, and processes, using other previously developed techniques of quality control [10]. TQM was slowly replaced due to the growing demand for zero-defect and on-time production. Methods such as **lean production** or **lean manufacturing**, which focus on efficiency of the production system and cutting down supplier-to-customer response time, were developed and utilized. They also apply forecasting strategies to predict supply-demand to request goods or materials only when needed, improving the amount of waste and cost [11].

At times, certain collections of standards or families of quality systems (or management systems) are used, such as in **ISO 9000** [12] or **Six Sigma** [13], well-known statistical tools for quality management. Some of them are regularly revised, updated, and used to this day.

Another approach is to use statistical methods to reach an improvement, called **statistical quality control**. It can be divided into three major parts: **designed experiments**, **acceptance sampling** and **statistical process control** [14].

Acceptance sampling relates to product testing in a way that we inspect a randomly chosen batch of products at any point in the process. Fig. 2.1 shows a system of acceptance sampling in a process. We call it an **incoming inspection** if we sample batches after receiving it from the supplier, **outgoing inspection** if we sample it before shipping it to the customer. **Rectifying inspection** is when a sampled lot gets rejected, but they are suitable for being reworked into something new.

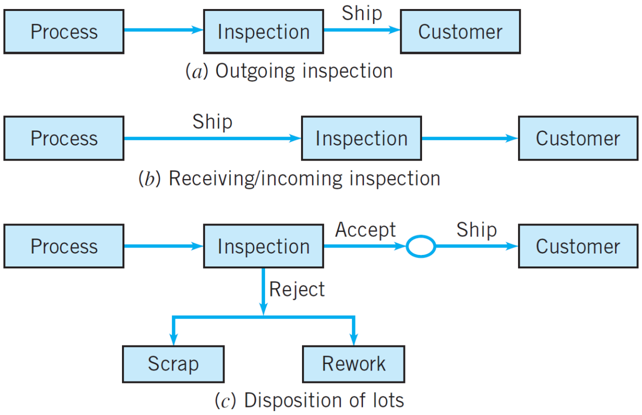


Figure 2.1: Acceptance sampling diagrams, adopted from [14]

Since acceptance sampling does not necessarily say much about the process and focuses more on the conformance-to-quality aspect of it, it is not used as much as SPC and design of experiments as the processes get increasingly complex. A well-designed experiment can be helpful as its goal is to determine key variables affecting the quality of the process. It does so by systematically shifting the variables and monitoring the change in variability. **Design of experiments** is mainly used at the development stage to decrease variability in the processes, signifying that it is an off-line quality control method. For the continuous monitoring of the process, methods of SPC are used.

2.4 Statistical process control

Statistical process control (SPC) is a method in quality control, which applies statistical methods to control and oversee processes in which specifications of conforming products can be quantified, typically by measuring. For that reason, SPC is regularly used in the manufacturing industries, although it can be administered in services or nonmanufacturing processes as well. One of the benefits of using SPC is in its core idea of correcting the process early instead of trying to correct a made mistake.

Fig. 2.2 is a more detailed depiction of Fig. 1.1 from Introduction (Chap. 1). Since the previous chapters already covered the necessary information, the flowchart now includes a better depiction of how SPC operates.

The machining process receives a process input that is cleared according to a predetermined control plan. A sample of CTQ characteristics are then chosen, and products containing the sample are transported to a measuring center for measurement. The OOC detection tool analyses the sample, and if a nonconforming product (or defect) is found, a diagnosis for an assignable cause is initiated. After identifying the assignable cause, a control action is selected and carried out to remove the cause.

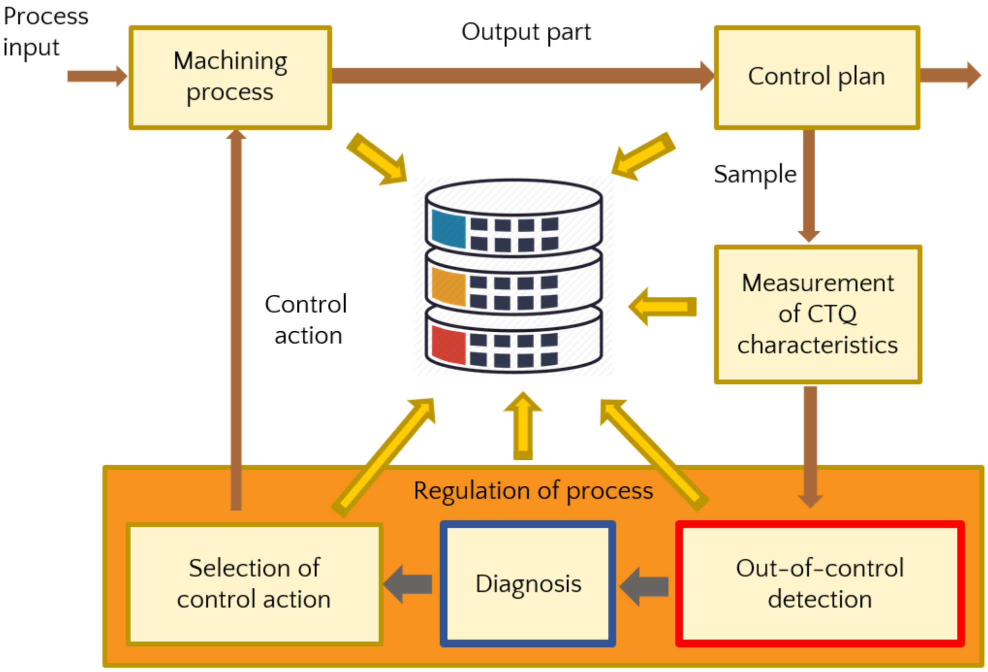


Figure 2.2: SPC diagram

It is clear that in reality, the SPC process is not as apparent and simple as is depicted in theory. Depending on the type of industry and each factory, many inevitable issues get in the way of accurate state detection.

In process control, we can identify two types of process inputs, **a controllable input** and **an uncontrollable input**. Fig. 2.3 shows a diagram of a manufacturing process. As the name suggests, the controllable inputs are process variables that we can control, such as applied pressure and feed rates. The uncontrollable inputs are mostly environmental factors that are either very difficult to control, or are uncontrollable, such as the behavior of the raw material. These inputs are all processed into a product possessing certain quality characteristics (CTQ characteristics) that are later on measured and evaluated.

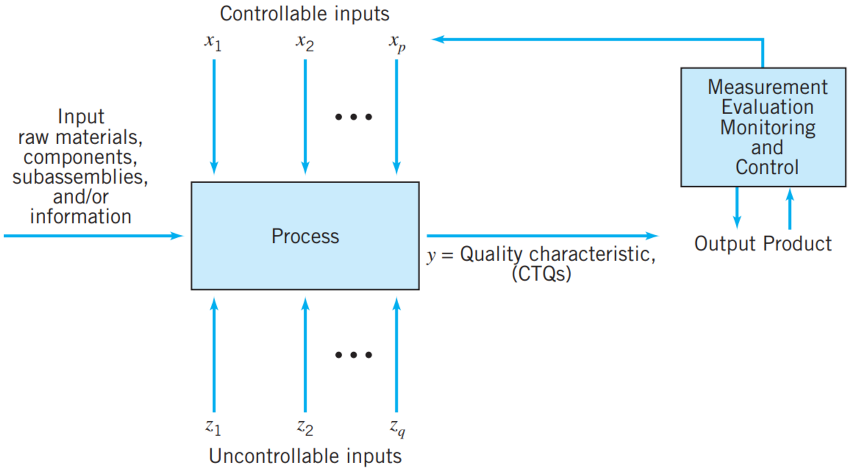


Figure 2.3: A manufacturing process, adopted from [14]

When selecting a method for an OOC detection, the manufacturing process as a whole should also be taken into great consideration, and it should be thoroughly discussed with the domain experts and employees. Looking again at the Fig. 2.2 of the SPC flowchart, some readers could easily assume that the measurement of CTQ characteristics would take a very short time, and no other components would be in the machining process at the moment, which is mostly not the case. For that reason, primarily in the beginning of the SPC application, the selected method needs to perform the detection quickly, in case the state was, in fact, in an OOC state. If the state were indeed in an OOC state, the fast detection would waste a smaller amount of components and materials compared to a method that would perform the detection slower.

2.4.1 Causes of variation

The previous chapters have already briefly mentioned the meaning of variation in statistical quality control. SPC differentiates between two sources of variation: a **common cause of variation** and a **special cause of variation**.

The common cause of variation is the natural variability in data that always occurs, no matter how stable and well-operated the process is also known as *background noise*. It is the result of small inevitable variations that have accumulated over time. A process consisting of only common causes of variations, or *chance cause of variation* is called to be **in-statistical-control**. A typical case of a common cause of variation is the humidity of the air.

The special cause of variation creates a significant shift in variance and should be minimized. Most of the time, the result of these mistakes produces a nonconforming product or even a defect. These causes are also called *assignable causes* of variation, and such a process consisting of such variation is called a process that is **out-of-statistical-control**.

Special causes of variation can be further differentiated into **sporadic** and **persistent** causes. Sporadic causes happen suddenly, affecting the process for a short period of time before disappearing.

These causes can reappear after some time if not taken care of. Persistent causes manifest themselves in changes in the observed parameters within the distribution of the observed variable (our CTQ characteristics) and typically affect the process until not managed. These causes can be caused by machine malfunctions or inexperienced staff.

Although most of the manufacturing processes that produce a conforming product usually operate with in-control processes, assignable causes will always happen, no matter how well-managed the process is due to natural occurrences (people make mistakes, devices break). Fig. 2.4 shows an in-control process with distribution μ_0 and σ_0. At time t_1, an assignable cause is present, shifting the distribution's mean to μ_1 for which holds $\mu_1 > \mu_0$. At time t_2, another special cause variation appears, which shifts the process distribution further.

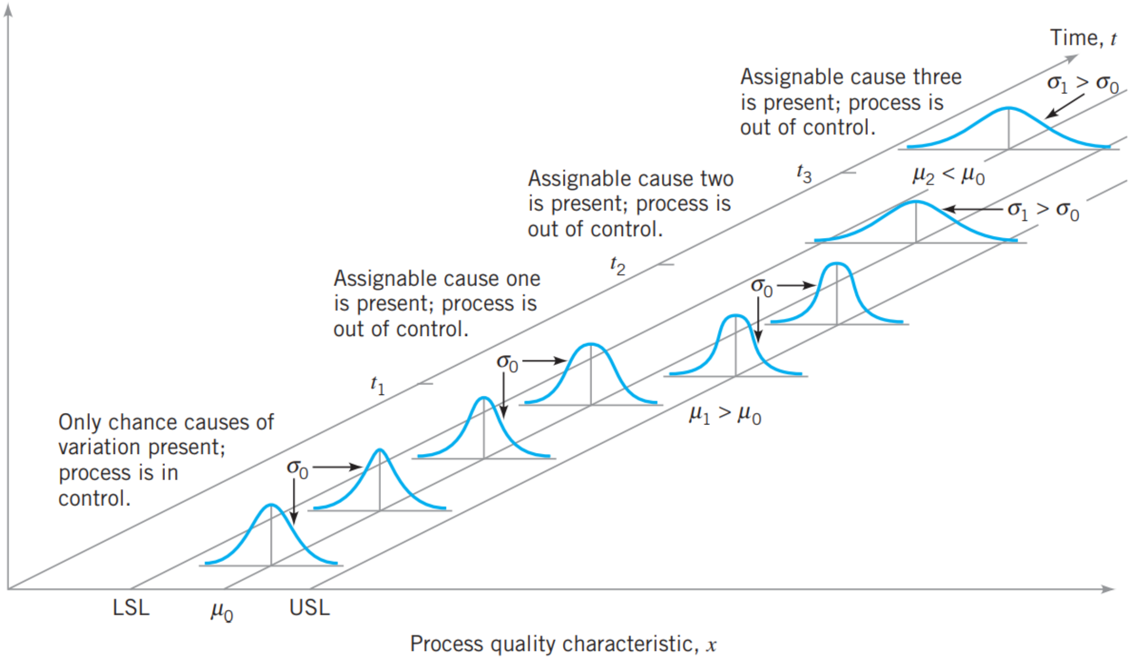

Time, t
Assignable cause three is present; process is out of control.
$\sigma_1 > \sigma_0$
t_3
$\mu_2 < \mu_0$
Assignable cause two is present; process is out of control.
$\sigma_1 > \sigma_0$
t_2
Assignable cause one is present; process is out of control.
σ_0
σ_0
t_1
$\mu_1 > \mu_0$
Only chance causes of variation present; process is in control.
σ_0
LSL
μ_0
USL
Process quality characteristic, x

Figure 2.4: Causes of variations, adopted from [14]

It is important to notice the lines denoted LSL (lower specification limit) and USL (upper specification limit), which were introduced in the chapter Quality Management (chap. 2). The graph indicates that if the process is in-control, the distribution primarily lies within the two limits. If the process is out-of-control, more of the distribution shifts outside the limits.

Having the process in-control also means that the *process is predictable* and we can foresee its behavior better. Out-of-control processes are unpredictable, unrepeatable and there is no way to predict its behavior.

As the goal of SPC is to minimize the variation between each manufactured product, the elimination of the assignable causes of variation is an important aspect. Unfortunately, it is not always possible to achieve a process without any assignable causes. In practice, the manufacturing process might be very complex and costly to adjust. In such cases, the company might decide that elimination of the assignable cause is too costly (and does not affect the quality of the product *too much*). [14]

2.4.2 Phases of SPC

Under normal circumstances, the process of establishing SPC in a company is divided into two phases, during which specific methods and actions are taken. The purpose of both of these phases is quite different, although it might not seem so at first glance.

Phase I is devoted to the retrospective analysis of historical data or, eventually, if they are not available, the collection of such data. Although the data collection task might seem trivial, considering that the manufacturing data is often confidential, the acquisition of the dataset can take up to a few months. During this period, it is necessary to learn as much as we can about the complex mechanical processes since the knowledge is necessary for the correct interpretation of the analysis.

A critical part of the analysis is determining whether the data were obtained from a process into statistical control. That is done by calculating the trial control limits (e. g. $\mu \pm 3\sigma$) from the data and constructing a control chart (see the following chapter 3 for detailed definition). The control chart displays the control limits computed from the data, the target value (given), and the data values. Data points plotted outside these control limits are considered nonconforming, and the process is recognized as out of statistical control. These nonconforming data points need to be then examined further to investigate anything suspicious of causing it. After the assignable cause is identified, the operating personnel work in conjunction with the mechanical engineers to eliminate the cause. Nonconforming data points (and possibly the product on which the data were measured) are then discarded from the dataset, and new control limits are calculated. Subsequently, new data is collected and plotted against the new control limits. The whole procedure is repeated until the process is stabilized. During this phase it is expected that certain circumstances in the process change due to the removal of assignable causes, action plan adjustments, et cetera. All of that is necessary to provide clean data for the second phase of SPC. Typically, at the beginning of any SPC application, it is assumed that the process is not in-control, and thus it is not uncommon to find in the literature that the objective of Phase I is to bring the process in statistical control.

Phase II is employed when the process has been reasonably stabilized. Although the process will always have some variability, the most harmful causes should have been removed by the corrective measures in Phase I. The objective of this phase is to monitor the process and remove smaller variabilities in data to achieve better quality. [15] [14].

Chapter 3

Out-of-control state detection

After a heavily theoretical chapter about the foundations of quality management and SPC, this chapter focuses solely on the state detection aspect of SPC. This chapter begins with a section Related work, where three generations of SPC are presented, and several important elements of SPC are discussed. Subsequently, the reader is finally introduced to the formal definition of the few most popular methods of SPC and a carefully selected machine learning algorithm. The rest of the chapter is dedicated to the data analysis of the real dataset from Škoda Auto and a discussion on it. This chapter also contains the evaluation method chosen for the following experiments.

3.1 Related work

This section is focused on the related work associated with SPC, introducing important methods used throughout history. It also discusses issues related to the growing complexity of data and processes, explains concerns of using classic methods of SPC on such data, and examines modern alternatives to standard statistical methods. It also discusses work related to variable selection, process changes, and the problematic side of interpretation and diagnosis of the SPC methods.

First generation SPC

Shewhart's control chart is generally considered a pioneering state detection tool. Although it was developed in the 1920s, making it a first-generation SPC tool, it is still currently used in process control. Despite its limitations in modern data due to its normality assumption and univariate character, we can still very often see variations of Shewhart used in practice, for example, in a 2017's paper where it was used to monitor clean ash during coal preparation [16], applying the central limit theorem despite not satisfying the normality assumption. Researchers also often discuss possible adjustments to Shewhart's control chart in order to improve its effectiveness in phase II of SPC by various control charting methodologies [17] or even fuzzy logic [18].

A control chart can also indicate that the process is out-of-control, although there is no data point plotted outside the control limits. Under such circumstances, the data point patterns exhibit certain non-random or systematic behavior. Despite the fact that there is no nonconforming data point to investigate, even such a situation can bring useful information. The symptomatic behavior of the pattern is a sign that something is wrong with the process, and nonconforming products may soon be produced. Trend analysis approaches the temporal patterns and evaluates the process from a dynamical point of view compared to the static approach of univariate methods. In 1956, Western Electric Company published a famous SPC handbook [19], which contains so-called Western Electric rules that distinguish control chart temporal patterns to natural and unnatural. Those rules are still in practical use, are implemented in SPC software, and were even extended. Between 1990 and 2010, the so-called "control chart pattern recognition" was intensively studied, and a related survey can be found in [20].

Second generation SPC

With the accomplishments of the first generation methods, the development of the second SPC generation was fueled by the growing ambition to reduce variability further and improve processes. As technology advanced, most systems evolved into a complex hierarchical structure of subsystems, creating a collection of varyingly intertwined datasets. As a result, correlations between quality characteristics slowly became a norm, exposing a detriment in univariate methods.

Around the 1980s, experts started pursuing the development of multivariate methods extensively. Process variables, however, violated many assumptions that were necessary for the application of multivariate methods, such as the absence of normal distribution, nonlinearity (failing to use linear transformation), multimodality (data comes from several distributions), the curse of dimensionality, and more [21]. Although today's processes call for extensive use of multivariate methods, many companies still utilize simpler univariate methods due to the previously mentioned issues. The second-generation SPC includes Hotelling T^2, multivariate CUSUM, MEWMA, U^2 multivariate control chart, and multivariate control charts based on projection methods such as Principal Component Analysis and Partial Least Squares Regression.

Third generation SPC

The third generation of SPC is trying to come up with a general approach to managing these previously mentioned problems. Undoubtedly, the amount of data will only increase in the future as automation in industries expands. Fortunately, with data, the field of data analysis and machine learning grows as well, allowing researchers to employ machine learning on SPC [22].

One of the methods that are frequently used to manage problems stated in the second generation SPC paragraph is Gaussian Mixture Model as in [23]. Intuitively, this unsupervised algorithm is an appropriate choice as it aims to find clusters in the data. Clusters represent the in-control state of the process, and SPC should normally contain a small number of out-of-control samples. However, researchers seem to forgo this method due to its dependency on the number of cluster selections.

State detection can also be transformed into a supervised learning problem in machine learning as in [24]. The authors generated synthetic nonconforming data, typically from the uniform distribution.

Using different labeling for the synthetic data and for the reference data, the method creates two different classes with two different labels, making it a two-class classification problem. In SPC, it is also called artificial contrast SPC.

Following up on the usage of clustering and classification on SPC, out of control state also offers an intuitive definition via one-class problem. In such case, we use data obtained solely from an in-control state to construct a model [25]. The convenience of this one-class setting is apparent when the out-of-control class might be poorly defined, or unavailable [26].

One of nowadays's most prominent methods based on one-class is one-class support vector machine (OSVM), also called support vector data descriptor. OSVM is typically linked with anomaly detection, which provides a very intuitive parallel to the state detection problem in SPC. It comes from the idea that nonconforming quality characteristics, or at least those suspected of nonconformity, usually do behave differently. The performance of OSVM is showcased in [27] where over 70 papers were examined. In 2003, Sun et al. [28] proposed a kernel distance-based control chart called k-chart. The shape of the irregular boundary formed by SVM is determined by the support vectors optimized via quadratic optimization. In 2018, [29] proposed a distance-based multivariate process control chart using support vector machines (SVM). They found that their model is more efficient than the random forest model [30] for high-dimensional and nonnormal processes. Other machine learning techniques used for SPC are neural networks [31], or nearest-neighbor algorithm [32].

Many of the machine learning methods are reported to be robust to nonnormality. A general survey on methods applied to process monitoring [33] contains a number of machine learning algorithms, mentioning that these methods show promise for monitoring large and diverse data sets.

Changing processess

All real-world processes are changing over time, and so do the statistical distribution of their data. Such changing processes can be evaluated using different approaches. A typical way of autocorrelated data handling is to apply an SPC method to the residuals of a time series model. Weese et al. [33] concluded their survey by stating that there is not "much literature on how to monitor auto-correlated data when the time series model is unknown". In [30], with each new observation, a new classifier is trained, and statistics (such as the error rate estimated from the classifiers) are monitored. The dynamic series of classifiers generate the statistics for the monitoring procedure. In machine learning, this phenomenon is also called concept drift, and a survey on possible solutions can be found in [34]. In the last four decades, adaptive control charts were researched and developed, in which at least one parameter (sample size, sampling interval, or control limit coefficient) changes in real-time according to the actual state of the sample statistics (see [35] for the survey).

Variable extraction

The usage of machine learning methods is closely connected to variable extraction, as is pointed out in another review from 2017 [36]. Variable extraction has a very important task – it needs to explore and identify representative features in the pursuance of avoiding irrelevant information and as a concequence, improve the classifier's performance. Variable selection tries to identify a subset of variables that carries most of the relevant information contained in the complete dataset. Integration of variable selection methods to MSPC approaches has become a promising research topic [37].

Interpretation and diagnosis

A quite important part of the research that is often neglected is the interpretation of positive detection (finding which characteristic is in charge of the alarm) and further diagnosis (finding the assignable cause). In common multivariate control charts based on Principal Component Analysis or Partial Least Squares [38], after an out-of-control (OOC) alarm is triggered, the projected point can be decomposed into its original variables, which can be then analyzed using contribution plots to determine which variables are responsible for the alarm. This is clearly more difficult for nonlinear data. In other multivariate charts, a sensitivity-based method can be used, where at each iteration, the influence of one characteristic on the final detection is removed, e.g., by replacement of its value by the mean computed over all available in-control data and the new value of the detection score or statistic is computed. The difference between this value and the value before characteristic influence removal approximately describes the importance of the characteristic for the decision. However, this method is approximate, and the sensitivity value depends monotonically on the number of removed characteristics, so it must be given how many characteristics we want to identify.

3.2 Methods used in SPC

The section about related research mentions many algorithms and techniques used for SPC. In reality, not all of them are commonly used because of their theoretical complexity or difficulty to use in practice. For that reason, the first generation's Shewhart control chart still remains one of the most popular methods due to its simplicity and easiness of interpretation. Another method worth mentioning would be the Hotelling control chart, which is the first multivariate SPC method. These two methods are frequently used in process control combined with the expertise of the operators and the mechanical staff members.

The most popular method among machine learning algorithms would be one-class support vector machine (OSVM). The nature of the algorithm offers a very intuitive definition of the problem and is suitable for both multivariate and multimodal data. We could easily imagine a parallel between machine learning and SPC as using OSVM in the retrospective phase I reminds the learning phase, and phase II is the use of the model. However, instead of a control chart, we are using a classification model of OSVM.

This section is be centered around the univariate Shewhart control chart, the multivariate Hotelling T^2 control chart, and a machine learning method, OSVM.

Shewhart control chart for individual measurements

Univariate methods analyze and monitor a single variable at a time, with the presumption that its mean and standard deviation come from the process that was in-control. As the process gets to the state of out-of-control, the values generated from such processes are suspected of being also out-of-control.

Although the popularity of univariate methods is still very apparent even today due to their simplicity and usefulness, the method of inspecting one variable at a time has its limitations. As technology advanced, process control began to involve related variables, creating the need to regard both variables simultaneously. Applying univariate methods to each variable separably can lead to a multiple comparison problem that manifests itself in the surge of false alarms. However, the presence of certain situations allow for the application of univariate methods to multivariate data, such as when multiple measurements are taken on the same unit of product or the availability of data is very slow, creating long intervals between measurement (as is in this case).

The complications that can arise from using univariate methods on correlated multivariate data are shown in Fig 3.1. The univariate method creates rectangular-shaped boundaries that can fail to recognize OOC data points and mistakenly categorize in-control datapoints as OOC.

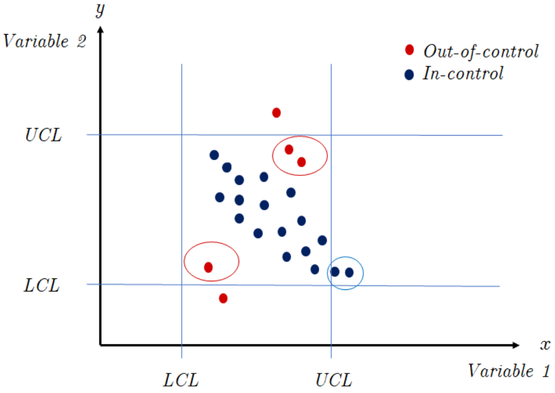


Figure 3.1: Univariate evaluation (Shewhart) on multivariate 2D data. Shewhart considers points inside the rectangular area as in-control.

The Shewhart control chart is a univariate method that is able to detect significant shifts in variability readily. Its x-axis represents time, while the y-axis depicts information we want to visualize: a specific statistic, e.g., mean or measured values of characteristics. The x-axis also suggests information about the order of the data sequence. The Shewhart control chart also follows the presumption that the historical observations were acquired from an in-control process.

Shewhart control chart contains a center line $\bar{x}$ bounded by an upper control limit (UCL) and a lower control limit (LCL):

$$UCL = \bar{x} + 3\frac{\bar{M}R}{d_2}$$
$$LCL = \bar{x} - 3\frac{\bar{M}R}{d_2}.$$

The center line $\bar{x}$ corresponds to the aspiring value of the observed characteristic. This value can be defined in several ways:

- The value is already defined by the regulations – it is the target/nominal value.
- We do not possess the target value. The center line is obtained by collecting the data while the process is statistically in control.
- The value is decided empirically on the basis of the past experience.

The limits UCL and LCL (control limits) are sometimes called *action limits* and are different from the limits mentioned above USL and LSL (specification limits). As we have already mentioned in the chapter 2, we have introduced the specification limits as in *most acceptable limits*. Note that they are *not the most acceptable limits in a way that they would create a defect if the value were exceeded*. A **specification limit** is an internal range agreed on by the company experts that does not *necessarily* call for action if exceeded. However, **control limits** says something about the real variation of the product. In such a case, if they are exceeded, the product is very likely defective and needs assistance. These control limits are selected to ensure that if the process is in control, all values are within limits.

Shewhart's control chart uses a moving range of two consecutive observations as the basis for estimating the variability of a process, defined in the following way.

$$MR_i = |x_i - x_{i-1}|.$$

Although it was developed over 100 years ago, the Shewhart control chart is still widely used in the industry due to its straightforwardness and ease of use [14].

Hotelling T^2 control chart

Hotelling T^2 control chart is one of the first multivariate methods developed. It extends the work of the univariate Shewhart $\bar{x}$ chart (chart depicting mean μ of the data), fitted for multivariate process control.

Let $\bar{\mathbf{x}}$ be the sample mean vector. The Hotelling T^2 statistic is defined as follows:

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-\mathbf{1}}(\mathbf{x} - \bar{\mathbf{x}})$$

where $\mathbf{S}$ is the estimation of covariance matrix

$$\mathbf{S} = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})'$$

where m is the number of observations, p is the number of parameters and $\mathbf{x_i}$ is the vector of quality characteristics with p-coordinates [39].

According to [40], the phase I limits should be based on the beta distribution; therefore, the limits should be defined as

$$UCL = \frac{(m-1)^2}{m}\beta_{\alpha,p/2,(m-p-1)/2}$$
$$LCL = 0$$

Hotelling T^2 control chart is a Shewhart-type chart that is also used during the statistical process control in phase I. The reason is that it detects more significant shifts, ignoring smaller shifts that are later dealt with in later phases of process control [14]. The boundary of the Hotelling T^2 control chart is elliptical as depicted in Fig 3.2.

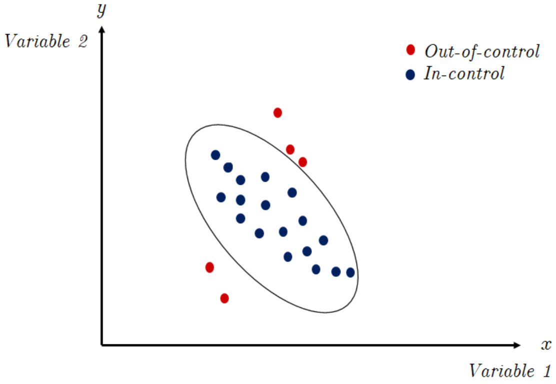


Figure 3.2: Evaluation (Hotelling) on multivariate 2D data. Hotelling considers the points inside the elliptical area as in-control.

One-class support vector machine

One-class support vector machine (OSVM) is an algorithm based on support vector machine (SVM) and a type of classification used primarily for anomaly detection, one-class classification. Before OSVM can be explained, it is necessary to discuss SVM and one-class approach.

Support vector machine (SVM)

SVM is a well-known machine learning algorithm that works based on supervised learning. It builds separating hyperplanes to split classes in the multidimensional space. The hyperplanes that it constructs are optimal, created from support vectors signifying the most influential data points in the set. SVM can be quite robust and effective as it performs well in the correct setting [41].

SVM can behave as a linear classifier for a binary classification problem or use a kernel trick to solve problems in n-dimensions. It can be used as soft SVM and hard SVM with the difference in how many wrongfully classified points are allowed. Due to this, SVM performs well when classifying both linearly separable and nonseparable datasets.

The optimal hyperplane is reached by solving a quadratic optimization problem, which maximizes the distance between two classes called margin and minimizes the training error. The quadratic optimization problem is solved by constructing a dual problem that minimizes Langragian L over $\mathbf{w}$ and b [42].

Hard margin SVM Hard margin SVM is used in the case of linearly separable data. The quadratic optimization problem in the following way [43]:

$$\min \tfrac{1}{2}||w||^2,$$
$$\text{s.t.: } y_i(\mathbf{w} \cdot \mathbf{x_i} \geq 1), \text{ for i} = 1,...,\text{N}$$

Soft margin SVM Soft-margin SVM is employed for datasets that are linearly nonseparable. It uses a slack variable $(\xi = \xi_1, ..., \xi_N)$ to allow missclassifications, since the dataset cannot be perfectly separated. It also introduces a constant C, which regulates the margin-training error tradeoff fixed by the user. We can define its quadratic optimization problem in the following way [43] [42]:

$$min\tfrac{1}{2}||\mathbf{w}||^2 + C\textstyle\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.: } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \text{ for } i = 1, ..., N$$
$$\xi_i \geq 0 \text{ for } i = 1, ..., N$$

Kernel trick Nonlinear classification with SVM was initially proposed in [44] using the kernel trick on the original hyperplane-maximizing SVM. In this case, every dot product is replaced with a nonlinear kernel function that fits the hyperplane into a transformed feature space.

Let us have a transformation $\Phi : x \rightarrow \varphi(x)$ and let us map every datapoint into high-dimensional space via this transformation Φ. Then the inner product between vectors $x_i \cdot x_j$ is:

$$K(\mathbf{x_i}, \mathbf{x_j}) = \varphi(\mathbf{x_i}) \cdot \varphi(\mathbf{x_j}),$$

where $K(\mathbf{x_i}, \mathbf{x_j})$ is called a kernel function .

One-class SVM As SVM is a supervised learning algorithm, it needs to learn to properly predict the correct labels by learning from a training dataset containing both. However, the one-class approach uses only one of the classes in the training approach, classifying everything else as an "anomaly" that belongs to another class [25]. The advantage of it is that it is trainable through a "clean" dataset without exploiting the negative sample from the other class. Despite that, in some cases, data points from other classes are utilized in a certain way to improve the classifier's performance [45]. One-class is therefore popularly used in anomaly detection.

The one-class approach can be used on SVM as well, creating a one-class support vector machine OSVM. OSVM builds a minimal hypersphere with radius r and center c containing all data points [46] [47]. The in-control data is then confined in the boundary created by the hypersphere.

OSVM is formally defined in the following manner:

$$\min_{r,c,\xi} r^2 + \tfrac{1}{\nu n}\textstyle\sum_{n}^{i=1}\xi_i,$$
$$s.t. : ||\phi(x_i) - c||^2 \leq r^2 + \xi_i, \forall i = 1, 2, ...n,$$

where ν is the positive parameter that indicates the tradeoff between sphere volume and the number of outliers, r is the radius of the hypersphere with center c [48].

3.3 Evaluation of methods

All SPC methods are typically experimentally analyzed on data in terms of well-known performance measures such as average length run, false positive rate, and false negative rate. There is, however, an issue with the availability of datasets. Although the quality and representativeness of such benchmark data are essential for the result's credibility, surprisingly little attention is paid to it. According to [27], 55% of the authors of the analyzed articles have used only simulated data in their research and 45% have used real (or real and simulated) data. The datasets mostly contain only units or tens of variables and a limited sample size. Moreover, no survey would summarize such benchmark datasets. For example, this problem is briefly mentioned in [37] and [49].

In our case, the evaluation of methods on real data also possesses an issue. Although we always evaluate data on the historical dataset provided to us, there is no way to obtain the ground truth. The ground truth in SPC should be the information about the presence of assignable causes. However, this is most likely not known. The decision of whether the characteristic is conforming or not is determined by the company's own process monitoring technique (which unfortunately cannot be disclosed). Although there have been various discussions about this problem, the procedure of acquiring the ground truth would have been too economically taxing to be realized. At first, we had no choice but to calculate our evaluation based on the Škoda method – to be more specific, the Škoda method was used as the ground truth for now, and we calculated our FNR rate according to it. Later on, we also decided to generate synthetic data with known ground truth to better grasp the performance of the methods.

Ultimately, we decided to use a false positive rate (FPR) and false negative rate (FNR) defined in the following way:

$$FNR = \frac{FN}{P}, \tag{3.1}$$

$$FPR = \frac{FP}{N}, \tag{3.2}$$

where FN is the number of out-of-control parts falsely classified as in-control, FP is the number of in-control parts falsely classified as out-of-control, P is the number of parts produced by the out-of-control process, and N is the number of parts produced by an in-control process.

We chose this evaluation method based on discussions with our colleagues from Škoda Auto on the importance of false positive and false negative rates. FPR is also analogous to the type I error rate, while FNR represents the type II error rate.

3.4 Analysis of real data

The dataset was provided by Škoda Auto, containing data about various parts of the engine heads currently produced. Unfortunately, due to security reasons, a lot of information about the data and domain knowledge has to stay unrevealed. For that reason, the figures and tables are anonymized, and some pieces of information will only be briefly stated – not followed by a graph or quantified proof of any kind. This section still hopes to provide as much information about data characteristics, so that the subsequent parts of the thesis will be as straightforward as possible.

In the beginning, we obtained a collection of measured values of various engine head parameters that came from a currently used monitoring system. This dataset was accompanied by metadata such as the date and time of measurement, certain physical features of the parameters, and other important engineering information. The data were clearly in the raw state and were presumed to be utilized by Škoda engineers (or mechanical operators) as they were not adequately described.

Despite that, after numerous teleconferences with our colleagues from Škoda, we were able to understand the meaning of the attributes to the point of being able to work with them. The first goal was to get a better overview of the dataset, hence, fundamental data analysis was performed to determine the statistical characteristics of the dataset better. Based on the discussion with Škoda employees, we have separated the data into specific groups based on the machine layout. These groups are identified by component labels, manufacturing machines, and tools of these machines. This separation results from the difference between the measured values of these groups (every component has several machines that operate on it, and every machine has several tools). Ultimately, the dataset contains hundreds of characteristics, creating a high-dimensional dataset. However, the data is fragmented due to the physical aspects of the component parameters. For that reason, the dataset suffers from the lack of certain types of data and the curse of dimensionality at the same time.

An analysis of the measurement sequence using the component characteristic's measurement time revealed that the measurements were irregular. The data also contains data point anomalies and unexpected absence of measurements. The reason for these occurrences was later explained, but it was revealed that this behavior could not be avoided. For that reason, we have decided not to conduct any time-series analysis due to the possibly misleading results. A plot of data measurements in time is shown in figure 3.3 with an anonymized set of characteristics. We can see from these four subfigures that the data already exhibit several types of trend (upward, downward, and normal). Although we cannot show the trend analysis for the whole dataset, this data behavior is only a fraction of what is occurring.

It is quite clear that we are working with very complicated processes, resulting in many common causes that may considerably change the data distribution. However, these causes are common because it is not economically profitable to manage them or it is simply impossible to do so.

After examining the dataset trends, it was revealed that the dataset actually exhibits many different trends, shifts, and other patterns, indicating that we are dealing with *multimodal processes*. These processes generate data from various distribution sources, making the behavior of the data unpredictable. Due to the difficulty of working with multimodal processes, the research on them has not been fully explored in SPC yet. Although machine learning-based methods show potential, no general SPC technique has been shown to be effective in practice, primarily due to the complexity of the technique in real-time monitoring. However, the fact that complex processes in Škoda generate multimodal processes is not surprising, since the automotive industry is known to be quite advanced and its production processes are complex.

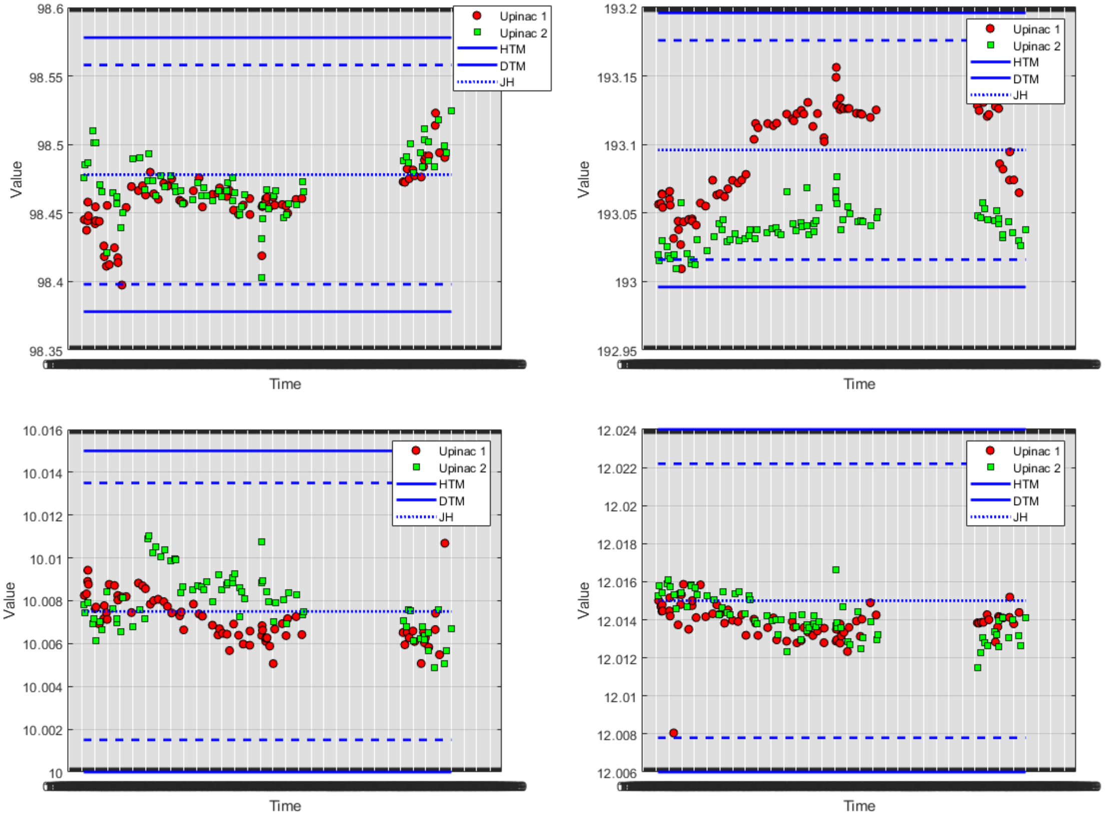


Figure 3.3: Examples of different types of temporal patterns inside the data

As a result of dealing with multimodal processes, we are working with data from multiple distribution sources, and thus the data cannot be assumed to be normal, violating the normality presumption of the multivariate methods. Figure 3.4 shows density functions of a group of randomly chosen anonymized characteristics (characteristics that possess the same component label, machine, and tool). A normal distribution graph was fitted to the component characteristic's histogram to better observe how the distribution compares to the normal distribution. The histogram is plotted as an x-axis containing the measured values and the number of occurrences of these values on the y-axis. The figure 3.4 shows that these component characteristics are not normally distributed.

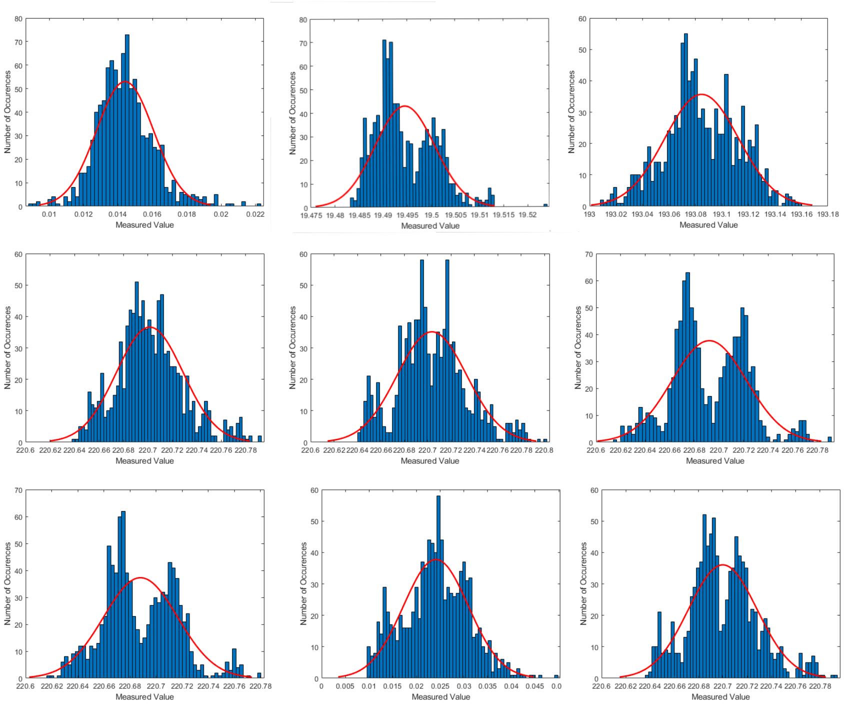


Figure 3.4: Distribution of component characteristics (each subplot corresponds to one characteristic)

The dataset also contains interesting cases of correlation depicted in the figure 3.5. The figure illustrates a situation where we have characters A and B that are almost linearly correlated. To get a better idea about their behavior, the left side of the figure also depicts measurements in time. The analysis of other characteristics discovered that several of them mimic each other's trends and behavior.

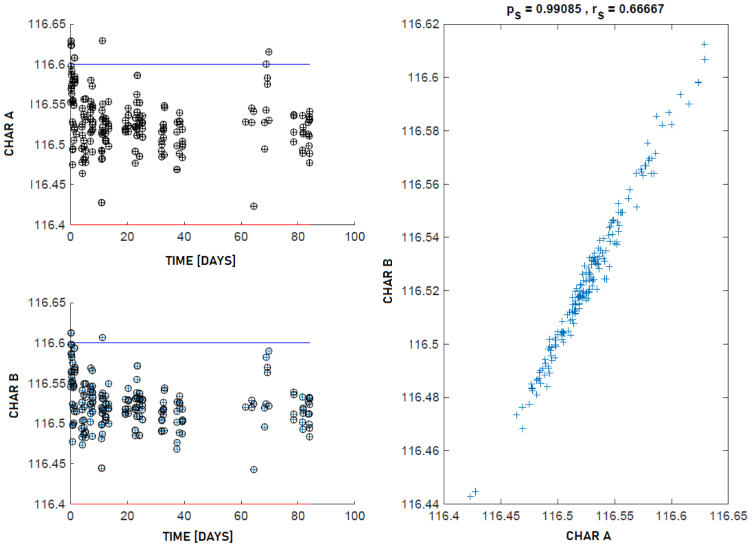


Figure 3.5: Correlation of two characteristics A and B

1	0.132	-0.0733	-0.0685	-0.0713	-0.0734	-0.0545	0.0358	-0.0723
0.132	1	0.3458	-0.0727	-0.0846	0.1779	0.1948	-0.5966	-0.0452
-0.0733	0.3458	1	0.7313	0.7174	0.8388	0.8555	-0.382	0.7522
-0.0685	-0.0727	0.7313	1	0.9757	0.9265	0.9372	0.0489	0.9834
-0.0713	-0.0846	0.7174	0.9757	1	0.9498	0.9436	0.0515	0.9954
-0.0734	0.1779	0.8388	0.9265	0.9498	1	0.9904	-0.1933	0.9605
-0.0545	0.1948	0.8555	0.9372	0.9436	0.9904	1	-0.205	0.9621
0.0358	-0.5966	-0.382	0.0489	0.0515	-0.1933	-0.205	1	0.0124
-0.0723	-0.0452	0.7522	0.9834	0.9954	0.9605	0.9621	0.0124	1

Table 3.1: Correlation matrix of characteristics

A correlation matrix 3.1 has been calculated on a selected anonymized group of 9 characteristics. The matrix has been color coded with a sequential color scheme to better observe the correlation between these characteristics. The higher the correlation between the characteristics, the darker the shade of red. Here, we can observe a very high linear correlation between several characteristics.

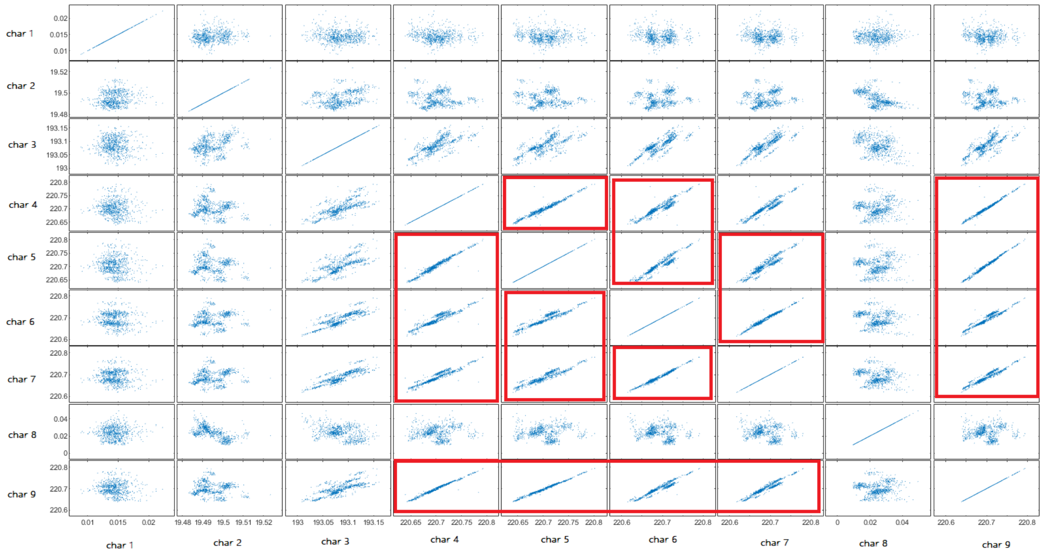


Figure 3.6: Scatter matrix of a group of component characteristics

A scatter matrix of the same group of characteristics in Fig. 3.6 marks the correlation greater than 0.9 with a red rectangle.

After collecting these correlated characteristics, we discovered that they were actually parameters coming from the same product part. An example can be a certain hole in the engine head – characteristic A may be a certain depth of the hole, and characteristic B could be another level of depth. In conclusion, these correlations were linear because physically they have to be, since characteristics A and B are parameters of one hole at different depths. Unfortunately, the appropriate identification labels for these characteristics were collected only after the correlation analysis, and this connection was not clear to us from the beginning. Also, it should be noted that there can be more than 2 highly correlated characteristics – the examples and figures of pairs of correlated characteristics are shown for simplicity.

At one point, a discussion was held on whether this analysis could help improve the mechanical process of product inspection. If the parameter A is nonconforming, then parameter B, which is strongly correlated, must be immediately checked as well. On the other hand, if parameter A is conforming, then parameter B might not have to be checked, freeing some of the capacity of the monitoring center [1].

Fig. 3.7 depicts how the correlation between variables affected the number of undetected nonconformities and how the number of removed variables influenced the result. It shows that if we were to set our correlation coefficient threshold to 0.985, we would be able to remove around 100 characteristics (out of 661) and still be able to detect all nonconforming characteristics, which supports our theory. It should be noted that we only calculated the correlation coefficient on such quality characteristics that *were present in 50 commodities and more* to avoid misinterpretation created from the lack of data.

This theory may also be applied to weaker correlations (to a certain extent), although they do not possess such a strong physical connection. Unfortunately, using the capacities of the monitoring center has not been allowed until now, so the theory has not been tested.

[1] In our case: Let us have a group of n correlated characteristics. We remove such number of correlated characteristics so that there exists at least one characteristic that is strongly correlated to them. We do this for every group of correlated characteristics found

Correlation coefficient threshold	Number of undetected nonconformities	Percentage of undetected characteristics	Number of removed characteristics	Percentage of removed characteristics
0.99	0	0	80	12
0.98	3	3.6	154	23
0.97	4	4.8	207	31
0.96	8	9.5	240	36
0.95	8	9.5	259	39
0.9	8	9.5	304	46

Table 3.2: Influence of correlation on the number of detected nonconformities. Any quality characteristics with their correlation coefficient above the threshold had a respected number of correlated characteristics removed.

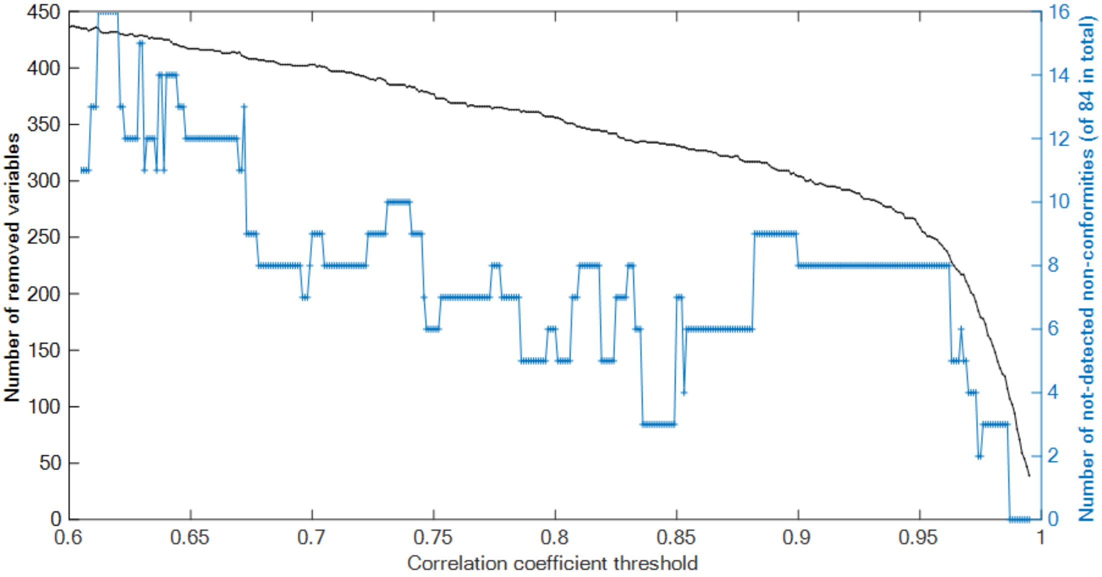


Figure 3.7: Influence of correlation on number of detected nonconfomities. The left axis describes the dependency of the number of removed variables on the correlation coefficient threshold; the right axis describes the number of undetected nonconformities out of 84 nonconformities in total. Any quality characteristics with a correlation coefficient above the threshold had a respected number of correlated characteristics removed.

3.5 Evaluation on real data

To control the quality of the components produced, Škoda uses its own process control method, which alarms the operators upon discovery of a nonconforming characteristic. The staff then investigates the characteristic and, based on their expertise, either decide that it was a false alarm or not. Since the whole manufacturing process of Škoda is very complex, the company relies on the expertise of the staff. The problem, however, remains that there is no definite ground truth because the decision is based on each operator. This also means that the performance of the method does not depend on an objective measurement, since no FNR or FPR rates can be calculated.

The issue of evaluation when there is no ground truth present was discussed and inspected for a long time. Measurement of all (or many) characteristics that are deemed nonconforming by the new methods is not feasible due to the capacity of the measuring center and the capacity of the staff.

The dilemma was later evaded by conducting a series of experiments by adding various shifts to the data and observing the method's performance. Unfortunately, these experiments with real data cannot be made public. However, to gain more flexibility and reliability of the method's performance, a new synthetic dataset based on the real dataset has been generated. The following chapter 4 contains a description of experiments conducted on the synthetic dataset.

Before the dataset shift experiments were conducted and synthetic dataset was generated, the only available ground truth was the Škoda method. For some time, the methods were tested and compared to the Škoda method to adjust the rate of alarms. A table 3.3 shows a FPR, FNR evaluated with respect to Škoda method as ground truth.

	Shewhart	Hotelling	OSVM
FPR	0.663	0.070	0.054
FNR	0.074	0.993	0.321
TNR	0.337	0.929	0.946
TPR	0.926	0.006	0.679

Table 3.3: Evaluation of methods with assumption checking

Let us note that these statistics are actually **not the real rates**. Since the Škoda method does not represent the real truth, this table only represents a table of similarity to each other. We should, however, remember that the company has been operating with the Škoda method for a long time, so it very likely works, although we cannot objectively estimate how much. For that reason, if the methods differ way too much from the Škoda method, the use of such method might not be accepted.

The evaluation is split into a table 3.3 and table 3.4. The previous section on methods in SPC (Sec. 2.4) mentioned that the Shewhart control chart and the Hotelling control chart have certain assumptions that the dataset should fulfill. The difference between these two tables is in the assumption control. In table 3.3 if the assumption is not fulfilled, the characteristic is skipped, and the next one is analyzed, while in table 3.4, these assumptions are completely disregarded, and every characteristic is analyzed.

Although it might seem useless to force the analysis despite not meeting the assumptions, it is believed that not many people check for assumption violation in practice.

The rates of the Shewhart control chart for the assumption checking table were 'better', meanwhile the Hotelling control chart's were mostly 'worse'. This could be explained as the assumptions for the Hotelling control chart are quite strict, compared to the Škoda method.

	Shewhart	Hotelling
FPR	0.676	0.045
FNR	0.134	0.994
TNR	0.324	0.955
TPR	0.866	0.005

Table 3.4: Evaluation of methods without assumption checking

Fig. 3.8 and Fig. 3.9 visualize the boundaries of all three methods in a 2D graph. Fig. 3.8 shows how the flexible boundary of OSVM adapts well to the irregular shape of the data, while the rectangular boundary of UCL, LCL ignores the shape of the data. Next, Fig. 3.9 shows the more rigid elliptical boundary of the Hotelling control chart, which categorizes many points as OOC.

OSVM continued to show good results throughout the research, adapting well to various irregular data shapes created by complex manufacturing processes of products. The method performs quite well even in cases where products are barely nonconforming, showed in figure 3.10.

However, OSVM has another practical value. Since OSVM can be easily adjusted and its limit can be tightened or relaxed, its boundary can be set to produce similar results to the Škoda method to reduce the initial shock in the beginning, from changing the process control techniques. The boundaries can then be slowly restricted over time.

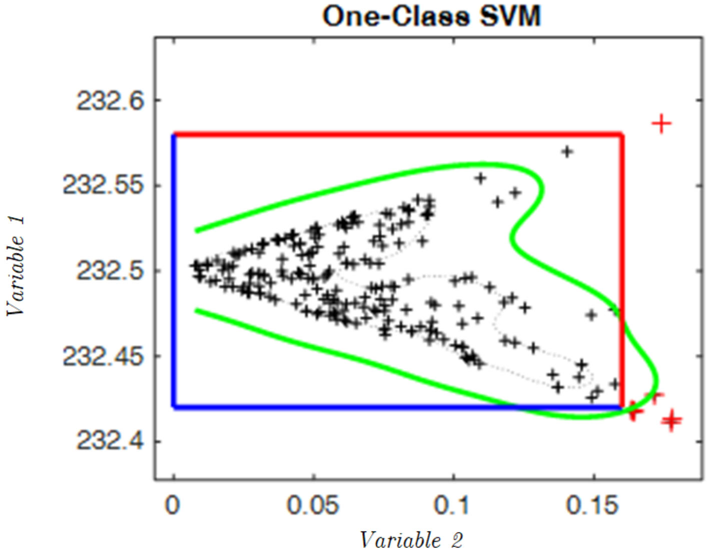


Figure 3.8: OSVM (green line) and Shewhart chart boundaries (red (UCL) and blue (LCL)). Variable 1 and variable 2 are quality characteristics.

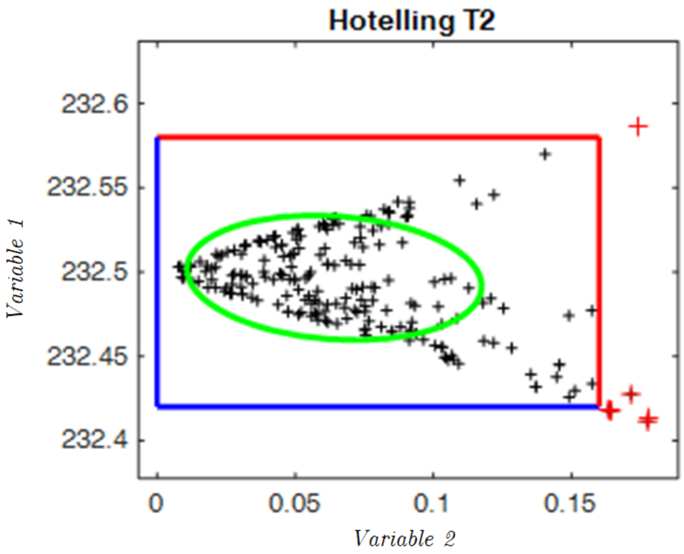


Figure 3.9: Hotelling (green line) and Shewhart chart boundaries (red (UCL) and blue (LCL). Variable 1 and variable 2 are quality characteristics.

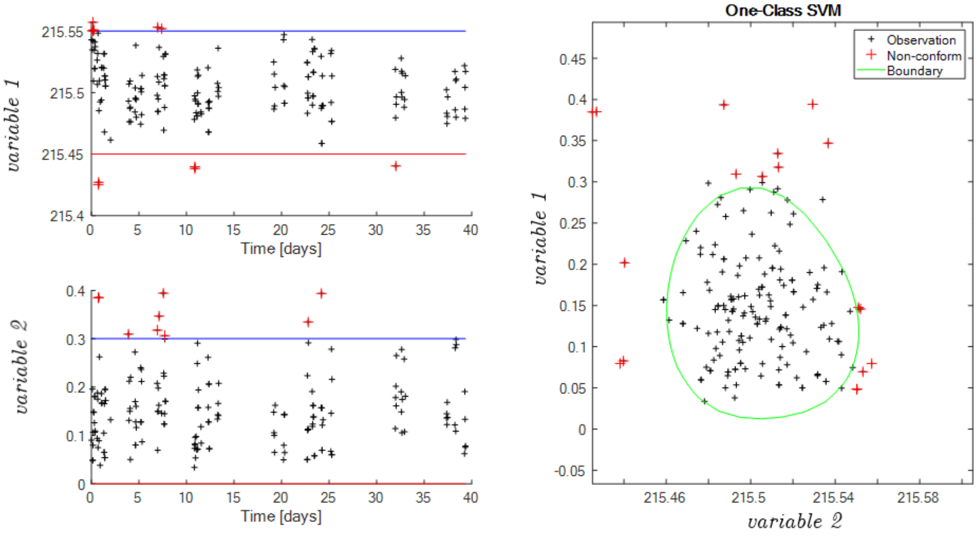


Figure 3.10: Flexible evaluation boundary of OSVM (green line) on the right. Variable 1 and variable 2 are quality characteristics.

Chapter 4

Experiments

This chapter is dedicated to experiments conducted on the synthetic dataset. As already mentioned in the previous sections, we currently cannot acquire the ground truth due to the complexity of the mechanical procedure of obtaining it. However, it was clear from the beginning that we cannot only rely on the existing Škoda method. Moreover, if we did not generate synthetic data, we would not be able to publish the results due to the confidentiality of the data.

The results of Experiment I were also administered in a conference paper at the 13th International Conference on Soft Computing and Pattern Recognition 2021 and as a result, published in Lecture Notes in Networks and Systems [1]. The results of Experiment II, as of now, were submitted to the International Conference on Systems, Man, and Cybernetics 2022 and are currently awaiting peer review.

4.0.1 Synthetic data generation

Synthetic data for both experiments are generated a bit differently due to the nature of the experiments. The first experiment focuses on the innate detailed performance of the methods in a 5-modal and multivariate environment of 2D data, where we can thoroughly visualize and examine the results. After an overview of the method's performance is gained, the second experiment utilizes a much more realistic and complicated dataset, where it is no longer easy to visualize or interpret the results in detail.

For Experiment I, we generated a bivariate distribution with two correlated variables. The first variable has a gamma distribution with shape parameter 3 and scale parameter 0.5, while the second variable has a gamma distribution with parameters 1 and 1. Subsequently, a normal cumulative distribution function (cdf) was applied to a standard normal random variable, resulting in a uniform random variable in the interval [0,1]. Using the theory of univariate random number generation, the inverse cdf of any distribution F on a U(0,1) random variable produces a random variable of which the distribution is exactly F. Applying the two-step transformation to each variable produces a dependent random variable with arbitrary marginal distribution. In this way, five datasets (representing 5-

modes), each consisting of 1000 data points, were created. Each such dataset will be further shifted to simulate five modes of different means as described in the next section.

For Experiment II, we again have 5 sets of generated data to generate the multimodality of the process (5 modes), and the multivariety of this experiment is increased to 12-dimensions. To represent a more complex correlation structure, we divided the 12 variables into three groups and generated highly correlated data for each mode within each such group, where the correlation coefficient was 0.9 as shown in Fig. 4.10. Therefore, we have 3 groups of correlated variables, where *each group was generated separately.* In this experiment, the modes have a normal distribution.

4.1 Experiment I

To create multimodal data, we generated five sets of data points consisting of 1000 points each, and these sets were shifted by the following small vectors:

$$\begin{aligned} v_1 &= [0, 0] \\ v_2 &= [1, 0] \\ v_3 &= [0, 1] \\ v_4 &= [2, 1] \\ v_5 &= [0, 0] \end{aligned}$$

Then every set represents one mode of the multimodal process.

The out-of-control part of the dataset was produced by applying a predefined shift S to both variables. That is done by adding $S = [s_1, s_2]$ to every out-of-control observation. This shift was applied to 2000 observations out of 5000 observations, resulting in 3000 observations labeled as in-control and 2000 out-of-control observations.

The previous procedure enables us to better control the process shift and evaluate the methods using our chosen evaluation techniques, FPR and FNR . It should be noted that the training of the machine learning-based OSVM corresponds to the first phase of SPC (Phase I, described in 2.4.2), since learning in OSVM is, in fact, *setting bounds for the control chart.* Learning was performed on 1000 in-control observations (since OSVM learns only on clear data), and evaluation was performed on the remaining 2000 in-control observations and 2000 out-of-control observations. All data for different sizes of the shift S are depicted in Fig. 4.5. One can observe the bivariate multimodal distribution of two correlated variables.

This experiment was carried out with three methods: Shewhart control chart, Hotelling T^2, and one-class SVM, where one is univariate, one is multivariate, and one is based on machine learning. Moreover, the goal of this experiment is to examine the performance of these methods on a complex multimodal process.

The performance of these three methods was tested by iterating through different directions and sizes of the shift, and the shift constant is constantly changed in the interval from -6 to $+6$. The FPR and FNR were calculated for 21 different values of the shift.

There are three following ways how to shift the data, considering the data mimics the real-world data, and the experiment hopes to preserve the correlation in the data:

- Diagonal shift

  $$s_1 \neq 0, s_2 \neq 0$$

  Both correlated variables shifted in the same direction.

- Horizontal shift

  $$s_2 = 0$$

  One variable is shifted.

- Vertical shift

  $$s_1 = 0$$

  One variable is shifted.

4.1.1 Results

The FPR of the testing set was selected to be the same for all methods to fairly test performance. First, Shewhart's FPR was measured, and at the same time, both the upper bound for the T^2 statistics and the threshold of OSVM were set in order to achieve the same FPR on the testing set. Despite using the testing set to set the FPR value, the results should not be impartial, as the test set was used, in fact, only to set the FPR value. Since the whole SPC theory is based on controlling false alarms, experiments measure performance through FNR.

A control area bounded by a decision boundary is presented in Fig. 4.5 for each method:

- The Shewhart control chart possesses a rectangular-shaped control area, created by its LCL and UCL.
- The Hotelling control chart has a control area in the shape of an ellipse, defined by its UCL.
- OSVM control area has the shape defined by its support vectors and area given by threshold for OSCVM score. It creates a flexible boundary around the data points.

Fig. 4.9 shows the graphical representation of the performance of each method for each shift. The first row of the x-axis represents s_1 and the second row represents s_2.

The blue line representing OSVM is almost at all points of the graph under the rest of the lines, which signifies better performance. As we can see, the FNR of Shewhart is already larger than $FNR = 0.1$ with such a small horizontal shift as $s_1 = -4.857$. OSVM doesn't reach $FNR = 0.1$ until between around $s_1 = -3$. For the vertical shift, the performance of the Hotelling control chart improved substantially, even being on par with OSVM for the second half of the graph (positive shift in s_2). For the diagonal shift, the Shewhart control chart also improved for the positive s_1 and s_2

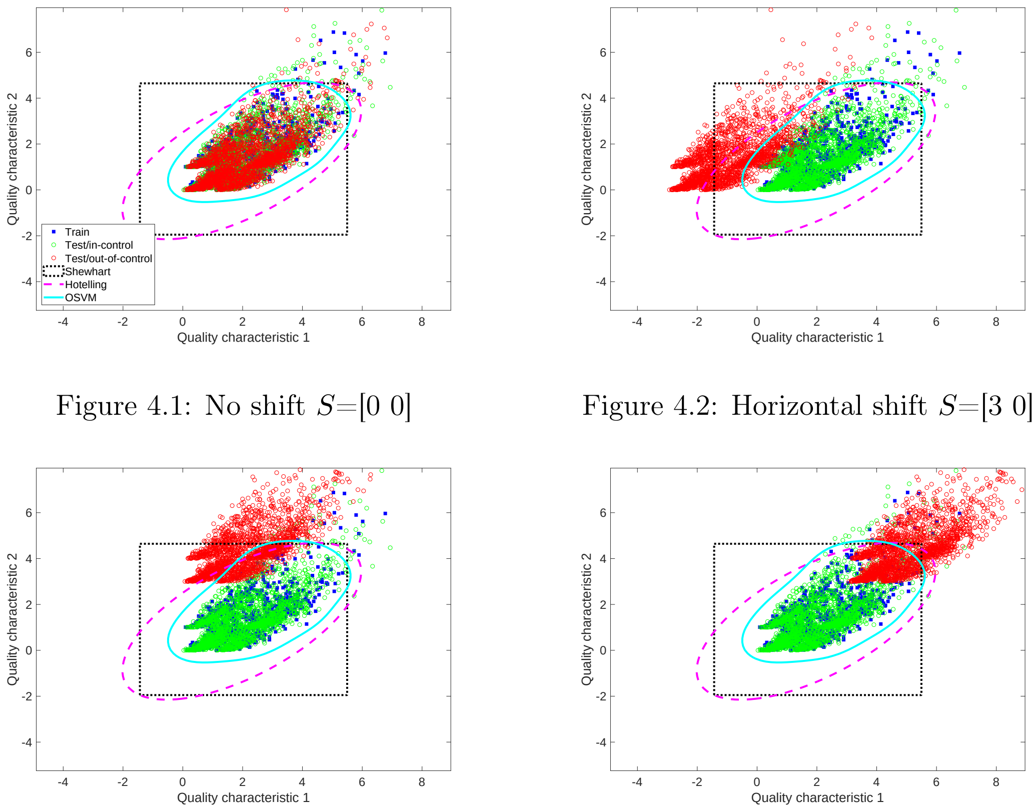


Figure 4.1: No shift S=[0 0]

Figure 4.2: Horizontal shift S=[3 0]

Figure 4.3: Vertical shift S=[0 3]

Figure 4.4: Diagonal shift S=[3 3]

Figure 4.5: Examples of different types of process shift

shifts, approaching the other two methods. However, for the negative shifts, it continues to perform poorly.

The univariate nature of the Shewhart control chart applied to the bivariate process failed to detect shifts of a single variable, either s_1 or s_2 corresponding to a horizontal and vertical shift. The diagonal shift was, however, noticeably better.

The most distinct moment remains the horizontal shift with the difference of OSVM and Hotelling control chart being around 30 % for most sizes of the shift. The considerable performance of OSVM in the negative shift detection in all shifts compared to the other two methods is caused by the positive skew of the data distribution.

4.1.2 Discussion

The synthetic data was purposely generated as multivariate, multimodal, and purposely correlated to authentically mimic the real data from Škoda Auto authentically. Due to that, the poor results of the univariate Shewhart control chart were not surprising, as the data violated almost every presumption of the method. The issue with using a method for monitoring a single variable for a multivariate dataset lies within the rise of false alarms caused by the multiple comparisons problem. In this case, false alarms increased almost ten times from the expected 0.0027 intended six-sigma type I error probability up to 0.026. The Shewhart control chart also performs especially poorly when the data are correlated.

However, the correlated data attributes are quite a norm in the manufacturing industry as perhaps no complex system can really function without a subsystem structure. Despite the fact, most multivariate methods require a normality assumption, including the Hotelling control chart, which in this experiment had its assumption strongly violated by the nonnormality caused by the multimodal nature of the generated synthetic dataset with five modes, each generated from a not normal distribution (gamma).

As multivariate SPC methods are quite often used in the industry, one of the possibilities for correct usage of Hotelling would be to know the real probability distribution of the data and have enough data collected to approach the distribution. That is, however, a challenging task. The real probability distribution is commonly unknown and is often extremely costly to estimate. Second, collecting enough data to approach the original distribution might pose a problem.

OSVM is a distribution-free method that uses support vectors to enclose the data points. The support vectors are based on the most important sample points in the dataset, which also offers some additional information. The support vector-based boundary is visibly much more flexible than the ones from the classical SPC methods. Due to its boundary, it is also more effective in higher dimensions [28].

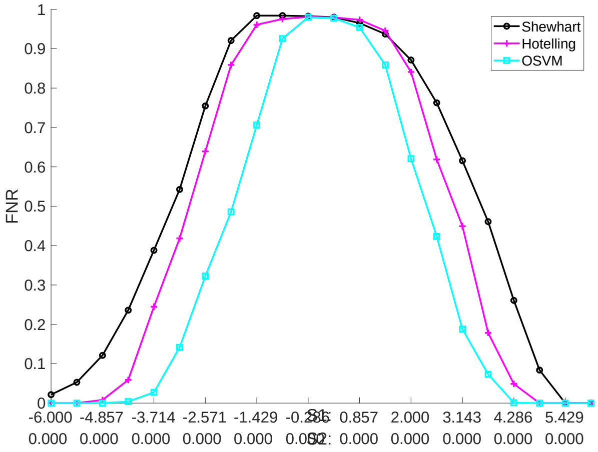


Figure 4.6: Horizontal shift

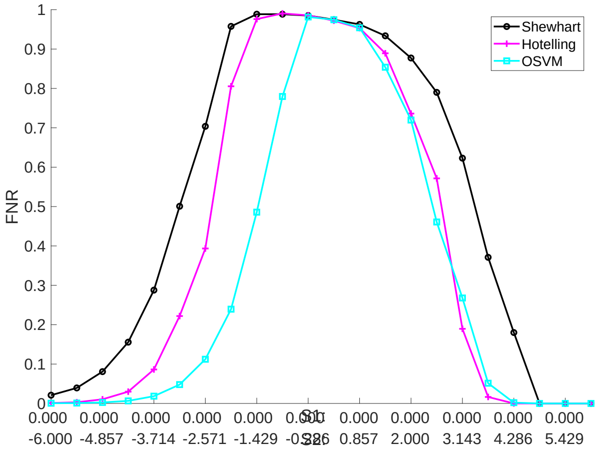


Figure 4.7: Vertical shift

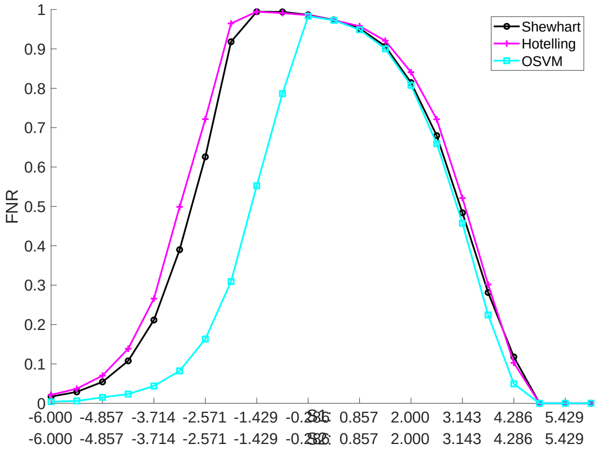


Figure 4.8: Diagonal shift

Figure 4.9: Comparison of the Dependence of FNR on the shift

Table 4.1: Dependence of FNR on the shift

s_1	s_2	Shewhart	Hotelling	OSVM
-6	0	0.021	0	0
-5.429	0	0.053	0	0
-4.857	0	0.121	0.008	0
-4.286	0	0.236	0.059	0.004
-3.714	0	0.388	0.245	0.027
-3.143	0	0.543	0.418	0.142
-2.571	0	0.755	0.639	0.323
-2	0	0.921	0.859	0.486
-1.429	0	0.984	0.961	0.706
-0.857	0	0.984	0.976	0.926
-0.286	0	0.983	0.981	0.98
0.286	0	0.98	0.98	0.978
0.857	0	0.965	0.974	0.954
1.429	0	0.937	0.946	0.859
2	0	0.871	0.841	0.621
2.571	0	0.763	0.619	0.423
3.143	0	0.615	0.449	0.188
3.714	0	0.461	0.179	0.073
4.286	0	0.261	0.049	0.001
4.857	0	0.084	0	0
5.429	0	0	0	0
6	0	0	0	0

Table 4.2: Horizontal process shift

s_1	s_2	Shewhart	Hotelling	OSVM
0	-6	0.021	0.001	0
0	-5.429	0.039	0.003	0.001
0	-4.857	0.081	0.01	0.002
0	-4.286	0.156	0.03	0.006
0	-3.714	0.288	0.086	0.018
0	-3.143	0.501	0.222	0.048
0	-2.571	0.704	0.394	0.112
0	-2	0.958	0.805	0.239
0	-1.429	0.989	0.976	0.486
0	-0.857	0.989	0.991	0.78
0	-0.286	0.985	0.986	0.982
0	0.286	0.975	0.972	0.975
0	0.857	0.963	0.953	0.954
0	1.429	0.934	0.889	0.854
0	2	0.877	0.736	0.72
0	2.571	0.79	0.572	0.461
0	3.143	0.623	0.19	0.268
0	3.714	0.371	0.016	0.051
0	4.286	0.18	0	0.002
0	4.857	0	0	0
0	5.429	0	0	0
0	6	0	0	0

Table 4.3: Vertical process shift

s_1	s_2	Shewhart	Hotelling	OSVM
-6	-6	0.017	0.021	0.004
-5.429	-5.429	0.029	0.037	0.006
-4.857	-4.857	0.054	0.07	0.015
-4.286	-4.286	0.108	0.139	0.023
-3.714	-3.714	0.211	0.266	0.044
-3.143	-3.143	0.39	0.499	0.082
-2.571	-2.571	0.626	0.722	0.163
-2	-2	0.918	0.965	0.309
-1.429	-1.429	0.995	0.995	0.553
-0.857	-0.857	0.994	0.991	0.787
-0.286	-0.286	0.987	0.986	0.983
0.286	0.286	0.974	0.973	0.973
0.857	0.857	0.952	0.958	0.949
1.429	1.429	0.907	0.921	0.9
2	2	0.814	0.841	0.807
2.571	2.571	0.68	0.721	0.659
3.143	3.143	0.484	0.521	0.457
3.714	3.714	0.281	0.302	0.224
4.286	4.286	0.118	0.103	0.049
4.857	4.857	0	0	0
5.429	5.429	0	0	0
6	6	0	0	0

Table 4.4: Diagonal process shift

4.2 Experiment II – Adding dimensions

The previous experiment examined the performance of methods on a simple multivariate, multimodal dataset with correlated variables. Additionally, this section expands the scope of the first experiment, bringing a dimension of size $d = 12$ to test the performance of the previously tested methods on the small-dimension dataset.

This dataset models the characteristics and tendencies that today's manufacturing data often have. This experiment focuses on the performances of the three previously tested methods – Shewhart control chart, Hotelling's control chart and OSVM on a much more complex and much more realistic set of data.

Out of 1000 samples for each set, 200 samples are used for testing and 800 samples for training. Out of the training samples, 400 samples were shifted by a randomly generated shift vector in the range from -4 to 4 to represent out-of-control data. The rest of the training samples are not shifted and they represent the in-control data samples. In summary, we have 5000 data samples in total, where 4000 are used for training, 1000 for testing.

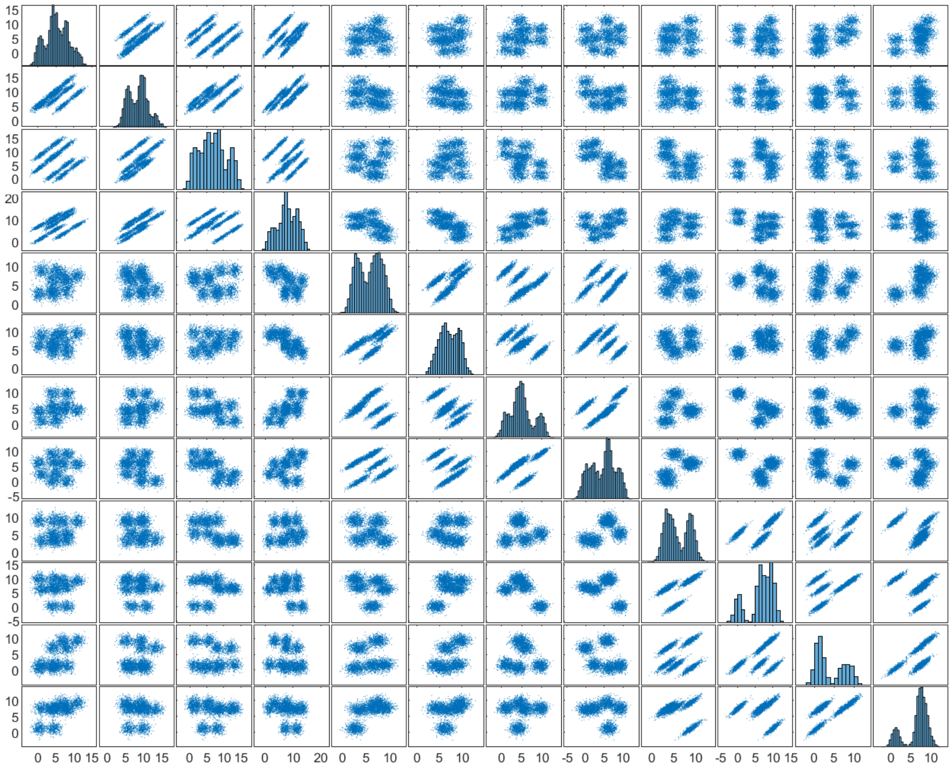


Figure 4.10: Scatter matrix of 12D generated data

In this experiment, the horizontal, vertical, and diagonal shifts now no longer make sense. Let us have a general randomly generated shift vector s that shifts around the correlated characteristics. With a much more complex dataset, several scenarios need to be tested. The experiment on shift s is conducted in the following manner [1]:

- variables 1-4 shifted as depicted in 4.11,
- variables 5-8 shifted as depicted in 4.13,
- variables 9-12 shifted as depicted in 4.12
- one single variable shifted as depicted in 4.14.

This setting was chosen deliberately as it imitates the behavior of the real data in the manufacturing industry. Correlation is, in fact, a very intriguing and common occurrence created by many different circumstances, and it is crucial to understand why specific characteristics correlate.

Certain characteristics are correlated because they are *related* in a *physical sense*, meaning that if one's value is physically shifted, then the other one's value is shifted as well. The previous sections have provided the reader with an example of the depth of holes which resulted in an almost linear correlation. Another realistic example would be the correlation created by deteriorating machinery. Groups of characteristics are always processed together, and if a dull gauge is used on them, all these characteristics will be correlated by a joint shift from a target value resulting from the same faulty gauge. This correlation is advantageous as it can quickly uncover possible nonconforming products or even groups of products. Similarly, as was mentioned before, with the linearly correlated characteristics, the capacity of measurement centers could be freed by taking advantage of correlations among characteristics. That is, of course, under the assumption that such intervention in the measuring processes was allowed.

[1] The x and y labels were intentionally excluded due to the size of the subplots

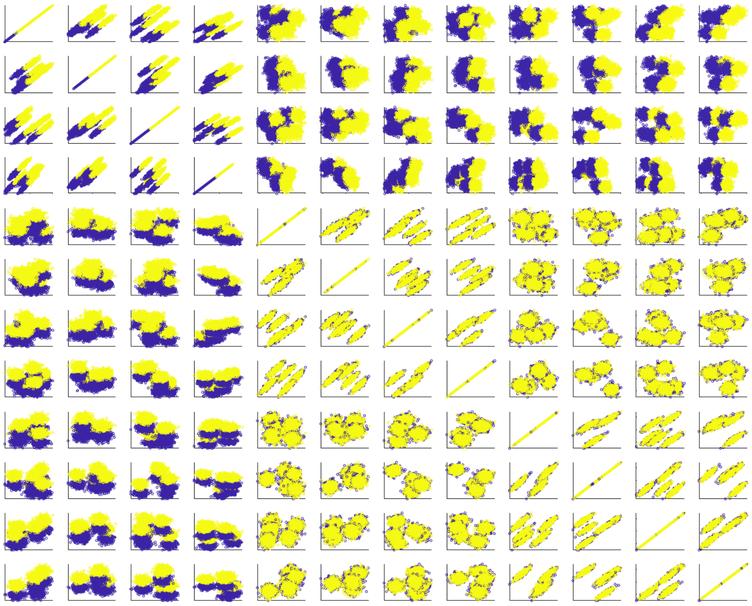


Figure 4.11: Variables 1-4 shift. Violet data points signify a shift.

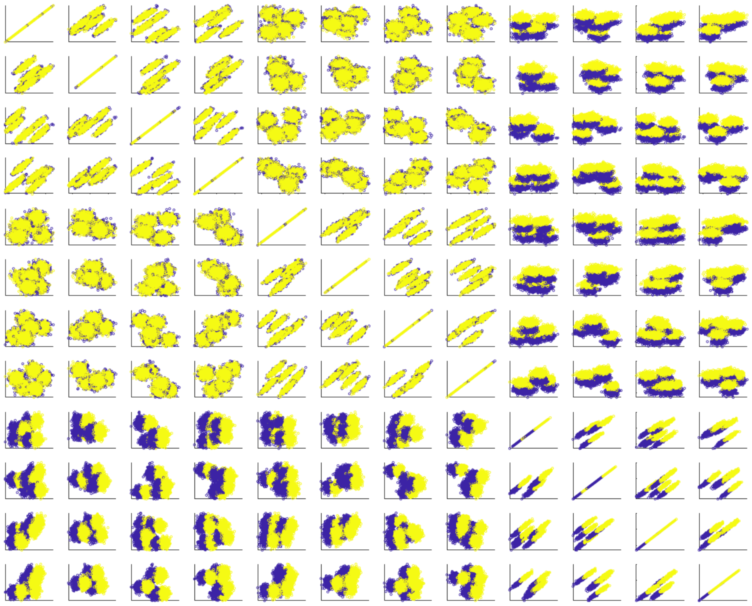


Figure 4.12: Variables 9-12 shift. Violet data points signify a shift.

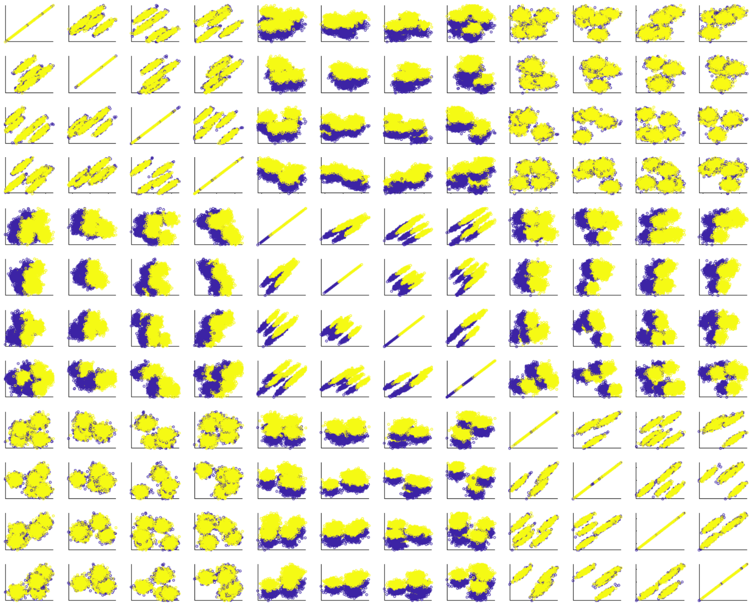


Figure 4.13: Variables 5-8 shift. Violet data points signify a shift.

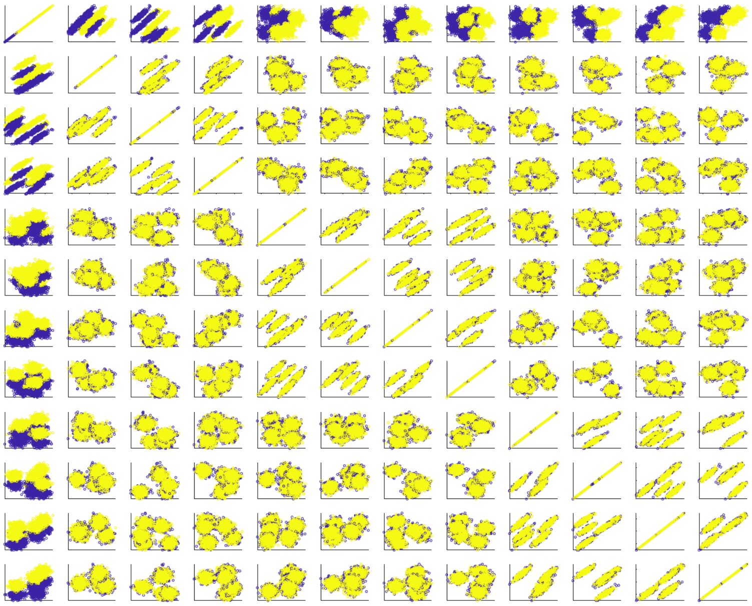


Figure 4.14: Only variable 1 shifted. Violet data points signify a shift.

4.2.1 Results

Since the methods are evaluated on a dataset with noticeably higher dimensionality, they performed much worse than their 2D counterparts. Shewhart control chart performed the worst out of the three methods, again due to the nature of the data as it evaluates each variable independently. In higher dimension and correlation between each 4 variables, the control chart's FNR already starts around $FNR = 0.8$ and never goes under 0.5. The more correlations the data have and the higher dimensions it operates under, the worse the performance of the Shewhart control chart.

In the single variable shift situation, the Hotelling control chart completely outperformed the other two methods. In this case, the assumption of independence of the variables acted as an immense advantage for the method, as shown in Fig. 4.18.

However, except for this moment, Hotelling control chart's FNR rate happens to be around $FNR = 0.4$ at best. In Fig. 4.15, Fig. 4.16 and Fig. 4.17, OSVM outperforms Hotelling quite considerably.

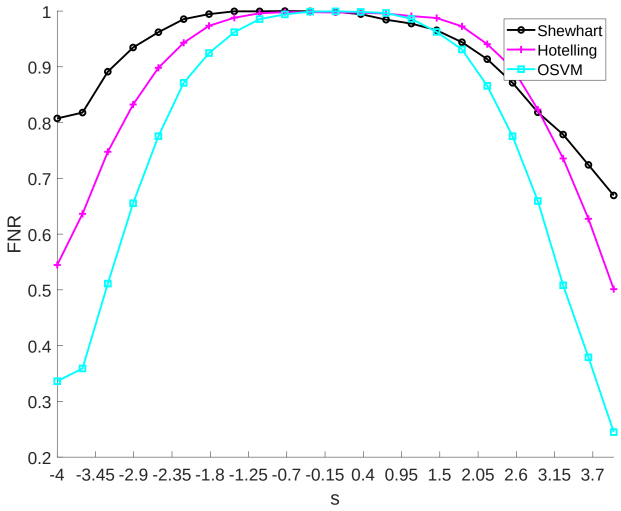


Figure 4.15: Performance of methods with the variables 1-4 shifted

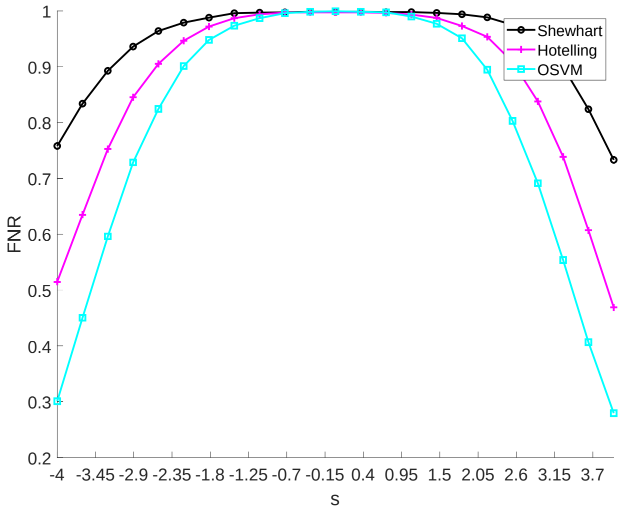


Figure 4.16: Performance of methods with variable 5-8 shifted

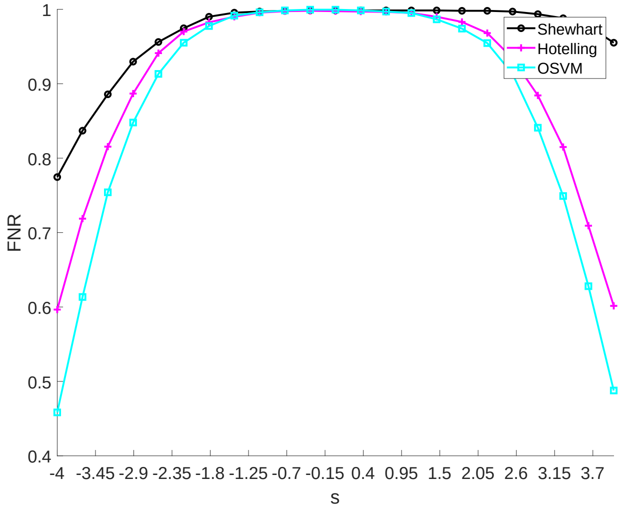


Figure 4.17: Performance of methods with variable 9-12 shifted

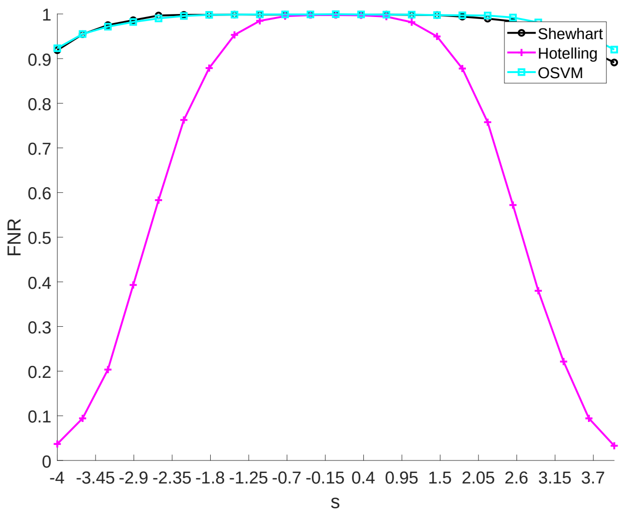


Figure 4.18: Performance of methods with variable 1 shifted

4.2.2 Discussion

Additional experiments were conducted on a 12-dimensional dataset of multivariate, multimodal, and correlated variables, where every four variables – variables 1-4, variables 5-8, and variables 9-12 were correlated to each other. Four settings were proposed to test the performance of the methods.

As the dimensionality of the dataset was raised, and several variables were correlated, the Shewhart control chart performed noticeably worse than in the previous experiment with 2D data. Seeing the poor performance of the Shewhart control chart on a dataset that is not that much more complex than the datasets that can be seen in real life, it is surprising how popular the method still is. However, in reality, the companies do not rely on SPC by themselves to assure the quality of the products. Most of the time, quality control is acquired by combining simple SPC methods and the experience of the operator and domain experts.

Although OSVM outperformed the other methods most of the time, the interesting situation depicted in Fig. 4.14 needs to be addressed. In this case, only one variable is shifted despite being a part of the 1-4 variable correlation. In other words, the shift is meant to describe a sole shift of variable 1 that is correlated with variables 2, 3, and 4, but the former variables are not shifted. Such a situation was not previously considered when discussing the nature of correlations in manufacturing industries, as under such circumstances, this portrays an unpredictable mistake.

For simplification, let us imagine a situation where a machining process is processing a mechanical nut that has a threaded hole, and each depth of the hole is its characteristic. Now, a few of these nuts fell on the ground during the transport. During the process control, one of the correlated characteristics is shifted, but the others are not because the dirt from the previous transportation mistake was trapped on one part of the threaded hole.

Considering the complexity of the dataset, the popularity of the Hotelling control chart is quite understandable. Despite the assumption violation and the high-dimensionality, it performs sufficiently well, and the independence assumption even acted as an advantage in certain situations.

Except for one situation, OSVM seems to be reasonably suitable for SPC in a complex dataset environment. Unfortunately, it cannot detect single-variable shifts that are correlated. Furthermore, the interpretability of the methods, defined as finding variables that are responsible for positive OOC state detection, is also important. Interpreting the results of multivariate methods is much more difficult than interpreting the results of univariate methods. This could be solved via recent methods of explainable AI, which is, however, out of scope of this thesis.

Chapter 5

Application CIRQUE

The continuous collaboration of the CIIRC research team and the Škoda mechanical engineer team called for a simple and effective means of result representation that can be reused at any time. That led to the development of the application CIRQUE that conveys the decision of the method in a more understandable intuitive graphical form for a quick check of results instead of analyzing abstract pieces of information “by hand”.

The goal of this application is to do the following:

1. load raw data files generated from a system in Škoda Auto,
2. analyze the data on the previously implemented methods,
3. visualize the results.

Results are saved upon finishing the analysis in a user-specified directory for future revision of these results. The flow of the application is depicted in a diagram 5.1

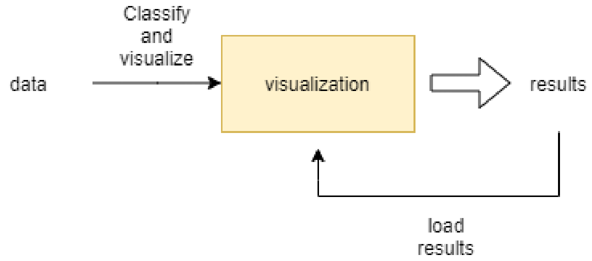


Figure 5.1: Application flow diagram

It is important to note that the data we firstly acquired to train our model are not the same as the raw data files. After the operator acquires the raw data files created from the processing system in Škoda auto, the file is enriched and processed into an Excel table with additional information. The final table contains organized, easy-to-read information, cleared from any redundant metadata. After some time, the raw data files are discarded.

The raw data files contain data in the most unprocessed state in a disorderly structure and are extremely difficult for a human to read. To process them into the application, the structure had to be continuously discussed with the head operator. By processing the raw data files into our application, we maintained more effective research as we could analyze each data file to quickly obtain the results and also because we did not need to wait for the Excel file to be produced. Another advantage proved to be that by understanding the data structure, we could update our model continuously, combating a potential data drift. The development of CIRQUE created a possibility to quickly analyze the most recent data and check the most obvious mistakes. It was a useful tool for showcasing our methods to the colleagues from Škoda Auto or other domain experts on various inside-project presentations.

5.1 Application architecture

The architecture of the CIIRQUE application is divided into three parts – the backend, the frontend and the analysis engine, which consists of the methods and trained models analysing the data. Since all of the research team members needed to interact and experiment with the models themselves, Matlab was chosen for the implementation of the statistical methods as all members were proficient in this language. Python was chosen to be a backend language, mainly due to its convenient Matlab integration as Matlab allows exporting its scripts as Python libraries.

Since the application needed to plot non-standard graphs and present non-standard visualizations, non of the graphical Python libraries were utilized. Instead, a framework based on JavaScript, HTML and CSS (Electron) was used as it provided all the necessary tools – it was beginner-friendly, highly customizable and could be run on any computer.

Backend

The backend of the application is divided into four python files:

- **data_matlab.py** attends to the extraction of the metadata and other necessary data about the characteristics from the standard Škoda file. It also pairs correct combinations of mechanical machining information, pairs correct character names and finally, it runs, exports and imports analysis.
- **main.py** initializes server necessities.
- **matlab_wrapper.py** contains instantiation of the matlab library, taking care of the initialization of the matlab runtime and handles the communication with the matlab runtime.
- **structs.py** contains structures for several methods, exporting necessary information for later use for graph plotting.

Frontend

The frontend of the application is divided into the following files:

- **main.css** contains all styling of the application.
- **main.html** contains layout of the application graphical interface.
- **main.js** instantiates the Electron application.
- **ooc-window.html** creates the out-of-control window for groups of characters upon mouse click.
- **renderer.js**

Analysis engine

The models and methods for analysis are contained in the following files:

- **hotelling.m** implementation of the Hotelling control chart.
- **shewhart.m** implementation of the Shewhart control chart.
- **one-class.m** implementation of the OVSM.
- **skoda.m** implementation of the existing Škoda method.

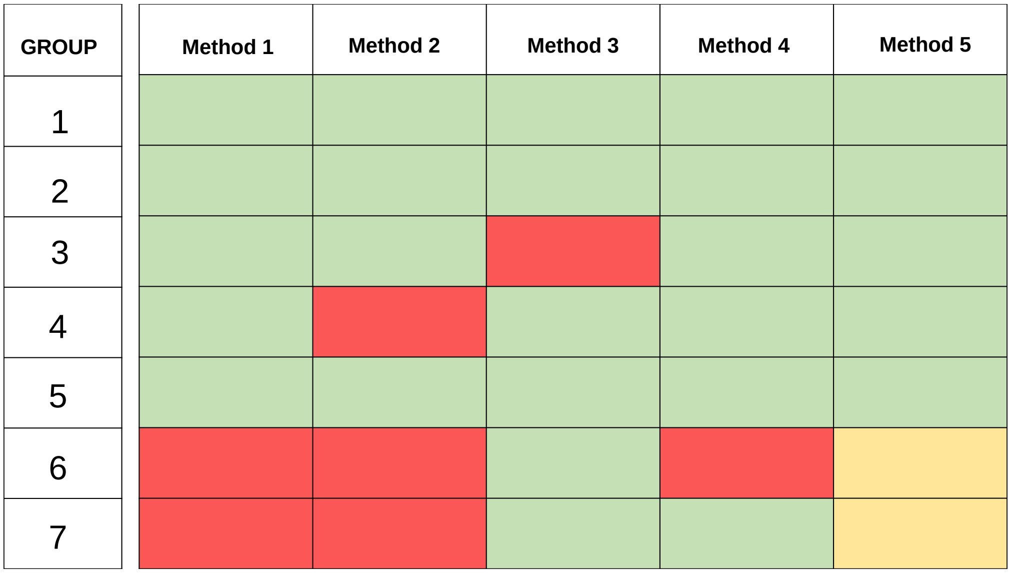
All scripts from the analysis engine were exported as a Python library and initialized in the matlab_runtime.py.

5.2 Visualization

It was crucial for the visual aspect to be intuitive and simple but still offer some interesting functions. These were the following conditions that needed to be considered:

- We need to show clearly whether the method decided that the character was nonconforming. We have several methods. How do we show this in a simple preemptive way?
- If the method decided that there was something wrong with the product, we should be able to analyze it further. What other information do we show?

To communicate the information effectively about the method's initial decision were used to show whether the method suspects the characters to be nonconforming. The results are shown in a tabular setting, where each column signifies the decision of the method, and each row is the group that the characteristic belongs to. The groups are predetermined by the machine layout. This visualization is used as an initial overview of the classification/method's results, as is shown on the design proposal 5.10.



GROUP	Method 1	Method 2	Method 3	Method 4	Method 5
1	🟩	🟩	🟩	🟩	🟩
2	🟩	🟩	🟩	🟩	🟩
3	🟩	🟩	🟥	🟩	🟩
4	🟩	🟥	🟩	🟩	🟩
5	🟩	🟩	🟩	🟩	🟩
6	🟥	🟥	🟩	🟥	🟨
7	🟥	🟥	🟩	🟩	🟨

Figure 5.2: Design proposal of result visualization.

The red square depicts that there is at least one nonconforming character in the group of characteristics. Green square depicts that all characteristics within the group are conforming. The yellow square is that crucial assumptions are violated and the method does not yield reliable results.

The visualization proposal presents the decision of the method promptly, with minimal room for misunderstanding due to the *large discriminability between the color hues.* The conforming group of characteristics (red square) is also clickable, presenting the characteristics that are the cause of the nonconformity. Certain methods are not so straightforward, such as those based on machine learning. For those, instead of presenting characteristics that *are the cause of nonconformity*, the application offers characteristics *suspicious of causing the nonconformity.*

Nonconforming characteristic graph

For some of these methods, if the group of characteristics is conforming (group shows red square), clicking the red square reveals exactly which character's fault it is. More specifically, it reveals a graph inspired by a statistical box plot for each character, as is shown in 5.3.

The graph presents the user with all of the available information about the characteristic and the information about its historically measured values. The green box revealing historically measured values could help to detect whether some particular characteristics have tendencies to skew. Regulation limits represent the upper and lower specification limits USL, LSL and action limits stand for the upper and lower control limits UCL and LCL. Currently, the methods alarm the user the moment the regulation limits are exceeded.

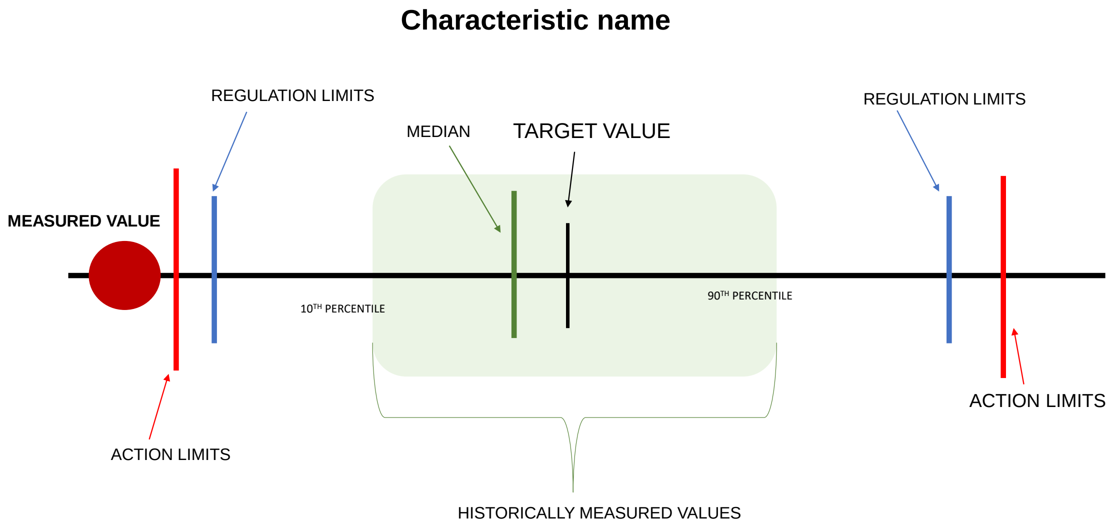


Figure 5.3: Proposal for visualization of nonconforming characteristic graph.

A version of this graph is used for the nonconforming characteristic graph for the OSVM algorithm as well. Since the machine learning algorithm was not very interpretable for our mechanical engineering colleagues, they expressed a desire to have a graph that would help them understand what the algorithm does. Although the OSVM graph might not make much sense to a person with a machine learning background, it has proved to help people from other backgrounds to orientate around the result.

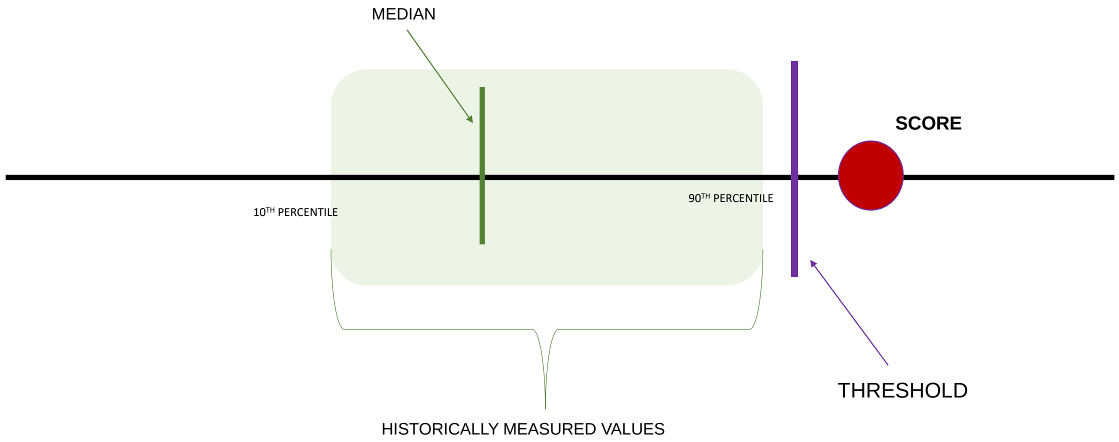


Figure 5.4: Proposal for visualization of nonconforming characteristic graph OSVM.

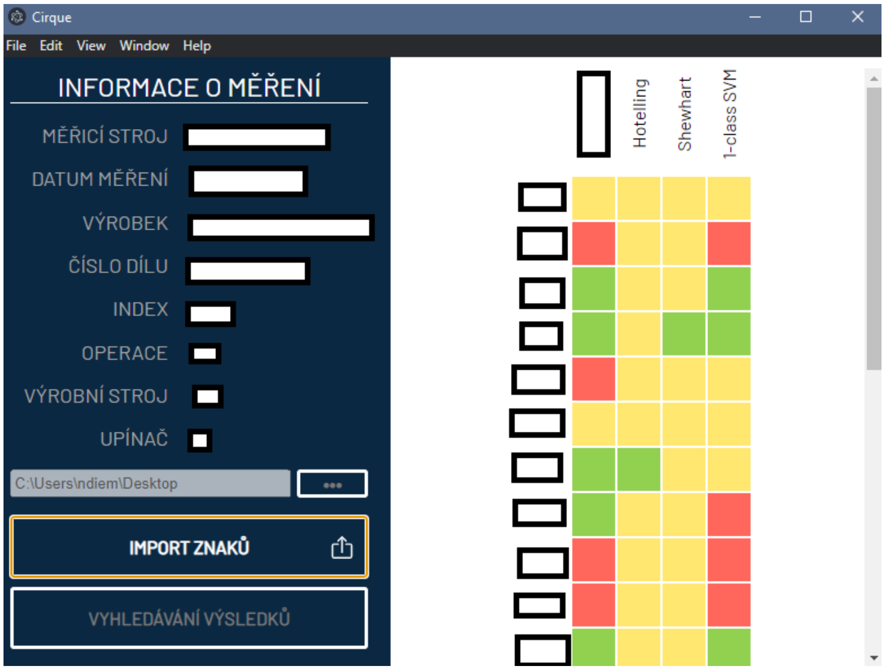
From the visualization theory point of view, this graph has made use of two identity channels in combination – **shape** and **hue**. Considering that the focus of the graph is either on the characteristic's measurement or OSVM score, the measured value and the score are mapped onto a *dot mark*, which is never used for any other information in the graph. For the original graph, both control and regulation limits were placed as a *line mark*, differentiated by color. The action limit marks are color mapped to red, and the regulation limits are color mapped to blue. Regulation limits marks are shorter in height to intuitively indicate that the limits are smaller in reality and the measured value could reach them sooner.

In the implemented version of both the graphs, the most important data is shown below, rounded to four decimals. Measurements and other digits are shown upon hovering on each visual mark. The OSVM graph also offers an explanation of the score to the user upon hovering. The implemented versions of the graph can be seen in Fig. 5.11 and Fig. 5.10 in the following section.

Secondary visualization and user interface traits

To effectively express the correct information, the application implemented various features from the theory of visualization. Besides the most crucial part of the application, the theory was applied to the following parts of the application as well:

- The left part of the application shows metadata so it is always clear and visible, what part we are currently analysing.
- While scrolling, method's header is anchored to avoid visual cluster.
- Hovering above a button triggers a short animation.
- While the program is working on the analysis, a revolving gear is shown to exhibit that it is still ongoing.



INFORMACE O MĚŘENÍ

MĚŘICÍ STROJ [redacted]
DATUM MĚŘENÍ [redacted]
VÝROBEK [redacted]
ČÍSLO DÍLU [redacted]
INDEX [redacted]
OPERACE [redacted]
VÝROBNÍ STROJ [redacted]
UPÍNAČ [redacted]

C:\Users\ndiem\Desktop

	[redacted]	Hotelling	Shewhart	1-class SVM
[redacted]	🟨	🟨	🟨	🟨
[redacted]	🟥	🟨	🟨	🟥
[redacted]	🟩	🟨	🟨	🟩
[redacted]	🟩	🟨	🟩	🟩
[redacted]	🟥	🟨	🟨	🟨
[redacted]	🟨	🟨	🟨	🟨
[redacted]	🟩	🟩	🟨	🟨
[redacted]	🟩	🟨	🟨	🟥
[redacted]	🟥	🟨	🟨	🟥
[redacted]	🟥	🟨	🟨	🟥
[redacted]	🟩	🟨	🟨	🟩

Figure 5.5: Metadata, buttons all on the left, classification on the right

This application was designed with an intent of deployment to Škoda Auto. It was supposed to be operated by the following type of people:

- The people operating the machines without any statistical or computer science background will use it for assistance in out-of-control detection.
- Mechanical engineers/other domain experts will use it for analysis to further improve the process.

5.3 Discussion

At first, the plan was to use the application at Škoda by domain experts and mechanical operators. Since the current raw data format is challenging to evaluate quickly, the application intended to provide visual feedback and analysis of the characteristics. The information and graphs of other methods were also supposed to provide a tool for mechanical engineers to examine further.

Since domain expertise and knowledge of inside information were needed for many parts of the implementation, the development of the application took place under the supervision of colleagues from Škoda Auto. The visualization and user interface proposal was extensively discussed before the implementation, and suitable adjustments were made. The implementation utilized the available data from mechanical processes and character parameters. Several datasets were being collected for months until they could be implemented into the application.

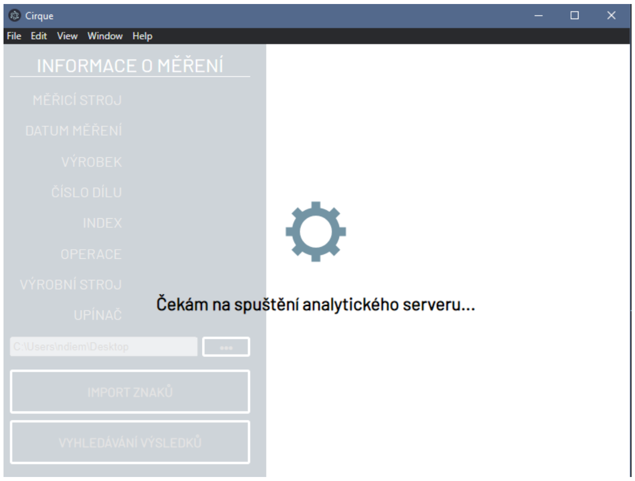
5.4 Example of application usage

The application, unfortunately, contains a considerable amount of inside information. The data, the know-how, and its code cannot be released publicly. Such information is interlaced throughout the application architecture, from the data we use and load to the underlying tables containing machine information for the correct visualization. For that reason, sole anonymization of the data regrettably does not suffice as the whole codebase would have to be rewritten for it to function.

This section at least provides an example of usage with a series of screenshots and descriptions of the application's behavior.

1. Initial launch

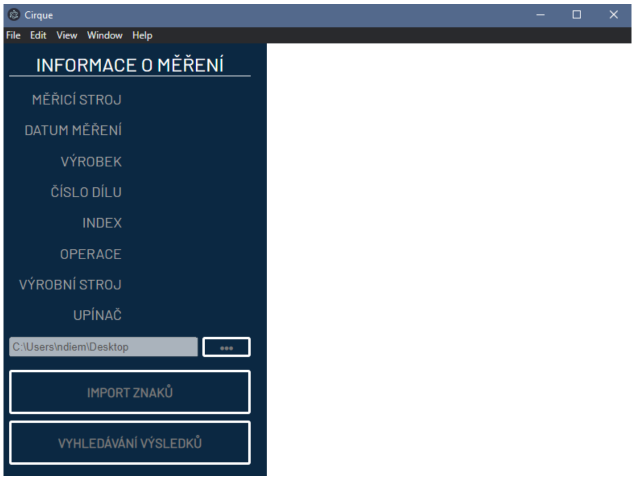
Upon the first initialization of CIRQUE, the application is prompting the user to wait while it is launching a Matlab server (analytic server for the user). The user interface follows the aforementioned visualization rule of interaction – it informs the user about its business by showing a revolving gear and a statement.



C:\Users\ndiem\Desktop

Čekám na spuštění analytického serveru...

Figure 5.6: Initialization of the environment



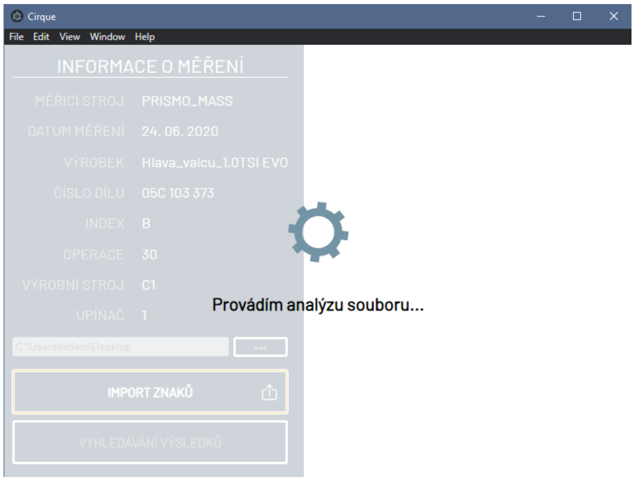
C:\Users\ndiem\Desktop

Figure 5.7: Initial launch of CIRQUE

After CIRQUE finishes setting the Matlab server (analytic server), the gear and the waiting statement fade into the background. The user is presented with the 'Information about the measurement' column on the left (intentionally left empty) and a blank space on the right. The user is now also able to interact with the buttons 'Import characteristics', 'Find results', and three-dot-button, which serves as a directory setting button.

The button 'Import characteristics' presents the user with a file directory, where the user is prompted to choose a file that contains a '.dfq' extension (a standard used by Škoda).

2. Analysis



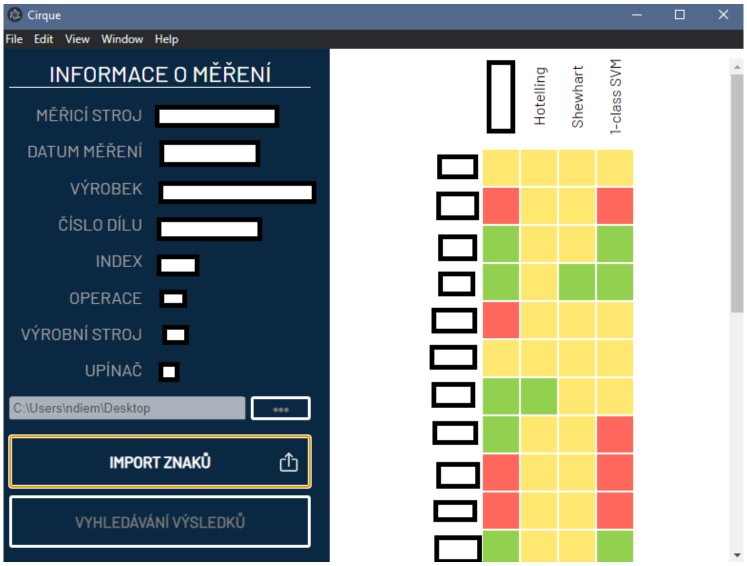
MĚŘICÍ STROJ PRISMO_MASS
DATUM MĚŘENÍ 24. 06. 2020
VÝROBEK Hlava_valcu_1,0TSI EVO
ČÍSLO DÍLU 05C 103 373
INDEX B
OPERACE 30
VÝROBNÍ STROJ C1
UPÍNAČ 1

C:\Users\ndiem\Desktop

Provádím analýzu souboru...

Figure 5.8: Analysis of the chosen characteristics

After the user chooses a '.dfq' file to run analysis on, Cirque executes the analysis on the chosen file and the user is again prompted to wait by a statement and a revolving gear.



MĚŘICÍ STROJ [redacted]
DATUM MĚŘENÍ [redacted]
VÝROBEK [redacted]
ČÍSLO DÍLU [redacted]
INDEX [redacted]
OPERACE [redacted]
VÝROBNÍ STROJ [redacted]
UPÍNAČ [redacted]

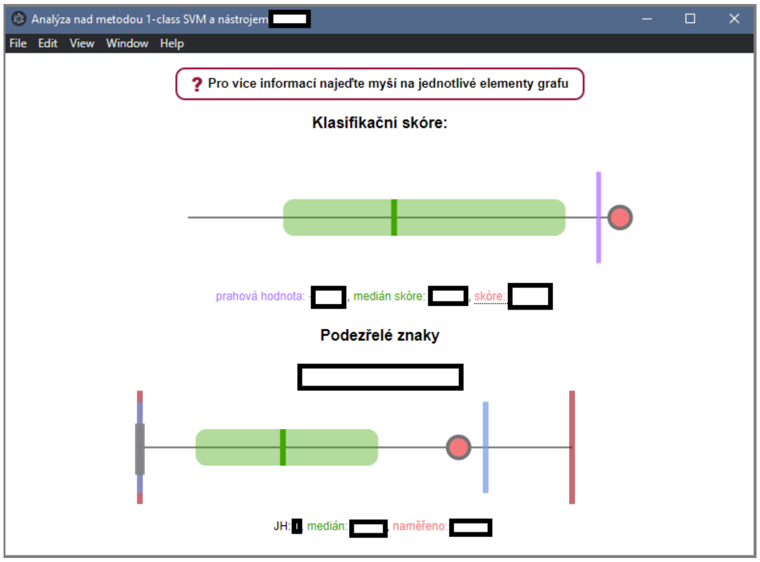
C:\Users\ndiem\Desktop

	[redacted]	Hotelling	Shewhart	1-class SVM
[redacted]	🟨	🟨	🟨	🟨
[redacted]	🟥	🟨	🟨	🟥
[redacted]	🟩	🟨	🟨	🟩
[redacted]	🟩	🟨	🟩	🟩
[redacted]	🟥	🟨	🟨	🟨
[redacted]	🟨	🟨	🟨	🟨
[redacted]	🟩	🟩	🟨	🟨
[redacted]	🟩	🟨	🟨	🟥
[redacted]	🟥	🟨	🟨	🟥
[redacted]	🟥	🟨	🟨	🟥
[redacted]	🟩	🟨	🟨	🟩

Figure 5.9: Final color-coded matrix

This particular color-coded matrix does not respond to the real results. The picture was meant just to display colors and the user interface.

Afterward, the method's decisions are shown in the color-coded matrix. These results are immediately saved to the predefined file directory shown above the 'Import characteristics' button. If the user ever wishes to load the analysis results again without going through the analysis, they can do so through the 'Find results' button. In case the 'dfq' file was misplaced or lost, the results of the analysis could still be kept. The results are saved in a .json file, which, if loaded, presents the same color-coded matrix from the previous analysis of the file.



Analýza nad metodou 1-class SVM a nástrojem [redacted]

? Pro více informací najeďte myší na jednotlivé elementy grafu

Klasifikační skóre:

prahová hodnota: [redacted], medián skóre: [redacted], skóre: [redacted]

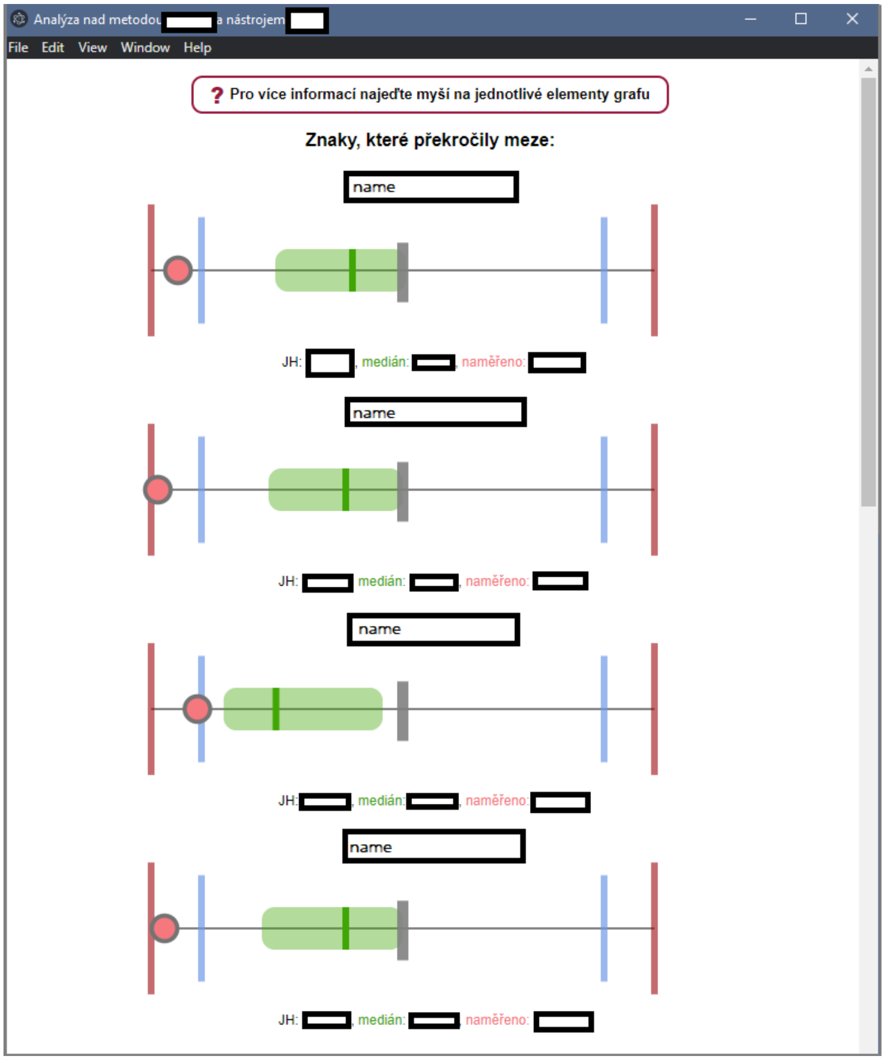
Podezřelé znaky

[redacted]

JH: [redacted] medián: [redacted] naměřeno: [redacted]

Figure 5.10: Analysis of nonconforming characteristics, OSVM

If the user wishes to look through the nonconforming characteristics, the red matrix window is clickable, presenting the nonconforming characteristic graph shown in Fig. 5.11 and Fig. 5.10. Compared to the design proposal of the OSVM graph, which only showed the classification score, the implemented version also shows the characteristics it deems suspicious. These characteristics are then presented in the same manner as the original graph so the user could interpret the reason behind the OSVM's decision better. As we see in the particular case of Fig. 5.10, the measured value of the characteristic is already nearing the regulation limits. Due to that, the OSVM decided that it was time to alarm the operator before the value exceeds any limits.



? Pro více informací najeďte myší na jednotlivé elementy grafu

Znaky, které překročily meze:

name

JH: [redacted], medián: [redacted], naměřeno: [redacted]

name

JH: [redacted], medián: [redacted], naměřeno: [redacted]

name

JH: [redacted], medián: [redacted], naměřeno: [redacted]

name

JH: [redacted], medián: [redacted], naměřeno: [redacted]

Figure 5.11: Analysis of nonconforming characteristics

Chapter 6

Conclusion

This thesis provides an overview of out-of-state detection and usage of machine learning on SPC. It introduces commonly used methods and related work to this field and presents research on issues that are commonly faced in the SPC. Although a lot of information could not be published, a very general and anonymized analysis of the data was conducted to provide the reader with the state of the data. Afterward, the implemented methods Shewhart control chart, Hotelling control chart, and OSVM were compared to the classification of the original method that is currently used in Škoda Auto. Subsequently, a series of experiments on a synthetically generated dataset was conducted to gain a better overview of the method's performance. The synthetic dataset was generated based on the characteristics of the real data; therefore, they are multivariate, multimodal, and heavily correlated.

The experiments showed that for data generated from such complex processes, OSVM performed the best most of the time. The univariate nature of the Shewhart control chart struggled with the multivariate aspect and even more so when the data dimension was raised in conjunction with further correlations between variables. Although the Hotelling control chart had all of its assumptions violated, its performance was satisfactory, in particular in the case of one variable shift where it outperformed even OSVM. A part of these experiments was organized into a paper submitted and admitted to the *13th International Conference on Soft Computing and Pattern Recognition* held in December 2021.

During the implementation and development of the methods suitable for the processes in Škoda Auto, an application was built for a more intuitive comparison between the methods. The application offers some useful features, such as intuitive visualization of the data chosen by the user, based on the implemented SPC methods.

Since this thesis is promoting the usage of machine learning in SPC, it continuously creates a parallel between SPC phases and the learning and testing aspect of machine learning. As this thesis is focused on phase I of SPC, which targets the retrospective analysis, it is analogous to the learning phase in a machine learning algorithm. In phase I, we are trying to adjust the method to create the right balance as we are trying to improve quality. In practice, we need to respect the existing processes as well, so a certain similarity between the new OSVM and the current Škoda method is needed. Although we lack the ground truth from the real dataset, we cannot expect Škoda to have the capacity to 'blindly' send every characteristic that the method deems nonconforming to the measurement center.

Fortunately, the visualization of the application seems to help with proving the effectivity of OSVM as it graphically shows the visualization of the measured value on a graph. The OSVM already deems the characteristic to be nonconforming when it nears the regulation limits; meanwhile, the original Škoda method does not yet notice since the characteristic has not exceeded the limits.

As for phase II of SPC that is parallel to the actual usage of the machine learning algorithm, we need to face several problems first. The modern processes and data frequently deal with the undocumented and unpredictable change of data structure, semantics, and infrastructure. This change is called a *data drift*, an inevitable result of the modern data architecture. As phase I of SPC focuses on the offline analysis, the current state of the OSVM is not prepared for the issue of data drift yet. One of the possibilities would be to employ online learning, automatically updating new models with newer data. Since many applications operate with real-time streaming data feeds, it could allow the method to learn directly from data streams. Another approach would include data drift recognition with model relearning to combat the problem.

Future Work

Although statistical process control is probably not the most popular research topic as of now, we believe it has immense potential, in combination with the developing field of artificial intelligence. Apart from the analysis itself, further work could be performed on the methods of how data is obtained and stored, how quickly it is accessible within the process and on improving the data structure in a way that would improve machine readability. Such improvements might allow us to analyze more properties, speed up the analysis or make the analysis more accurate. Lastly, we would like to generalize the usage of these methods outside the scope of the automobile industry and explore the possibilities of utilizing other machine learning algorithms for OOC detection.

Bibliography

[1] M. Macas, D. H. Nguyen, and C. Panuskova, “Support Vector Machines for Control of Multimodal Processes”, in International Conference on Soft Computing and Pattern Recognition, Springer, 2021, pp. 384–393.

[2] B. +. A. Conference. (2018). How much money could predictive analytics truly save your company?, [Online]. Available: https://biaconference.com/data/how-much-money-can-predictive-analytics-truly-save-your-company/ (visited on 05/12/2021).

[3] Metrology and Q. News. (2020). Artificial Intelligence Supports BMW Quality Assurance, [Online]. Available: https://metrology.news/artificial-intelligence-supports-bmw-quality-assurance/ (visited on 05/12/2021).

[4] F. W. Taylor, Scientific management. Routledge, 2004.

[5] D. C. Montgomery, Statistical quality control. Wiley Global Education, 2012.

[6] R. E. Barlow and T. Z. Irony, “Foundations of statistical quality control”, Lecture Notes-Monograph Series, pp. 99–112, 1992.

[7] W. E. Deming, “Lectures on statistical control of quality”, Nippon Kagaku Gijutsu Remmei, 1950.

[8] A. Mitra, Fundamentals of quality control and improvement. John Wiley & Sons, 2016.

[9] D. Garvin, “Competing on the eight dimensions of quality”, Harv. Bus. Rev., pp. 101–109, 1987.

[10] D. Ciampa, Total quality: a users’ guide for implementation. Addison Wesley Publishing Company, 1992.

[11] J. P. Womack and D. T. Jones, “Banish waste and create wealth in your corporation”, Recuperado de http://www. kvimis. co. in/sites/kvimis. co. in/files/ebook_attachments/James, 2003.

[12] Y. Tsim, V. Yeung, and E. T. Leung, “An adaptation to ISO 9001: 2000 for certified organisations”, Managerial Auditing Journal, 2002.

[13] G. Tennant, Six Sigma: SPC and TQM in manufacturing and services. Routledge, 2017.

[14] D. C. Montgomery, Introduction to statistical quality control. John Wiley & Sons, 2020.

[15] J. Eva and N. Darja, Pokročilejší metody statistické regulace procesu. Grada Publishing as, 2015.

[16] X. Fu, R.-f. Wang, and Z.-y. Dong, "Application of a Shewhart control chart to monitor clean ash during coal preparation", International Journal of Mineral Processing, vol. 158, pp. 45–54, 2017.

[17] M. Koutras, S. Bersimis, and P. Maravelakis, "Statistical process control using Shewhart control charts with supplementary runs rules", Methodology and Computing in Applied Probability, vol. 9, no. 2, pp. 207–224, 2007.

[18] A. Faraz and M. B. Moghadam, "Fuzzy control chart a better alternative for Shewhart average chart", Quality & Quantity, vol. 41, no. 3, pp. 375–385, 2007.

[19] Statistical Quality Control Handbook. Western Electric Co., 1956.

[20] W. Hachicha and A. Ghorbel, "A survey of control-chart pattern-recognition literature (1991–2010) based on a new conceptual classification scheme", Computers & Industrial Engineering, vol. 63, no. 1, pp. 204–222, 2012.

[21] Y. Wang, Y. Si, B. Huang, and Z. Lou, "Survey on the theoretical research and engineering applications of multivariate statistics process monitoring algorithms: 2008–2017", The Canadian Journal of Chemical Engineering, vol. 96, no. 10, pp. 2073–2085, 2018.

[22] S. J. Qin, Process data analytics in the era of big data, 2014.

[23] T. Chen, J. Morris, and E. Martin, "Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring", Journal of the Royal Statistical Society: Series C vol. 55, no. 5, pp. 699–715, 2006.

[24] W. Hwang, G. Runger, and E. Tuv, "Multivariate statistical process control with artificial contrasts", IIE transactions, vol. 39, no. 6, pp. 659–669, 2007.

[25] M. M. Moya and D. R. Hush, "Network constraints and multi-objective optimization for one-class classification", Neural networks, vol. 9, no. 3, pp. 463–474, 1996.

[26] S. S. Khan and M. G. Madden, "One-class classification: taxonomy of study and review of techniques", The Knowledge Engineering Review, vol. 29, no. 3, pp. 345–374, 2014.

[27] A. Apsemidis, S. Psarakis, and J. M. Moguerza, "A review of machine learning kernel methods in statistical process monitoring", Computers & Industrial Engineering, vol. 142, p. 106 376, 2020.

[28] R. Sun and F. Tsung, "A kernel-distance-based multivariate control chart using support vector methods", International Journal of Production Research, vol. 41, no. 13, pp. 2975–2989, 2003.

[29] S. He, W. Jiang, and H. Deng, "A distance-based control chart for monitoring multivariate processes using support vector machines", Annals of Operations Research, vol. 263, no. 1, pp. 191–207, 2018.

[30] H. Deng, G. Runger, and E. Tuv, "System monitoring with real-time contrasts", Journal of Quality Technology, vol. 44, no. 1, pp. 9–27, 2012.

[31] L.-J. Kao and C. C. Chiu, "Application of integrated recurrent neural network with multivariate adaptive regression splines on SPC-EPC process", Journal of Manufacturing Systems, vol. 57, pp. 109–118, 2020.

[32] S. B. Kim, T. Sukchotrat, and S.-K. Park, "A nonparametric fault isolation approach through one-class classification algorithms", IIE Transactions, vol. 43, no. 7, pp. 505–517, 2011.

[33] M. Weese, W. Martinez, F. M. Megahed, and L. A. Jones-Farmer, "Statistical learning methods applied to process monitoring: An overview and perspective", Journal of Quality Technology, vol. 48, no. 1, pp. 4–24, 2016.

[34] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation", ACM computing surveys (CSUR), vol. 46, no. 4, pp. 1–37, 2014.

[35] T. Perdikis and S. Psarakis, "A survey on multivariate adaptive control charts: Recent developments and extensions", Quality and Reliability Engineering International, vol. 35, no. 5, pp. 1342–1362, 2019.

[36] S. Cuentas, R. Peñabaena-Niebles, and E. Garcia, "Support vector machine in statistical process monitoring: a methodological and analytical review", The International Journal of Advanced Manufacturi vol. 91, no. 1, pp. 485–500, 2017.

[37] F. A. P. Peres and F. S. Fogliatto, "Variable selection methods in multivariate statistical process control: A systematic literature review", Computers & Industrial Engineering, vol. 115, pp. 603–619, 2018.

[38] M. Fuentes-García, G. Maciá-Fernández, and J. Camacho, "Evaluation of diagnosis methods in PCA-based Multivariate Statistical Process Control", Chemometrics and Intelligent Laboratory Systems, vol. 172, pp. 194–210, 2018.

[39] H. Hotelling, "Multivariate quality control", Techniques of statistical analysis, 1947.

[40] N. D. Tracy, J. C. Young, and R. L. Mason, "Multivariate Control Charts for Individual Observations", Journal of Quality Technology, vol. 24, no. 2, pp. 88–95, 1992. DOI: 10.1080/00224065.1992.12015232. eprint: https://doi.org/10.1080/00224065.1992.12015232. [Online]. Available: https://doi.org/10.1080/00224065.1992.12015232.

[41] V. Vapnik, The nature of statistical learning theory. Springer science & business media, 2013.

[42] L. Wang, Support vector machines: theory and applications. Springer Science & Business Media, 2005, vol. 177.

[43] G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning. Springer, 2013, vol. 112.

[44] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers", in Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

[45] O. Y. Rodionova, P. Oliveri, and A. L. Pomerantsev, "Rigorous and compliant approaches to one-class classification", Chemometrics and Intelligent Laboratory Systems, vol. 159, pp. 89–96, 2016, ISSN: 0169-7439. DOI: https://doi.org/10.1016/j.chemolab.2016.10.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743916302799.

[46] Z. Noumir, P. Honeine, and C. Richard, "On simple one-class classification methods", in 2012 IEEE International Symposium on Information Theory Proceedings, IEEE, 2012, pp. 2022–2026.

[47] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification", in Irish conference on artificial intelligence and cognitive science, Springer, 2009, pp. 188–197.

[48] D. M. Tax and R. P. Duin, "Support vector data description", Machine learning, vol. 54, no. 1, pp. 45–66, 2004.

[49] J. Colin and M. Vanhoucke, "Developing a framework for statistical process control approaches in project management", International Journal of Project Management, vol. 33, no. 6, pp. 1289–1300, 2015.