Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Sciences

Online process monitoring using a multivariate CUSUM approach with winsorization

A thesis presented by

María Fernanda Márquez Alderete

Submitted to the

School of Engineering and Sciences

in partial fulfillment of the requirements for the degree of

Master of Science

In

Manufacturing Systems

Monterrey Nuevo León, June 12th, 2020

To *José*; for your love, patience, and encouragement.

Acknowledgements

To God for his care and protection.

To my family for supporting me always.

To my advisors for guiding me during this period.

To Tecnológico de Monterrey for supporting my tuition.

To CONACyT with the support for living.

Online process monitoring using a multivariate CUSUM approach with winsorization

by

María Fernanda Márquez Alderete

Abstract

With the development of Industry 4.0 (I4.0), companies are transforming the way products are designed, manufactured and distributed. The application of new technologies in production and data acquisition exacerbates the need to foster quantitative approaches in the quality management of manufactured products, such as statistical process monitoring (SPM). A measuring system machine for evaluating die-casted workpieces was designed following the previous trend. This machine already applies part of the theoretical concepts of I4.0. The presented thesis complements the application of I4.0 concepts to the device, by using SPM methods, specifically, a multivariate CUSUM to assess small and sustained shifts; where winsorizing was used to create robustness over isolated changes that can be detected using complementing Shewhart-type charts. Additionally, an online dashboard was created to display the plotting statistics in real-time.

List of Figures

List of tables

Content Table

Chapter 1. Introduction

When a product is manufactured, it is usually intended to meet or exceed customer expectations. The process should be capable of operating with little variability around the nominal dimensions of the product produced in order to achieve these requirements. Many statistical methodologies have been developed to assure the quality of products; a central part of these methodologies is statistical process control/monitoring (SPC/SPM).

Although quality control and improvement methods are essentials components of our society, corresponding systematic theory and methods were introduced barely last century. In 1907, the Western Electric Company was reorganized by AT&T for inspection and testing of different products. By 1924 the inspection department of the Western Electric Company had more than 5000 members, including Joseph M. Juran. In 1925, a new Inspection Department was created in the newly formed Bell

Telephone Laboratories, in which several important statistical concepts and terminologies were created. For example, this is where Walter Shewhart first developed the concept of a control chart as a strategy to reduce the cost of quality assessment [1].

The US involvement in World War II (WWII) resulted in a rapid expansion of the manufacturing industry and, to assure the quality of the manufactured, goods training programs in the quality inspection were established. In the 1940s, research groups in quality control were established across the country. Later in 1947, the Hotelling's T^2 statistic for multivariate processes was proposed by Harold Hotelling [1].

Shewhart control charts are handy during early stages of a process monitoring, where the process is likely to be out-of-control (OC) due to several an isolated assignable causes of variation. Shewhart charts assess sampled observations individually, ignoring any information given by the entire sequence of points. This characteristic makes the Shewhart chart suitable when dealing with significant changes, but unsuitable for small sustained shifts (usually 1.5 standard deviations or less). Alternatives were created to avoid the problem of detecting small shifts. Two effective alternatives are the cumulative sum (CUSUM) control chart and the exponentially weighted moving average (EWMA) control chart [2].

With the development of Industry 4.0, industries are transforming the way products are designed, manufactured and distributed. The application of new technologies in production and data acquisition should be able to merge with basic and new methodologies of SPM, to create a real-time process control.

1.1 Background

In 2017 a prototype measurement machine was designed at ITESM [3, 4]. The machine tells the operator if a previously die casting workpiece is within tolerance, displaying the shape of the workpiece with the 19 measured points, shown in Figure

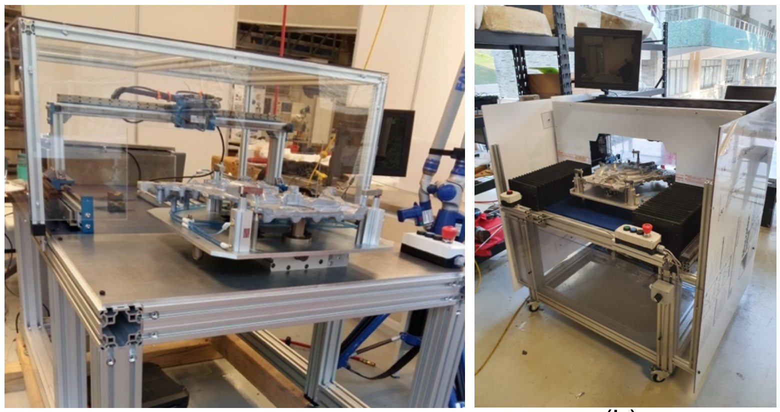
1, specifying which of the points are within tolerance and which ones are not. This function is possible with the use of a laser sensor that measures 19 points. Additional to this, a Bosch XDK sensor captures environmental temperature, humidity, pressure, and the maximum acceleration reached by the gantry. Also, an accelerometer located with the laser sensor records the maximum acceleration of the gantry for every measured point.

Figure 1. Die-casting workpiece

The machine was built to test novel features in the field, and a second version was built by modifying the original design; both versions are presented in Figure 2. The first version created by Castro and Guamán [3, 4], consisted of a moving plate that holds the workpiece with three clamps and four guiding pins. A piston moves the plate to its position where three kinematic couplings position the plate so that the laser can measure the piece. The machine was resting over wooden tables.

For the second version, the piston was removed, and the moving plate was replaced by a fixed plate. The gantry guides and the gantry itself in this new version were protected by bellow covers, and the machine was held over wheels.

After thousands of cycles, particular components of the machine failed, Specifically: the positioning and locating pins got loose, and the linear motors could not track their positions because of debris. While not catastrophic, these problems caused delays and motivated the development of a procedure to detect early changes to anticipate a major fault partially. Figure 3 presents how data from the final stage of the measurement process was accquired during the first test of the prototype in an industrial environment; each sample represents the overall error of the laser measured, hence, the lower the distance from the origin, the better. From sample 600 onwards, there are some points that are far from the the ± 0.40 mm limit.



(a) (b)

Figure 2. (a) First version. (b) Second and current version

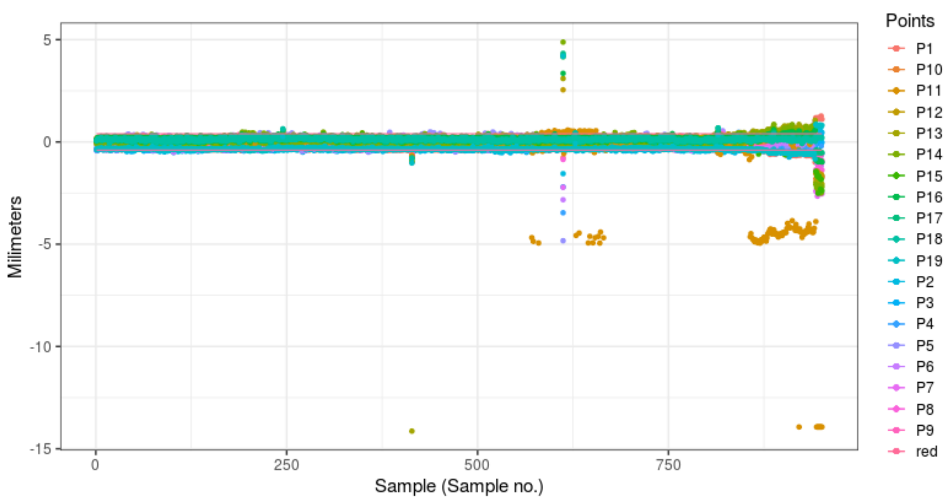


Figure 3. Raw data of the 19 points

From the literature review presented in Chapter 2, several multivariate CUSUM methods were analyzed. Specifically, Healy's CUSUM [5] was found helpful to address specific out-of-control situations. However, because of the nature of the industrial environment, external noise variables increases the risk of a false alarm, which in turn hinders the usefulness of the monitoring approach to detect sustained shifts. It was necessary to make the monitoring chart robust to isolated perturbations without sacrificing its ability to detect sustained changes in order to address this issue.

Due to the previous conditions, extreme values can be detected by the monitoring scheme. To avoid considering situations like abrupt movements as false positives, limiting the change rate of those values can be useful. This procedure is called winsorization and is thoroughly explained in chapter 2 and 4.

The average run length (ARL) is a metric to characterize the control chart, and it is defined as the average number of samples that must be plotted before a point indicates an out-of-control condition [2]. This metric might be used to evaluate the chart's performance when compared to an expected ARL value.

1.2 Problem Statement and Context

Since the introduction of I4.0, industries have been moving toward a digitalized environment. The first part of this process towards digitalization consisted of data generation from manufacture and machine performance. However, data is worthless if it is not analyzed to create knowledge that foster informed decision making.

Processing data to address the newly implemented I4.0 processes might result in savings related the quality costs [2]. There are quality costs related with the inspection and testing of all materials and products, not analyzing data coming from the inspection might not affect the process performance negatively; however, by not analyzing data, more significant benefits may be lost. The previous enhancements may vary depending on the industry in context, and these can go from preventing maintenance to time/costs savings.

As machines become more capable of obtaining vast amounts of data, the analysis of this data becomes increasingly difficult. This situation is where data science comes as an essential part of the fourth industrial revolution, where data dimensionality, speed, complexity and size present new challenges in knowledge creation. In this research, part of the theoretical concepts of I4.0 is implemented in a machine that already applies other I4.0 concepts [3, 4].

Machines working on an industrial environment tend to present failures and problems during a regular work shift. For this case, a laser measurement machine prototype has presented some problems working on an industrial environment, although these problems compromise the measurement capabilities of the machine, fixing them was not very complicated. Fortunately, these problems did not cause any damages in the following processes, but if these problems are not early detected, they could cause a delay in production and a loss of time and money. For this reason, it is essential to monitor not only the production process but also the machine performance.

The available data from the machine is related to its measuring performance. Hence, by analyzing this data, the overall performance of the machine can be estimated and forecasted.

Events that directly affect the machine, such as impacts or loose components, are reflected in the output data from the machine; conditions such as shifts on the variance and mean could have a physical representation as one of the events previously mentioned. However, it is not possible to determine the origin of these anomalies in the data, as they might be a consequence of changes in the manufacturing process.

Several control charts with different features and characteristics are available for data scientists to use, many of them with particular conditions for them to be used. For this particular case, subtle and sustained shifts are desired to be detected; therefore, the CUSUM control chart is considered to be adequate. Nevertheless, punctual shifts are also significant; for this reason, a Hotelling T^2 will also be analyzed.

This thesis deals with the problem of designing a multivariate CUSUM monitoring scheme to detect sustained changes while being robust to isolated perturbations and with the use of I4.0 technologies to monitor and report the machine's performance.

1.3 Research Questions

- How can process data can be used to forecast machine performance?
- How can reports be produced promptly?
- What is the performance difference of the winsorized in-control chart in comparison with Healy's original proposal, for isolated changes?
- What is the performance difference of the winsorized out-of-control chart in comparison with Healy's original proposal?

1.4 Hypothesis

- Industry 4.0 tools can be used to facilitate monitoring and timely reporting of potential failures
- The proposed SPM method will enhance the error detection capabilities of the prototype.
- The effect of isolated changes will be smaller when the chart is winsorized. The in-control average run length will be closer to the expected performance on the winsorized chart compared to the chart without winsorization.
- The winsorization will not change significantly the ability to detect sustained changes compared to Healy's original control chart.

1.5 Objective

The overall goal of this work is to implement SPC to the data obtained by the machine and to display of the resulting information in the form of an online dashboard and make data available for the future development of a digital twin.

This work proposes to:

1. Measure the in-control performance of the proposed approach in the presence of isolated changes and compare with the performance of Healy's original chart.
2. Measure the out-of-control performance of the proposed approach, over different sustained shifts, and compare results with the corresponding performance or Healy's original chart.
3. Code a real-time application based on the previous algorithms.
4. Design a dashboard and deploy it as a web App.

1.6 Scope

The solution includes the implementation and use of SPM methods to the data obtained to detect errors in the process and the machine before they become problematic in order to avoid line stopping and machine failures. Also, the experimentation with winsorization to evaluate different scenarios. The charts will be displayed in a friendly user dashboard. The display of the data from the newest version of the machine is out of the scope of this project; as it is not available in the online database. However, statistical analysis was done for both the previous and new version of the machine.

1.7 Research Contribution

The main contributions of this work are the application of multivariate control charts for practical and industrial use and measuring the effect of winsorization on Healy's multivariate CUSUM. All these combined as a system that will make information available on a friendly user manner using free access and low-cost tools. All this work guided to the creation of a digital twin for the measurement machine as the same approach used for developing the dashboard might be well suited for the digital twin to have a closed-loop feedback system.

1.8 Thesis Organization

This work is laid out in six chapters, including this one as an introduction. Chapter 2 deals with the literature review to study concepts and related works. The application of the methods studied in Chapter 2 and the construction of control charts is developed in Chapter 3. Chapter 4 presents an analysis of how winsorization will affect the detection of shifts. As for Chapter 5, the control charts created in Chapter 3 are shown in an online dashboard. Finally, Chapter 6 presents conclusions and future work.

1.9 Used Scripts

All the codes used for this thesis are available in this repository: https://github.com/MariaMarquezA/SPC_Dashboard. The part referring to the request of data from the Hosting service was omitted for security reasons.

Chapter 2. Literature Review

In this chapter, the literature reviewed to create this work is presented. In section 2.1, the I4.0 state-of-the-art is shown, focusing on the topics with the most substantial relation to this work. In section 2.2, a review on multivariate process control is also presented, where the methods used with the proposed approach are thoroughly explained. Finally, in section 2.3, brief research on dashboards is presented.

2.1. Industry 4.0

After the first Industrial Revolution, manufacturing methods have changed radically. From water and steam-powered machines to electrical and digital automated production. Manufacturing processes have become more sophisticated, automatic and sustainable, allowing people to operate machines simply and efficiently [6], with the integration between manufacturing operations systems and Information and Communication Technologies (ICT) – Internet of Things – forming the so-called Cyber-Physical Systems (CPS) [7]. With the development of these technologies, a new concept, Industry 4.0 (I4.0), was introduced in Germany during the Hannover Fair event in 2011, which symbolizes the beginning of the 4th Industrial Revolution [6, 8].

The trend of manufacturing process digitalization which characterizes I4.0 is based in 8 enabling technologies [9], shown in Figure 4: Internet of Things (IoT), Cyber-Physical Systems (CPS), Additive manufacturing, Cobots, Augmented reality, Machine learning, Cloud computing, Big Data, and Real-time optimization. This thesis will focus aboard the last two technologies.

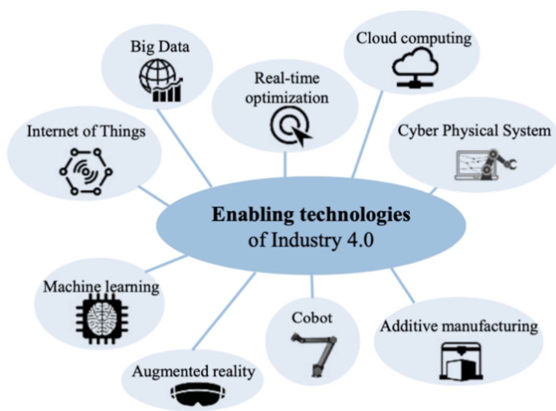


Figure 4. Enabling technologies of I4.0 [9]

2.1.1 Big Data

The concept of big data has been used since the earliest days of computing; originally meant for the volume of data that could not be processed by traditional database methods and tools [10]. Big data analytics is the applications of advanced analytic techniques, such as data mining, statistical analysis, or predictive analytics on big datasets, creating one of the most profound trends in business intelligence [11]. The enormous amount of data creates new opportunities of improvement; therefore many organizations started to develop and enhance their big-data analytics capabilities to gain a better understanding and uncover vital points hidden in the data [12]. The study of big-data is continuously evolving, but the main attributes are now enclosed into the "5V" concept, which consists of volume, velocity, variety, verification, and value [13]. The overall process of extracting insights from a dataset is divided by Gandomi and Haider [14] into five stages, which are grouped into two main sub-processes: data management and analytics. Data management includes the stages of acquisition and recording; extraction, cleaning and annotation; and integration, aggregation, and representation. As for the Analytics sub-group, the included stages are modeling and analysis; and interpretation. These stages are graphically represented in Figure 5.

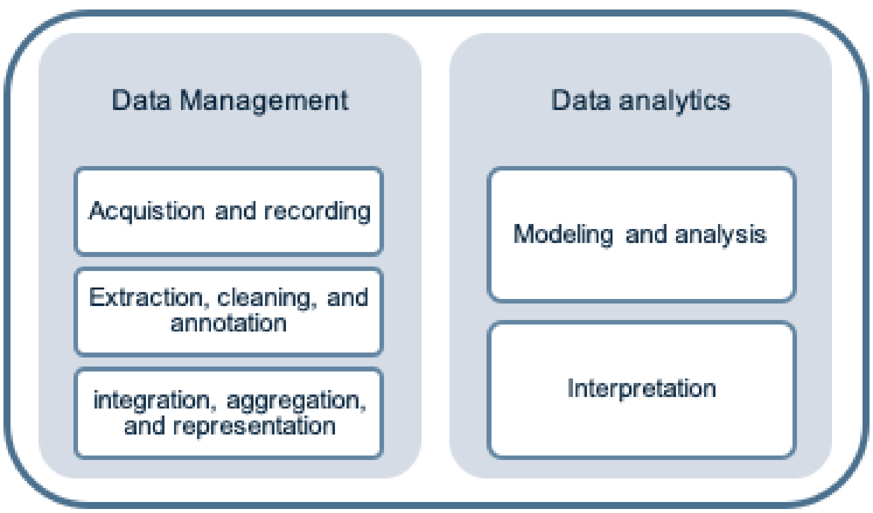


Figure 5. Processes for extracting insights from big data [14].

2.1.2 Real-time optimization

The objective of a real-time optimization is to respond to an essential quality characteristic in a continuously evolving environment, within a specific time frame, using limited resources and information which are usually incomplete or uncertain. Real-time decision problems typically include stochastic and dynamic components. In this context, system correctness depends on how appropriate the response is and also on how much time it took to get that response. Therefore a balance between the quality of the response and its computation time must be achieved. A real-time decision system (RTDS) must be able to support, assist or replace human operators for real-time decision making [15].

2.1.3 Digital Twin and Digital Twin Shop-floor

The concept of Digital Twin (DT) was introduced by Grieves [16] in 2003 at the University of Michigan during a Product Lifecycle Management course. The DT concept contains three main parts: the physical products in real space, virtual products in virtual space, and the connections of data and information that ties both.

Based on the DT concept, the Digital Twin Shop-floor (DTS) is proposed [17], making the physical and virtual parts optimized and combined effectively. Data from the physical and virtual sides, as well as the fused data, are provided to manage the production. DTS is conformed by four components, Physical Shop-floor (PS), Virtual Shop-floor (VS), Shop-floor Service System (SSS), and the Shop-floor Digital Twin Data (SDTD).

PS consists on a series of entities, including humans, machines and materials, existing in a physical space; from this, the VS is formed by models built-in multiple dimensions, including geometry, physics, behavior and rule. SSS is an integrated service platform, which includes the functions of Enterprise Information System (EIS), computer-aided tools, models and algorithms.

SDTD includes PS, VS and SSS data, the fused data of the three parts, and the existing methods for modeling, optimizing, and predicting. The construction of SDTD begins with the conversion of all data into a unified form, and then the data is cleaned by detecting and removing dirty data or missing values. Finally, to achieve consistent and comprehensive interpretations, virtual and physical data are converged to form fused data. SDTD is optimized iteratively, and historical data are updated and expanded with the real-time data, while real-time data can be tested and corrected with the historical data. Another important aspect is that physical data can be evaluated and simulated by virtual data, and virtual data can be compared with real data to confirm the accuracy [18].

2.2 Multivariate Process Monitoring

SPM techniques are widely used in the industry to monitor and improve production processes. Various control charts have been developed to monitor process parameters and detect out-of-control conditions [19].

Since the design of the first control chart by Shewhart in 1924, many control schemes were introduced in the literature. The CUSUM chart of Page [20], the EWMA scheme of Roberts [21], Healy's CUSUM [5] or Castagliola, Celano and Fichera CUSUM-S^2 [22]. Bodnar and Schmid [23] analyzed control charts for the mean of a multivariate Gaussian process using the generalized likelihood ratio approach and the sequential probability ratio test, proposing various types of CUSUM control charts and comparing them with several other control schemes available in the literature. Noorosana and Vaghefi [19] studied how autocorrelation can deteriorate the performance of MCUSUM control charts, using the ARL criterion. The proposed solution was based on a time series model. Zaman et al. [24] proposed an adaptive EWMA control chart based on the CUSUM accumulation error, for detecting imbalanced mixed range shifts in process dispersion.

Miekley et al. [25] applied MCUSUM control charts for the early detection of mastitis and lameness diseases in cows. The values of each input variable were pre-processed with wavelet filters or a multivariate vector autoregressive model to exclude biological trends and obtain independent observations; the residuals generated were transferred to both classic and self-starting MCUSUM. Haq, Munir and Khoo [26] proposed a dual MCUSUM (DMCSUM) and a mixed DMCUSUM (MDMCUSUM) chart for monitoring the mean of a multivariate normal process.

2.2.1 Hotelling T^2

The first work concerning multiple variables of interest was done by Harold Hotelling. In 1947, he applied his procedure to bombsight data during WWII. It is a direct analog of the univariate Shewhart $\bar{x}$ chart. For this work, the individual observations version is of interest, this means a subgroup $n = 1$, with m samples and p number of quality characteristics that are observed in each sample. A summary of Hoteling's work follows, as shown in [2]. Let $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ be the sample mean vector and covariance matrix, respectively. The Hotelling T^2 statistic for this case is presented in equation (1) [2].

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

With upper control limit (UCL) and lower control limit (LCL):

$$UCL = \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha,p,m-p} \tag{2}$$

$$LCL = 0$$

When the number of m samples is larger than 100, many practitioners use an approximate control limit, presented in equation (3) and (4); Tracy, Young and Mason [27] proposed limits for $n = 1$ based on a beta distribution. These limits are shown in equation (5).

$$UCL = \frac{p(m-1)}{m-p} F_{\alpha,p,m-p} \tag{3}$$

$$UCL = \mathcal{X}^2_{\alpha,p} \tag{4}$$

$$UCL = \frac{(m-1)^2}{m} \beta_{\alpha,p/2,(m-p-1)/2} \tag{5}$$

$$LCL = 0$$

2.2.2 CUSUM

The first CUSUM control chart was proposed by Page in 1954 [20]. Since then, many authors have contributed to the theory. In 1987 Healy [5] proposed a testing method for the shift in the mean of a multivariate normal, where the CUSUM procedure was reduced to a univariate normal CUSUM, this procedure is optimum when the in-control mean, and the out-of-control mean are known. By optimum, we mean that this chart is the fastest chart, in term of average run length, to detect a change to the out-of-control mean vector when it happens.

Assuming the multivariate data $\boldsymbol{x_n}$ comes from a multivariate normal distribution with an acceptable mean $\mu\mathbf{0}$ or a multivariate normal distribution with a non-acceptable mean $\mu\mathbf{1}$; being F_0 and F_1 multivariate normal distributions with mean vectors of $\mu\mathbf{0}$ and $\mu\mathbf{1}$ and a known common covariance matrix $\sum$. All vectors have p number of elements, and all matrices are $p \times p$, being p the number of quality characteristics that are observed in each sample. The CUSUM statistic proposed by Healy is presented in equation (6).

$$S_n = \max(S_{n-1} + \boldsymbol{a}'(\boldsymbol{x_n} - \boldsymbol{\mu_0}) - 0.5D, 0) > CL, \tag{6}$$

where

$$\boldsymbol{a}' = \frac{(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'\boldsymbol{\Sigma^{-1}}}{[(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'\boldsymbol{\Sigma^{-1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})]^{1/2}}, \tag{7}$$

and

$$D = \sqrt{(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'\boldsymbol{\Sigma^{-1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})}. \tag{8}$$

The expression $\boldsymbol{a}'(\boldsymbol{x_n} - \boldsymbol{\mu_0})$ has a standard univariate normal distribution when $\boldsymbol{x_n}$ has its mean equal to $\boldsymbol{\mu_0}$. When $\boldsymbol{x_n}$ has a mean equal to $\boldsymbol{\mu_1}$, $\boldsymbol{a}'(\boldsymbol{x_n} - \boldsymbol{\mu_0})$ has a univariate normal distribution with mean *D* variance 1. Therefore, for detecting a shift in the mean, the CUSUM procedure reduces to a univariate normal CUSUM procedure, which means all of the theory for calculating the average run length (ARL), Control Limit (CL) and the initial value of the CUSUM for a univariate normal CUSUM can be used for a multivariate CUSUM (MCUSUM).

Healy's procedure only looks in one particular direction from $\boldsymbol{\mu_0}$ to $\boldsymbol{\mu_1} = \boldsymbol{\mu_0} + \delta$, if changes are also expected in the opposite direction, this means from $\boldsymbol{\mu_0}$ to $\boldsymbol{\mu_1} = \boldsymbol{\mu_0} - \delta$, the author recommends creating another one-sided CUSUM; this way, two one-sided CUSUM will be able to detect changes in both directions. More univariate charts can be used for other pre-specified types of control changes the practitioners might be interested in detecting quickly.

The CUSUM chart starts at an initial state C_0^+, from this point, it may stay on the axis, or it may move to positive values. When a plotted point crosses the CL indicates that a shift has occurred and actions must be taken, and the CUSUM will be restarted. The sequence from the starting point to the first point out-of-control is called run. The number of observations appearing in the run is called run length. Sometimes signals will be generated even if no significant shift has occurred, and this is an example of a Type I error. On the order side, sometimes the chart might not detect a significant shift; this is a Type II error. The ARL, defined by Page [20] as a performance measure of a control chart, is used to balance both errors. Even though a run length is highly variable, the ARL is a good summary number of the general tendency.

Although the ARL does not guarantee false alarms is an easily interpreted and a well-defined measurement.

Because Healy's CUSUM can be treated as a univariate control chart, existing functions can be used to obtain a CL for a given a *D*, and ARL. Simulations can be used to corroborate results [28].

2.2.3 Winsorization

A fundamental assumption in variables charting is that, in the absence of special causes, the readings tend to be normally distributed. The effectiveness of the ARL depends on this assumption. When the data does not follow a normal distribution, the calculated ARL can be significantly different from the actual ARL. For example, when the natural distribution of the data tends to have outliers it creates a heavy tail; causing more false alarms [29].

To achieve a long-in-control ARL with a quick response to persistent shifts, 'robustifying' CUSUM charts was proposed to make the statistic less sensitive to isolated outliers, but still sufficient for detecting genuine shifts [29].

Lucas and Crosier [30], proposed a procedure in which a single suspected outlier is ignored, but two outliers in a row are considered to be an out-of-control signal.

A second procedure investigated by Hawkins [31] and also by Lucas and Crosier [30] consists of replacing any signal exceeding a pre-established threshold value with that value.

Winsorizing is an attractive manner of limiting de effect of outliers without not taking them into account for the control chart [29].

To evaluate and compare the different scenarios, Pignatiello and Runger [32] proposed comparing the ARL performance of the different CUSUM charts, using different standardized distances: 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0, from a target mean vector of $\boldsymbol{\mu 0} = \mathbf{0}$, changing *p* variables of the process. Here $\mathbf{0}$ is the origin in a vector space.

2.3 Data Visualization

A dashboard is a single screen display that shows essential information about a company and processes so that the whole panorama can be quickly understood by providing real-time information across all levels of the company. As digitization acquire more importance in manufacturing industries, dashboards are becoming more commonly used [33].

Nadoveza and Kiritsis [34], presented an idea based on identifying the information relevancy by capturing the manufacturing context of the users since the system collects large amounts of information is vital to show the appropriate information to the specific user at a particular time. Mazumdar et al. [35] also exposed the importance of providing the right visual means to explore and analyze datasets flexibly and efficiently to the right end-user.

Tokola et al. [33], published the results of a survey of manufacturing companies on preferred key performance indicators (KPIs) for dashboards and the use of them. Essential aspects to take into account are, display of the dashboard should be just a single screen display, where the most critical information should be on the top left and centre. Eckerson [36] emphasizes how different users must have different dashboards; for example, executives want to monitor the status of the company; managers analyze the performance-related information and shop-floor workers in production details. As conclusion [33] , designed three dashboard proposals, one for each of the three specific users, a strategy dashboard for executives, a tactical dashboard for production managers, and an operational dashboard for workers.

Chapter 3. Control Charts construction

The methods reviewed in Chapter 2 were applied to the data obtained by the measurement machine. Section 3.1 presents the acquisition of data and the first analysis with Hotelling T^2, to confirm and obtain an in-control dataset. On Section 3.2, Healy's CUSUM is implemented in all the dataset. Finally, in Section 3.3, the same analysis made on sections 3.1 and 3.2 is replicated for a dataset obtained from the new version of the machine.

3.1 Data Acquisition and cleansing with Hotelling T^2

Data acquired by the machine is stored on a remote database; further explanation about it is presented in Chapter 5. The data is then imported to R for further analysis. The data analyzed in this exercise is from the first version of the prototype.

Initially, it was assumed that the environmental parameters given by the Bosh sensor would show a correlation with the resulting data from the prototype, but an analysis was done using these parameters and resulted in no apparent correlation between the measurement error and the data gathered by the sensors. This result might not mean that there is no correlation between the mesurement error and the environment, but the lack resolution from the sensors might be the an indicator of this result. Appendix A includes the resulting matrix plots from the correlation analysis. From the previous analysis it was concluded no to used the environmental parameters for the analysis.

A first table was created with data considered to be in-control (IC). After eliminating ten measurements samples (for assignable causes, in this case, measurements that clearly show a wrong location of the piece) the resulting table is analyzed with equation (1), with a CL given by equation (5). With a confidence level of 99.73%, the resulting control chart is presented in Figure 6.

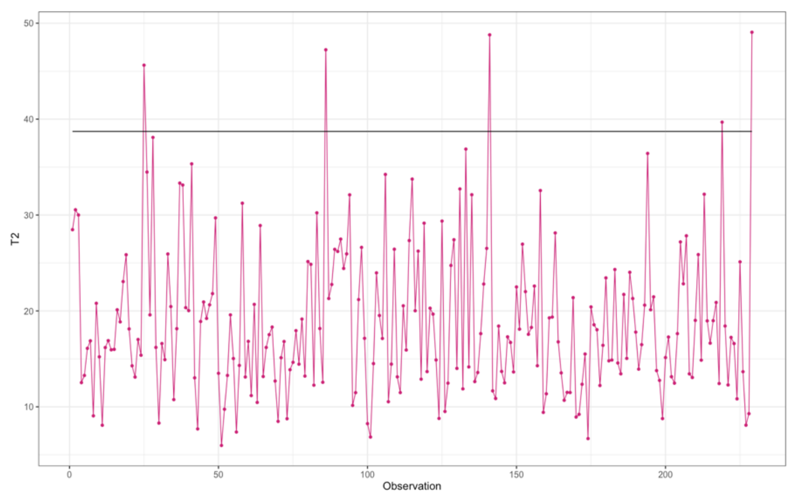


Figure 6. Hotelling T^2 chart for all 19 variables (1^{st} version)

Figure 6 shows five points which are out of control, with a CL of 38.71543. Consequently, each variable is analyzed with a univariate Shewhart control chart; the resulting 19 charts are presented in Appendix B. These charts demonstrate that before the 40^{th} observation at least 13 variables present irregular variations, also called violations. These violations indicate that the machine was not working in a steady-state; hence, the first 40 observations were eliminated from the in-control table, and the last observation, since it was from the next day. The Hotelling T^2 is analyzed again. The resulting chart is presented in Figure 7, where the CL changed to 38.28504 with 186 observations. Even though three points are still out of the CL, no special causes are detected, and observations are left as part of the IC dataset. If in fact, those three observations were due to special causes, the power of the chart or detection capability will be reduced.

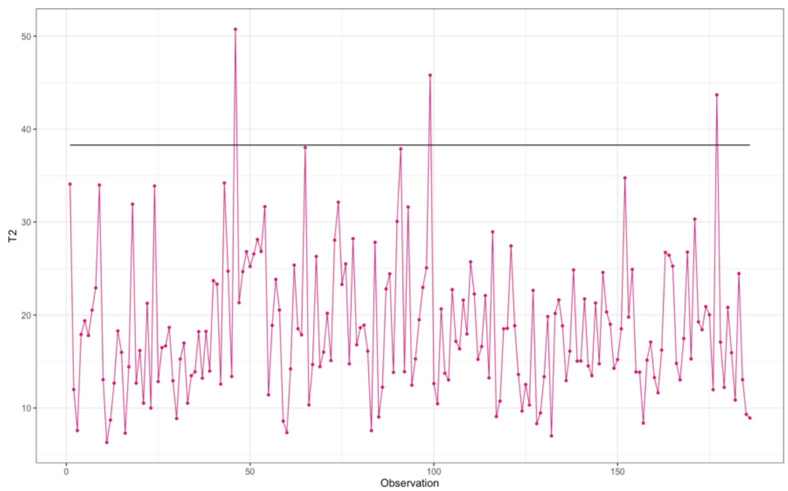


Figure 7. New Hotelling T^2 chart for all 19 variables (1^{st} version)

To evaluate the multivariate normality a Shapiro-Wilks test is implemented using the `mvnormtest` package in R. The result of the test gives a p-value of $1.8e^{-10}$ since this value is lower than 0.05 the test rejects the normality assumption, even tough as individual fifteen of the variables act as normal. A script was written to evaluate every possible combination of variables (2^{19} combinations) and detect which combinations resulted in a p-value equal or greater than 0.05. The code outputs a table with the combination number and the p-value. The combination number is the decimal representation of a binary number, each digit of the binary number represents the used variables; where the least significant digit corresponds to the first variable. The results yielded more than one thousand possible combinations, and Table 1 shows the first thirteen combinations with the most variables.

Table 1. Multivariate normality test p-value

p-value	Included variables
0.081475	1,2,5,6,7,8,11,12
0.058856	3,4,5,6,7,8,11,12
0.059664	1,2,3,4,6,8,12,13
0.064806	1,3,4,5,6,7,12,13
0.085901	1,2,5,6,7,8,11,16
0.089736	3,4,5,6,7,8,11,16
0.083509	1,2,5,6,8,11,12,16
0.050421	1,2,3,6,8,11,13,16
0.062248	1,2,5,6,10,11,13,16
0.052698	1,5,6,7,8,11,12,18
0.063778	4,5,6,7,8,11,12,18
0.053805	1,3,4,5,7,8,13,18
0.051684	1,5,6,7,11,12,15,18

Even though observations are not multivariate normal, Figure 8 shows how a T^2 chart would behave when introducing a small part of a known out-of-control situation, where it can be easily noticed the change between the IC data and the out-of-control data.

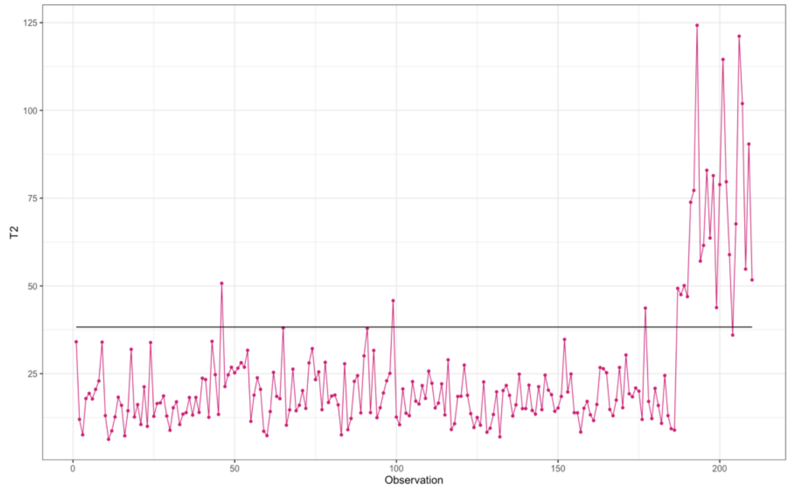


Figure 8. Hotelling T^2 test chart (1^{st} version)

3.2 Healy's CUSUM implementation

After obtaining the in-control dataset, the next step is to apply the CUSUM statistic. For coding the procedure, it is necessary to obtain the mean vector and the covariance matrix of the IC dataset, named $\boldsymbol{\mu 0}$ and Sigma respectively. The shift that is wanted to be detected, for which the chart is optimum, is given by $\boldsymbol{\mu_1} = \boldsymbol{\mu_0} + \delta$, being δ the shift expected to be detected. After obtaining these matrixes equations (6-8) can be applied for obtaining the CUSUM chart. The IC and OC datasets can be concatenated to obtain a chart that graphs both datasets.

For the CL the `R` function `getH` from the `CUSUMdesign` package was used. For validating the obtained results for both study cases, simulations were done. With an ARL of 370, for the case where the shift is $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} \pm 0.01$, the resulting CL is 4.868675, and when the shift is $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} \pm 0.02$ the obtained CL is 2.713268.

The first two study cases analyzed where $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} + 0.01$, and the second one where $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} + 0.02$, resulting graphs are shown in Figure 9 and Figure 10, respectively.

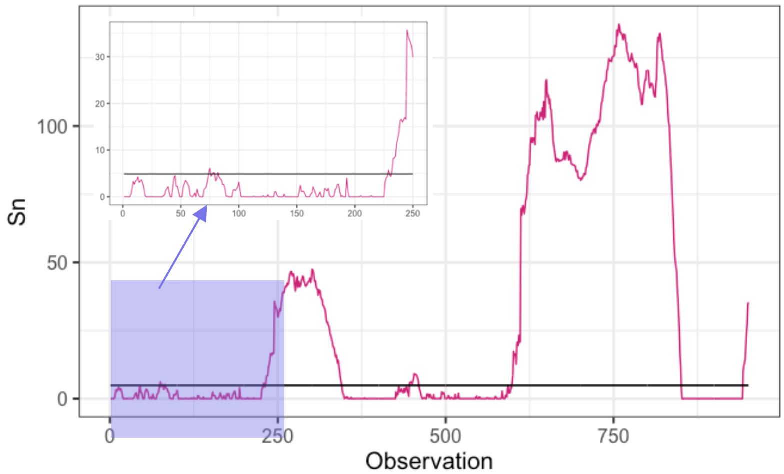


Figure 9. CUSUM chart calibrated to be optimum for a shift to μ0+0.01 in all variables (1st version)

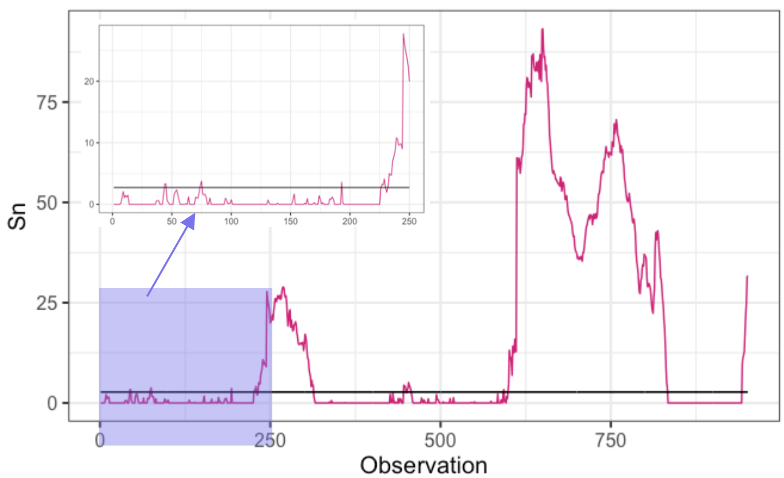


Figure 10. CUSUM chart calibrated to be optimum for a shift to μ0+0.02 in all variables (1st version)

As Healy recommends, two more one-sided CUSUM charts were designed, now for the negative direction, Figure 11 shows a shift of $\boldsymbol{\mu0} - 0.01$, while Figure 12 shows a shift of $\boldsymbol{\mu0} - 0.02$. The CL's for both charts remain the same as their counterpart.

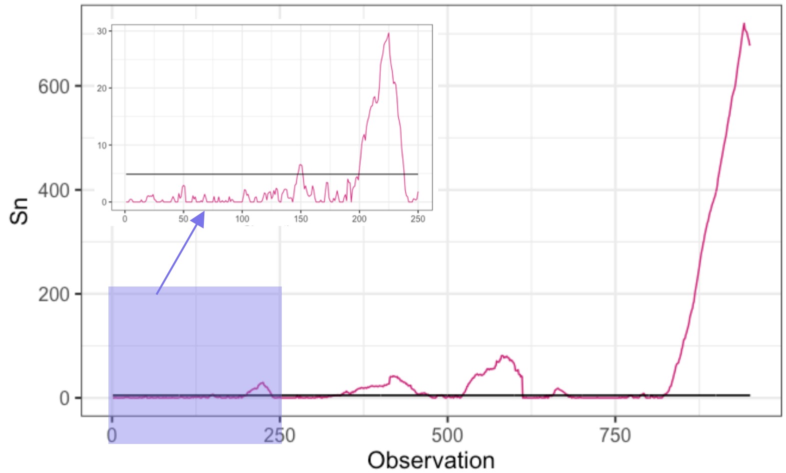


Figure 11. CUSUM chart calibrated to be optimum for a shift to μ0-0.01 in all variables (1st version)

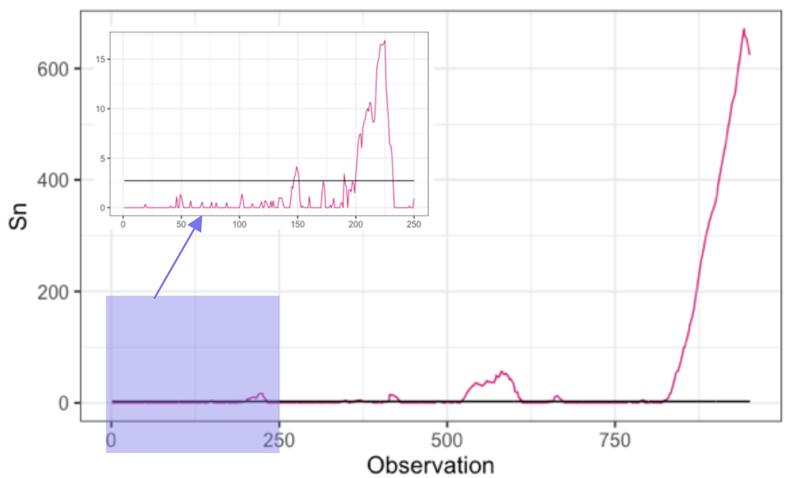
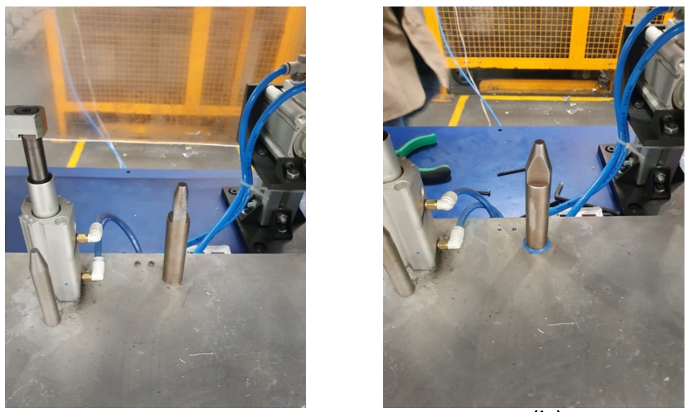


Figure 12. CUSUM chart calibrated to be optimum for a shift to μ0-0.02 in all variables (1st version)

From all the previous charts, it is noticeable that a shift in the measurements is detected. Even the negative side of CUSUM shows a considerable change that makes difficult to evaluate all observations; the four charts present a close-up of the first 250 observation including the 186 IC table, where It can be seen the first change; this matches with a problem detected while the first version of the machine was being tested in the plant for a complete working shift. The second day of operation, a guide pin was found loose, but it was incorrectly adjusted shown in Figure 13, causing some incorrect measurements of the workpieces, especially at the end of the work shift.



(a) (b)

Figure 13. Guide pin positioning (a)Incorrect and (b) Correct

A day later, a kinematic coupling got loose and fell as shown in Figure 14 (a), causing a significant problem in the measurements. Also, Figure 14 (b) shows a bad adjusting of the kinematic couplings, where they are not touching their supports.


(a) (b)

Figure 14. Kinematic coupling. (a) fallen and (b) incorrect position

For proving Healy's assumption that the expression $\boldsymbol{a}'(\boldsymbol{x_n} - \boldsymbol{\mu_0})$ has a standard univariate normal distribution when $\boldsymbol{x_n}$ has a mean equal to $\boldsymbol{\mu_0}$, the IC dataset was analyzed with the previous expression with a given a' from (6). The histogram is presented in

Figure 15, proving that the dataset behaves as Healy's proposed. The central limit theorem affects here, as the cumulative effect of the CUSUM normalized data, even though the original set was not normal.

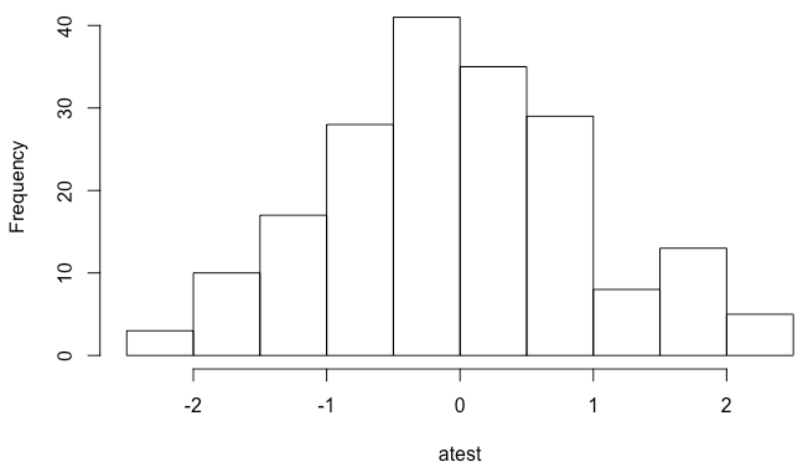


Figure 15. a'test histogram (1st version)

3.3 Second version model analysis

Since the machine was remodeled (changes explained on 1.1) the previous analysis is no longer accurate, this is because any change in the structure may affect the measurement and the relation between variables. This second version was created to prevent some of the failures presented on the first version and optimizing the takt time previously measured as 29 seconds on average. For this new version, the takt time was reduced to 22 seconds.

The analysis exposed on 3.1 and 3.2 was replicated for a new dataset obtained on another round of work shift testing with the second model of the prototype. During these test days, it was also reported a loose guide pin, which was readjusted.

After detecting and suppressing six observations for assignable causes, leaving 185 observations, the resulting Hotelling T^2 chart is presented in Figure 16 with a confidence level of 99.73% and a UCL of 38.25968.

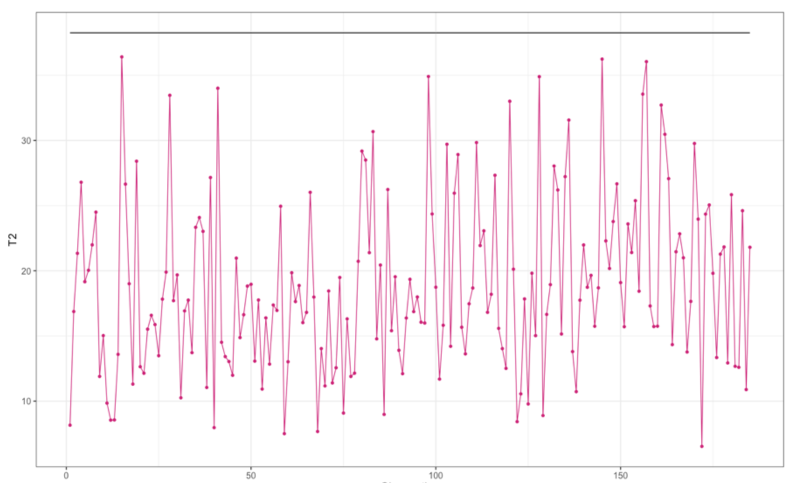


Figure 16. Hotelling T2 chart for all 19 variables (2nd version)

Even though Figure 16 those not present any points out of control, an analysis of every variable is done. The resulting charts are contained on Appendix C. From the 19 charts it can be seen an irregular behavior on the 50 first observations these are eliminated to avoid using data that was obtained when the machine was not yet working stable. Also, between observations 85 to 100 on eight variables, despite this, observations were left since not assignable causes were detected. The new Hotelling T^2 chart is presented in Figure 17, with a UCL of 37.38131 and 135 observations remaining.

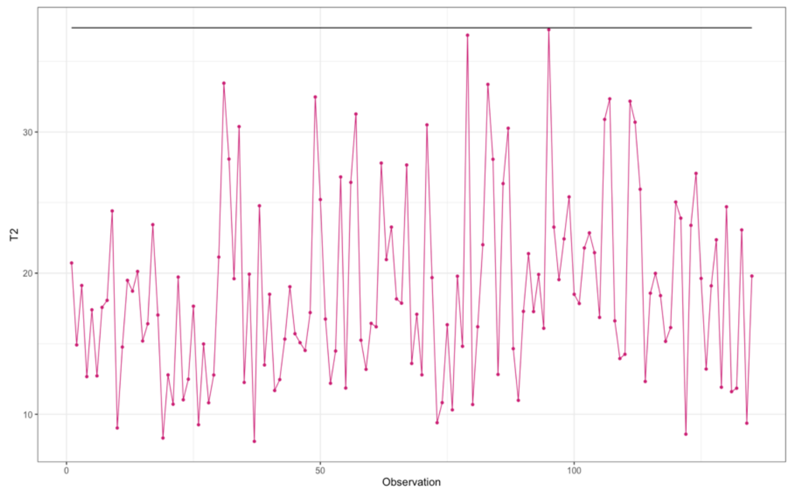


Figure 17. New Hotelling T2 chart for all 19 variables (2nd version)

After having the IC dataset, the CUSUM statistic can be applied using equations (6-8). Once again four scenarios were analyzed, for the positive shift $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} + 0.01$, and $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} + 0.02$, with a CL of 4.815231 and 2.674335 respectively. The resulting charts are presented in Figure 18 and Figure 19. For the negative shift $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} - 0.01$ and $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} - 0.02$, chart are shown in Figure 20 and Figure 21 respectively. The CL's for both charts remain the same as their counterpart.

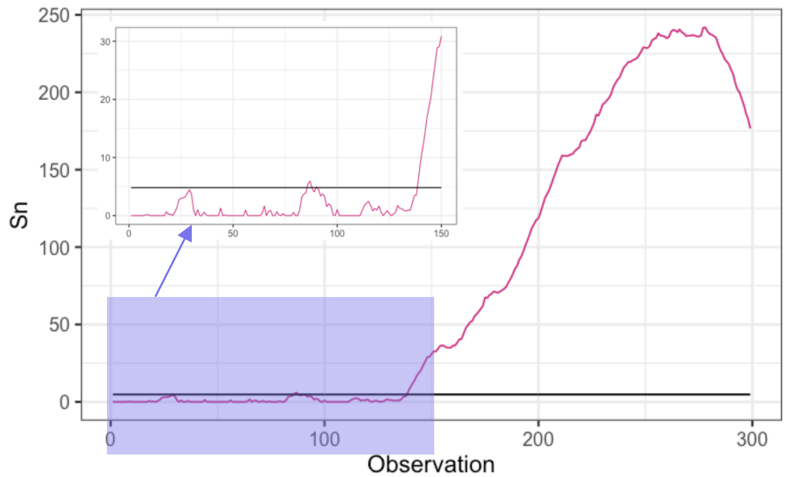


Figure 18. CUSUM chart calibrated to be optimum for a shift to μ0+0.01 in all variables. (2nd version)

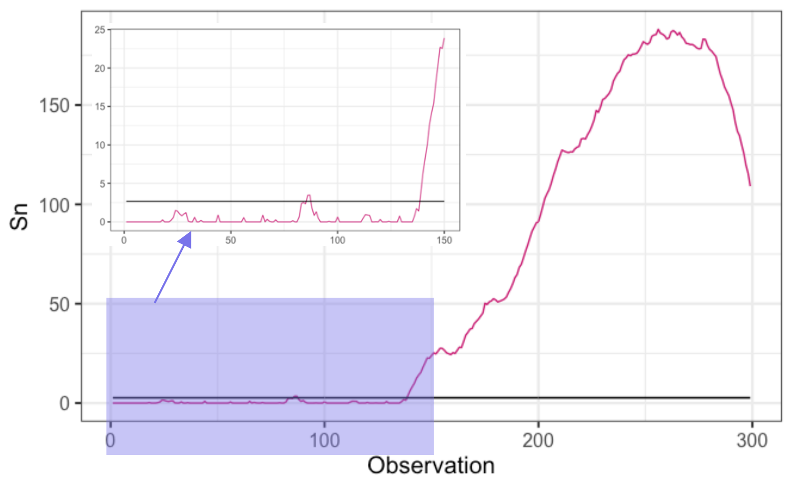


Figure 19. CUSUM chart calibrated to be optimum for a shift to μ0+0.02 in all variables. (2nd version)

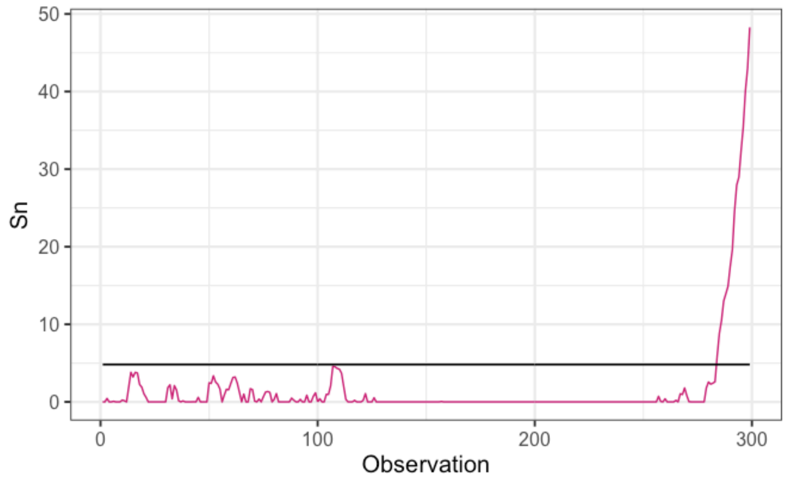


Figure 20. CUSUM chart calibrated to be optimum for a shift to μ0-0.01 in all variables. (2nd version)

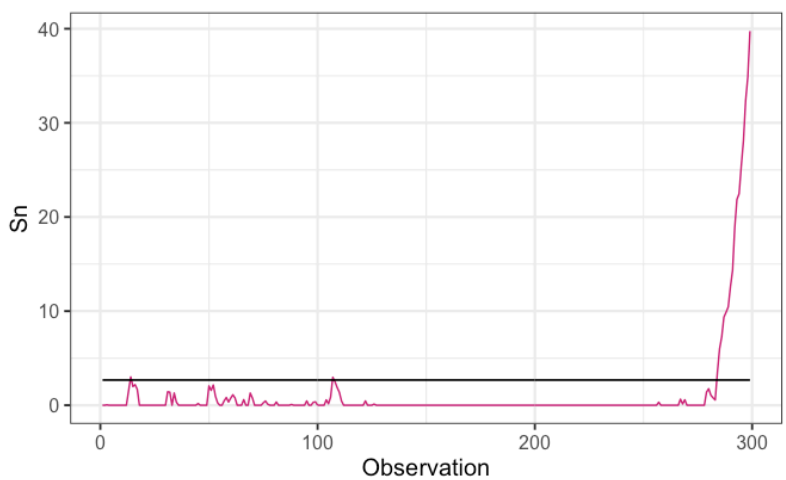


Figure 21. CUSUM chart calibrated to be optimum for a shift to μ0-0.02 in all variables. (2nd version)

A previously stated, some violations were detected between observations 85 and 100 on the individual Hotelling T^2 charts. These violations are visible in both Figure 18 and Figure 19. The shift detected near observation 150 is closely related with the loose guide pin. The control charts show that after the readjustment, neither the machine nor the process returned to its regular operation.

Chapter 4. Winsorization

As explained in subsection 2.2.3, winsorizing is limiting the effect of extreme values in the statistical data to reduce the risk of false alarms due to occasional outliers. This chapter will present experimentation made to evaluate the effect on Healy's performance when winsorization is applied.

Based on Healy's CUSUM, a winsorized model is proposed. Equations (7) and (8) are still used. For generating the winsorized CUSUM equation (9) is proposed:

$$S_n = \max(S_{n-1} + Z - 0.5D, 0) > CL, \tag{9}$$

where

$$Z = \begin{cases} \boldsymbol{a}'(\boldsymbol{x_n} - \boldsymbol{\mu_0}), & W > \boldsymbol{a}'(\boldsymbol{x_n} - \boldsymbol{\mu_0}) \\ W, & W \leq \boldsymbol{a}'(\boldsymbol{x_n} - \boldsymbol{\mu_0}) \end{cases}, \tag{10}$$

where W is a specified robustifying limit, and Z the standarized distance of the measurement.

Two scenarios were studied using equations (9) and (10), the first one presented in Section 4.1, analyzes the effect on the ARL when random OC outliers appear on an IC simulation. In Section 4.2, winsorization is analyzed in OC data and how the winsorized data might affect the earlier detection of shifts. Finally, in Section 4.3, winsorization is applied in the dataset analyzed in Section 3.2.

4.1 The effect of winsorizing random outliers in the in-control performance

An experiment was made to analyze the effect of spurious outliers appearing randomly in an IC dataset. Random data was generated with a 19-element mean vector of $\boldsymbol{\mu_0} = 0$ and a $p \times p$ identity covariance matrix.

Given a particular change with a Mahalanobis distance D given by equation (8), the probability of random outliers appearance, the size of those random outliers, and the winsorization threshold; the ARL was estimated in several simulations. The D distance defines the control chart calibration in order to detect the desired change in the mean. By detecting more significant changes in the mean, the ARL decreases. Outlier measurements were used to test the performance of the control chart against observations with a mean value different to the IC mean target. The outlier observation is composed of two parameters, the probability of appearance during runtime and the size of its mean given by a specific standard deviation. As the number of outlier observations increase, the ARL decreases, also depending on the outlier size. As stated before, winsorization limits the effect of extreme values; hence, a smaller winsorization value will increase the in-control ARL.

Table 2, Table 3 and Table 4 contain the results of the simulations given different winsorization conditions. As seen in the Tables, when the winsorization value is relatively small, as shown in Table 3, the in-control ARL becomes the largest of all the scenarios. Table 3 and Table 4 show that from a specific standard deviation size, the in-control ARL behaves similarly. For instance, when the chart is calibrated with $D = 0.50$, and the probability of an outlier is 0.01, an outlier of 6 standard deviations generates an in-control ARL of 81.0 when no winsorization is used. When winsorization is used, the same scenario produces an in-control ARL of 273.5 and 220.0 using $W = 2$ and $W = 3$, respectively. The same comparison can be made for any other combination.

The Mahalanobis distance, referred to as d_m, is defined as the amount of standard deviations times $\mu = 0.3162278$. The constant μ is defined in order to make the Mahalanobis distance as an integer. For example, by multiplying the mean vector by 2, the resulting Mahalanobis distance is 2.

Table 2. Random outliers without winsorization

		Probability			
D	***d_m***	**0.005**	**0.01**	**0.05**	**0.1**
0.5	**2**	295.7	236.7	80.4	39.1
	3	235.2	166.6	47.1	22.8
	4	181.0	119.3	32.3	16.9
	5	143.3	89.9	22.9	12.7
	6	132.7	81.0	19.5	10.2
1	**2**	286.4	214.3	70.5	31.9
	3	188.1	129.1	32.7	17.7
	4	130.4	79.5	19.0	9.8
	5	131.6	78.2	19.3	9.9
	6	130.3	79.0	18.9	9.9
2	**2**	273.7	205.8	67.8	32.2
	3	130.7	80.7	19.4	9.6
	4	131.5	79.7	19.5	9.9
	5	132.3	80.1	19.1	9.7
	6	128.0	78.6	19.0	9.7

Table 3. Random outliers using a winsorization limit W=2

D	d_m	Probability			
		0.005	**0.01**	**0.05**	**0.1**
0.5	**2**	321.0	276.3	117.7	58.1
	3	320.8	274.0	117.1	58.1
	4	318.1	278.6	116.6	60.7
	5	322.8	273.8	116.1	58.8
	6	314.4	273.5	119.0	58.8
1	**2**	313.9	270.8	112.5	55.4
	3	314.6	266.1	114.8	54.6
	4	315.7	271.9	110.0	54.7
	5	311.1	269.8	113.5	54.8
	6	314.0	278.1	113.5	54.7
2	**2**	302.9	253.8	94.8	45.7
	3	308.0	257.4	95.5	44.9
	4	307.3	254.7	95.3	45.1
	5	310.7	255.8	96.7	45.6
	6	303.1	260.2	95.0	45.6

Table 4. Random outliers using a winsorization limit W=3

D	d_m	Probability			
		0.005	**0.01**	**0.05**	**0.1**
0.5	**2**	288.7	240.5	79.5	38.5
	3	283.1	222.0	70.6	33.5
	4	278.2	224.0	70.3	33.5
	5	275.8	224.6	71.6	34.7
	6	280.4	220.0	69.9	34.3
1	**2**	277.5	220.2	68.4	32.2
	3	264.7	198.9	58.8	29.0
	4	255.8	197.3	59.8	28.0
	5	258.0	200.3	60.9	27.3

	6	261.5	196.9	59.2	28.0
	2	257.1	194.0	60.2	29.4
	3	236.4	166.6	49.1	24.1
2	**4**	232.2	165.3	48.5	24.1
	5	232.4	164.7	47.7	24.0
	6	235.5	167.0	49.1	24.7

4.2 The effect of winsorizing over the out-of-control performance

The second experiment consisted of analyzing how long it would take to Healy's CUSUM to detect OC observations with and without winsorization. Random data was generated with a 19-element mean vector, and a varying mean vector depending on the desired shift on the mean. The covariance matrix remains as a $p \times p$ identity matrix. The control chart was calibrated for a specific D value, with different μ_1 mean vector values, applying winsorization, and modifying the number of variables changing to the μ_1 mean vector, being μ_1 the OC mean vector.

Table 5, Table 6 and Table 7 present the results from the simulations in terms of ARL given the distinct D values. The effect of winsorization is slightly visible in the three tables, being the W=2 simulations the ones that had a more noticeable effect on the ARL, due to its increased detecting allowance. The ARL performance decreases considerably depending on how many variables are affected with μ_1. As the amount of variables reduces, the control chart looses its capability of early detection. Finally, as the μ_1 increases, the control chart became more capable of detecting the shift, and this is not affected by the winsorization. As a summary of the previous statements, winsorization affects the resulting ARL enough for the chart to be considered robust to drastic changes while keeping the effect of subtle changes.

For example, assuming the chart is calibrated using $D = 0.50$ and every variable present a change in the mean of 0.2294 (representing a Mahalanobis distance of 1.0), an out-of-control signal is generated for an ARL of 9.7, without winsorization. When winsorization is used, the same scenario produces an ARL of 9.9 and 9.8 using $W = 2$ and $W = 3$, respectively. The same comparison can be made for any other combination.

Table 5. OC simulation, D=0.5

		Not winsorize					2					3				
Shift	**mu1**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**
0	0	6.7076	366.6	363.7	371.3	373.2	6.2397	367.1	368.9	373.0	365.2	6.6756	365.8	373.4	365.9	371.1
0.50	0.1147	6.7076	23.5	36.7	59.1	128.6	6.2397	24.0	36.8	58.5	129.2	6.6756	23.4	36.1	58.0	128.3
1.00	0.2294	6.7076	9.7	14.2	22.0	58.3	6.2397	9.9	14.2	22.8	58.4	6.6756	9.8	14.2	22.1	57.8
1.50	0.3441	6.7076	6.1	8.6	12.9	33.1	6.2397	6.5	8.9	13.1	33.4	6.6756	6.2	8.6	13.0	33.1
2.00	0.4588	6.7076	4.5	6.2	9.1	22.2	6.2397	5.2	6.7	9.4	22.3	6.6756	4.6	6.2	9.1	22.0
3.00	0.6882	6.7076	3.0	4.0	5.7	12.9	6.2397	4.2	4.8	6.2	13.1	6.6756	3.3	4.2	5.8	12.9

Table 6. OC simulation, D=1

		Not winsorize					2					3				
Shift	**mu1**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**
0	0	4.0955	373.8	375.0	366.8	368.8	3.6928	369.8	366.1	368.7	374.8	4.0633	370.7	374.2	366.7	366.8
0.50	0.1147	4.0955	27.6	46.7	76.7	160.5	3.6928	28.0	46.7	76.3	160.2	4.0633	27.7	47.2	75.9	157.8
1.00	0.2294	4.0955	8.5	14.0	25.2	75.9	3.6928	8.9	14.3	25.9	78.2	4.0633	8.6	14.1	25.6	77.6
1.50	0.3441	4.0955	4.9	7.4	12.5	41.1	3.6928	5.2	7.6	12.7	42.0	4.0633	4.9	7.4	12.3	41.7
2.00	0.4588	4.0955	3.4	4.9	8.0	25.4	3.6928	3.9	5.3	8.2	25.7	4.0633	3.5	5.0	8.0	25.3
3.00	0.6882	4.0955	2.2	3.0	4.5	12.5	3.6928	3.1	3.6	4.9	12.9	4.0633	2.4	3.2	4.6	12.6

Table 7. OC simulation, D=2

		Not winsorize					2					3				
Shift	**mu1**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**	**CL**	**100% (19)**	**75% (14)**	**50% (10)**	**25% (5)**
0	0	2.1756	373.3	370.2	374.3	370.3	1.7034	362.5	373.0	366.2	373.5	2.0530	373.9	374.3	375.8	365.3
0.50	0.1147	2.1756	47.7	77.1	118.9	203.0	1.7034	46.4	76.3	115.0	201.8	2.0530	44.6	73.7	113.8	201.1
1.00	0.2294	2.1756	11.3	22.0	43.9	117.0	1.7034	11.6	22.0	42.2	116.7	2.0530	10.8	20.8	41.0	114.9
1.50	0.3441	2.1756	4.7	8.9	18.5	69.8	1.7034	5.0	9.3	19.3	68.3	2.0530	4.8	8.7	17.9	66.9
2.00	0.4588	2.1756	2.9	5.0	9.9	43.6	1.7034	3.3	5.3	10.2	42.4	2.0530	3.1	4.9	9.6	41.0
3.00	0.6882	2.1756	1.7	2.5	4.4	18.8	1.7034	2.2	2.9	4.6	18.7	2.0530	2.1	2.7	4.4	18.4

4.3 Winsorizing Healy's CUSUM

The charts presented in subsection 3.2 are now replicated applying equations (9) and (10). Two winsorized escenarios were analyzed. For a change of $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} \pm 0.01$, the charts presented in Figure 22 and Figure 23 have a $W = 2$, both with a CL of 4.456167. Then, for the same change in the mean but, with a $W = 3$, charts are shown in Figure 24 and Figure 25, both with a CL of 4.836532.

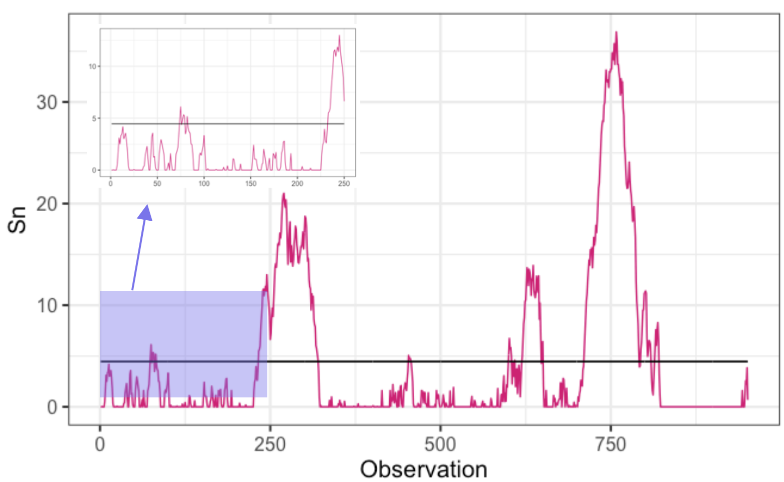


Figure 22. CUSUM chart calibrated to be optimum for a shift of µ0+0.01, W=2 (1st version)

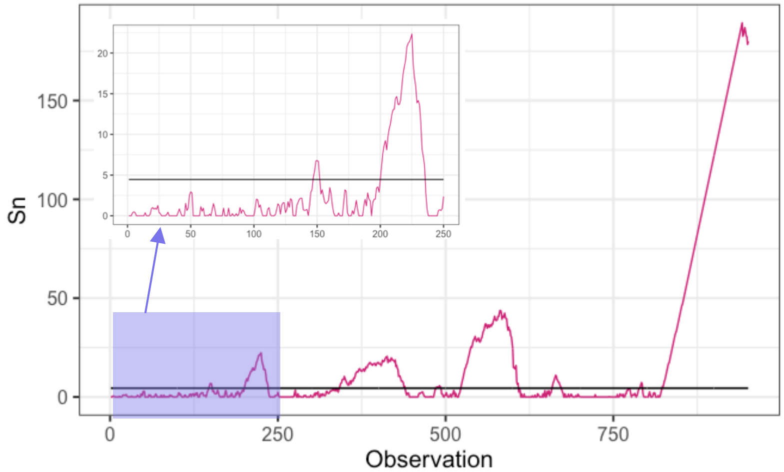


Figure 23. CUSUM chart calibrated to be optimum for a shift of μ0-0.01, W=2 (1st version)

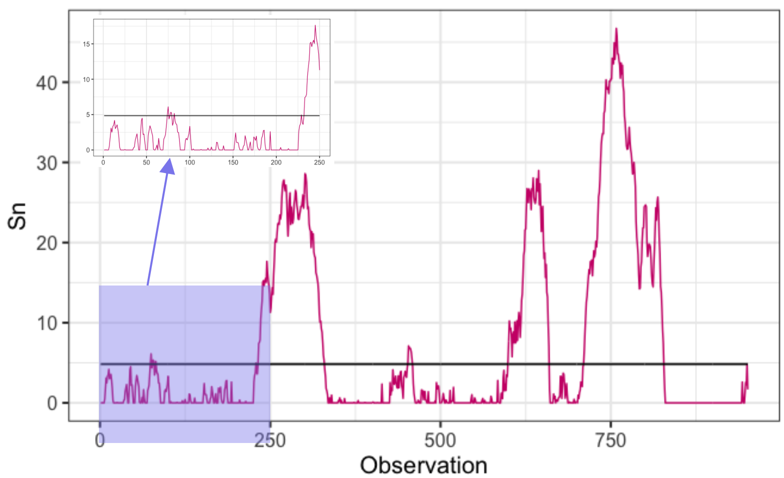


Figure 24. CUSUM chart calibrated to be optimum for a shift of μ0+0.01, W=3 (1st version)

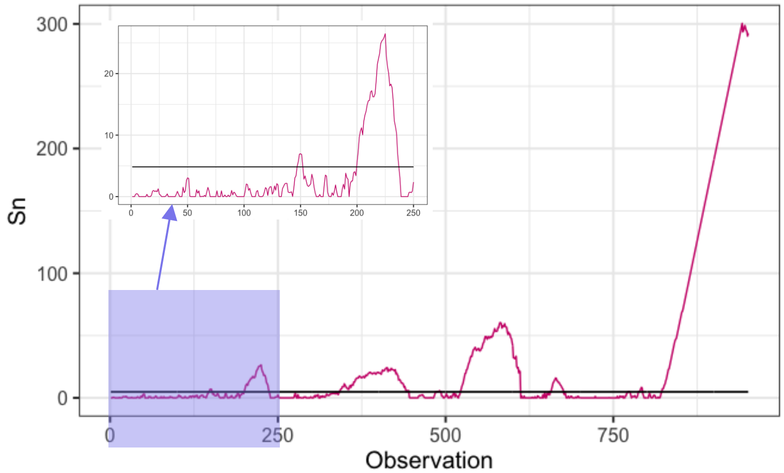


Figure 25. CUSUM chart calibrated to be optimum for a shift of μ0-0.01, W=3 (1st version)

The next charts present a change of $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} \pm 0.02$, for a $W = 2$, presented in Figure 26 and Figure 27, both with a CL of 2.298777. Also, for a change of $\boldsymbol{\mu 1} = \boldsymbol{\mu 0} \pm 0.02$, with $W = 3$, charts are shown in Figure 28 and Figure 29, both with a CL of 2.656877.

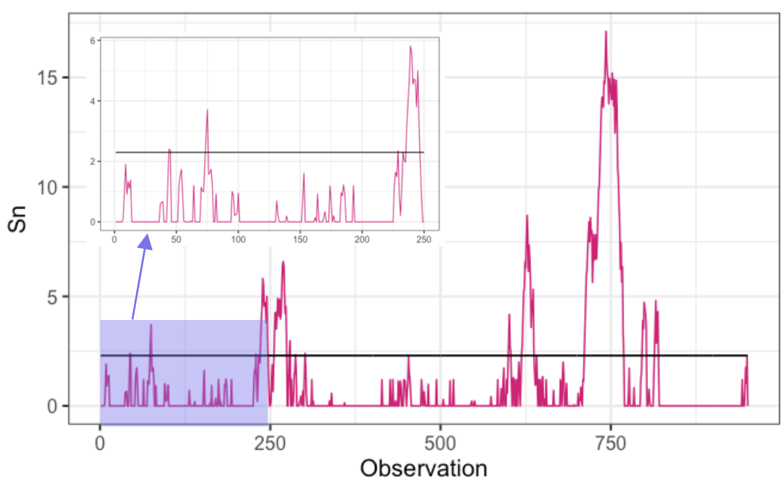


Figure 26. CUSUM chart calibrated to be optimum for a shift of μ0+0.02, W=2 (1st version)

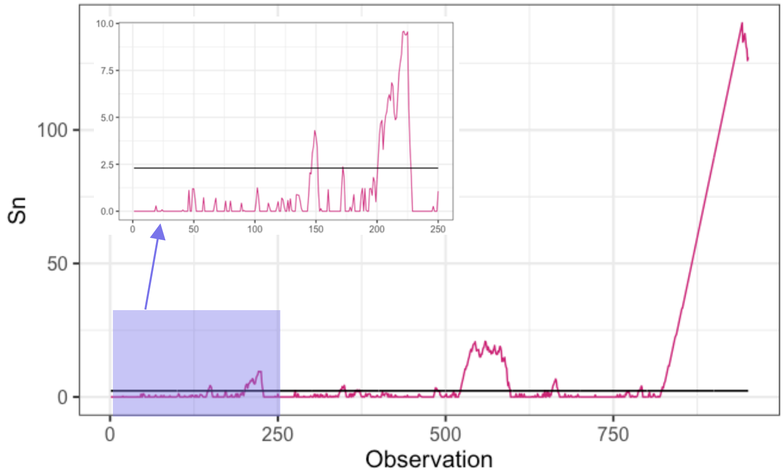


Figure 27. CUSUM chart calibrated to be optimum for a shift of μ0-0.02, W=2 (1st version)

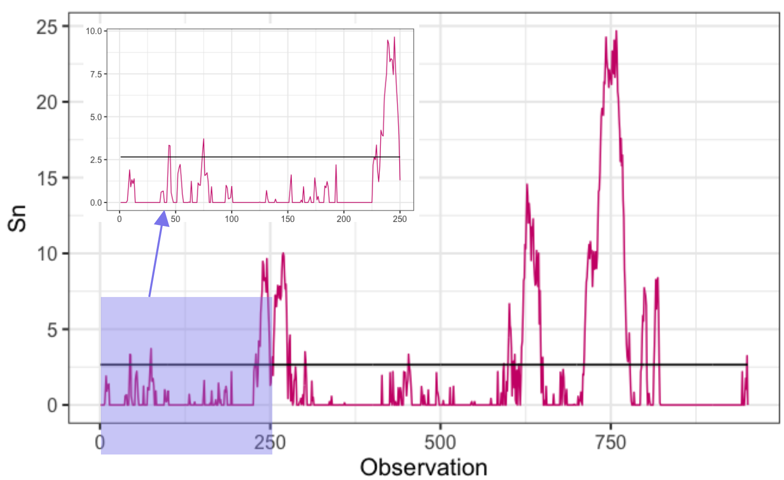


Figure 28. CUSUM chart calibrated to be optimum for a shift of μ0+0.02, W=3 (1st version)

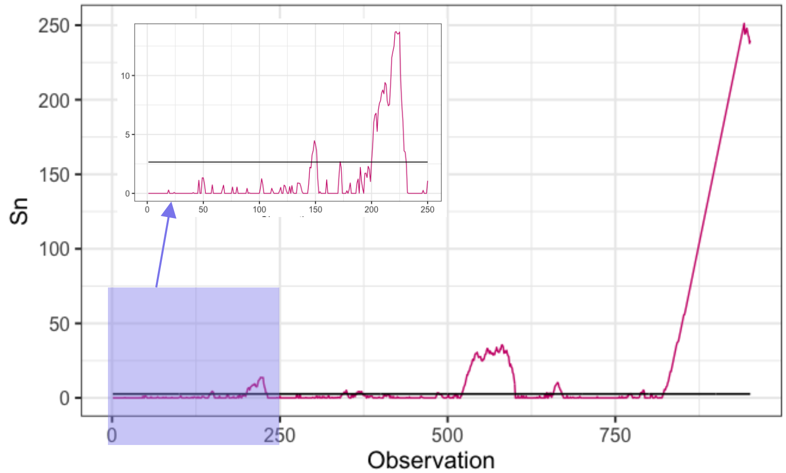


Figure 29. CUSUM chart calibrated to be optimum for a shift of μ0-0.02, W=3 (1st version)

Although the shape of graphed points does not changes, the value of each Sn is smaller; this can be seen when comparing the y-axis of the winsorized and not winsorized charts. It can be seen how winsorization limits the effect of extreme values, especially in the charts having a winsorization value of two. The detected changes by all charts are more likely to have been sustained over a short period rated than isolated changes.

Chapter 5. Data Visualization

As previously stated, a dashboard is an efficient tool for presenting information clearly and practically. This chapter is divided into two sections; the first one presents the procedure to obtain data from the machine until it is transformed into an R data frame. The second section explains the general view of the designed dashboard.

5.1 Obtaining data from the database

Many communication protocols are available to connect with an existing database, some of them consist of a specific protocol to connect to the database, but some others use an intermediary protocol such as `HTTP`. Since the existing solution already has a database interface, the proposed solution implies using and extending the features of the existing system.

Being `R` a high-level language, many extended features not related to statistical analysis are available as if it were another programming language. Notably, the `httr` library allows developers to perform `HTTP` requests such as `GET` and `POST`. The `HTTP` protocol is widely used for the average internet usages as browsers do; however, a new trend of online services called Rest API uses the `HTTP` protocol to interface with online programs. A Rest API is a program waiting to be triggered that is running in an online environment, an `HTTP` request carrying the appropriate parameters can trigger the API and reply to the request accordingly to the given parameters.

A small Rest API was coded and deployed into the hosting service provider environment in order to interface the R application with the database in a cautious manner. The Rest API is deployed as a specific subdomain of the Consortium website and is executed whenever a `POST` request is received; no additional parameters are required. When triggered, the hosting controller running a `PHP` script gathers a subset of the measurement dataset that includes the nineteen measurements points and the date and time when uploaded. Then, the controller parses the subset as a comma-separated value (CSV) string and outputs it as the replied body for the `HTTP` request.

As the CSV format is widely used for programming purposes, the `R` language natively supports this kind of format and is capable of converting the incoming CSV response into a data frame.

APIs like this one are widely used by large service providers to link two pieces of software together for applications like IoT. This technique will also be critical for the development of a digital twin of the machine capable of monitoring and forecasting not just its conditions, but the status of the process as the virtual twin may be fed with an additional training set different from the incoming samples from the die casting process to enhance the performance of the real twin.

The API developed for this application may be reused and extended to support various other links the twins may need, as a two way communication between the two.

5.2 Creating the Dashboard

Often, deploying a Rest API written in a high-level language is hard to do, due to the hosting service providers restrictions and the complexity of making the application compatible with the hosting environment. Fortunately, RStudio provides the Shiny App framework which allows to develop and test both front and back ends of a Web App offline and only using the `R` language.

The Shiny App framework is divided into two main scripts: the server and the user interface (UI). The UI section provides a simple interface for developing a graphical user interface that may include plots, buttons, checkboxes, and order standard widgets for the user to use as an interface. The server section runs a given program and can use inputs from the UI widgets and output results as plots or labels; any R script can be executed in this section. Once the application is ready to be deployed, the RStudio publish tool will include any dependency on the RStudio hosting environment, and finally, the application will be online and ready to be used.

A dashboard was designed to execute the statistical algorithms explained in Chapter 3. The process data comes from the existing database by using the procedure described in section 5.1. By default, the last fifty observations are displayed on the graphs, but the user can select a custom lapse using the calendar and time input boxes. Figure 30 presents the general view of the dashboard when accessing via computer, with the following parts:

1. Date selection range
2. Time selection range
3. Execute button (Necessary to press after every change)
4. “X-axis as sample” checkbox: when checked the x-axis on the graphs is the number of observation, when not checked the x-axis shows the hour in which the observations was registered.
5. The number of observations presented in the actual graph

6. Utilization percentage
7. Tabs for each graph
8. Graph section

The following algorithm was used to obtain the utilization percentage:

1. Equation (12) describes the active time:

$$AT = 3 \cdot mean(\Delta t_n), \ \ \Delta t_n \leq 100 \qquad (12)$$

Where Δt_n is a vector containing the elapsed time between observations.

2. Get the ratio of the active and the available time of the machine and multiply by one hundred to get the percentage.

Figure 30. Dashboard default view

Figure 31 shows the connection diagram of every model in the current solution. As stated before, the measurement system uploads the observations to the existing database using the `HTTP` protocol (red arrows), while the Shiny App request data from that same database using the custom made `readFromDB` function.

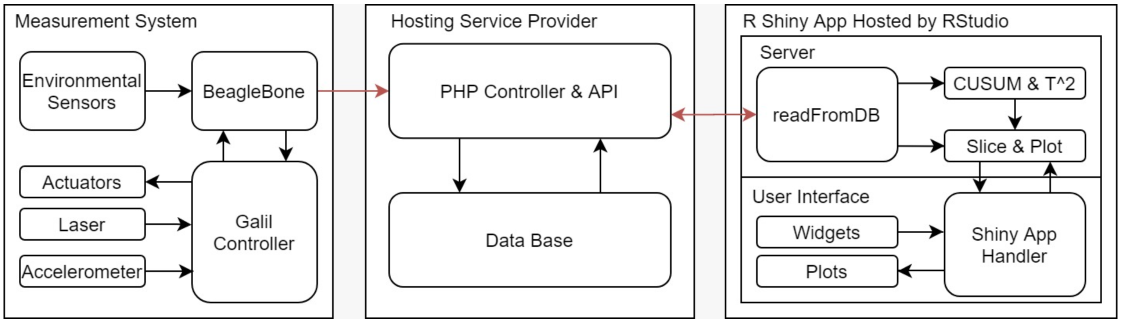


Figure 31. Interconnection block diagram

Chapter 6. Conclusion and future work

6.1 Conclusions

The thesis presented the statistical analysis of a multivariate process dataset using Healy's CUSUM and the study of how winsorization affects the performance of the used control chart to limit the effect of outliers. Also, the implementation of SPM in the measurement machine data and the visual presentation of the resulting charts. The three hypotheses were proven and did answer the research questions.

Regarding how data can be used to forecast machine performance and how reports can be produced on time, Healy's CUSUM is a good option for detecting subtle and sustained shifts for this case, as the already known situations were detected earlier before becoming noticeable in the raw data. However, this solution is not able to distinguish between an event caused by the process or the machine as the statistic only highlights an out-of-control situation. Also, the online dashboard has essential information for allowing any person to observe the general behavior of the process and the machine.

For the question regarding the performance difference of the winsorized in-control chart, winsorization does limit the effect of isolated changes in an IC chart, resulting in an ARL closer to the expected performance. However, the effect is slightly visible with a $W = 3$ compared to the not winsorized scenario.

As of the performance difference of the winsorized out-of-control chart in comparison with Healy's proposal, winsorizing has minimal effect on the ARL when detection OC samples, the most noticeable difference is found by changing the number of variables affected by the change of mean.

Finally, this work used techniques that will be useful for the development of a digital twin. Using an online API to get the data from an online hosting service would allow a seamless communication between the real and the virtual twins.

6.2 Future work

More variables can be added to the CUSUM analysis if these are significant for the process performance (i.e. environmental variables such as temperature, humidity, and pressure).

As the dashboard designed in this thesis is a preliminary version, improvements can be made to it. First, replace the IC dataset with at least 200 observations from different workpieces under a control situation; this means observing when the pieces are being measured to corroborate all the measurements and suppress any assignable causes. The CUSUM statistic will be more accurate using the previous recommendations.

Other control charts can be used, for example, a multivariate EWMA, which is also known to work with subtle and sustained changes. Other robustification methods can be investigated, analyzed and compared with this work's proposal; some examples of these methods could be trimmed mean, Huber, turkey bi-square, and Hampel [37].

Even though the Shapiro Wilks test did not show normality on data, by using Healy's CUSUM, the problem could be treated as a normal scenario. Evaluating the dataset with other multivariate control charts might present the need for pre-processing data before the analysis.

The work done in this thesis can be repeated using the previously proposed control chart and robustification methods when the machine collects enough data from different scenarios, including different types of perturbations. Also, the analysis can be repeated by evaluating the effect of robustifying methods using a non-normal dataset

Machine learning might be a considerable followup for the automatic classification of the data. This work didn't use Machine Learnng algorithms due to the simple nature of the analyzed data. However, if new sensors are included, many other combinations and patterns might be present in the output data, and it might not be as easy to identify for the a human as it would be for a Machine Learning classifier.

The presented dashboard contains information from three days only and with known problems. As the machine is expected to work during regular work shifts, it is vital to be able to discriminate junk data; this because it has been proved that operators tend to trigger the measurement process without a workpiece and the control chart can detect this measurement as a shift when it is, in fact, a false alarm.

Another critical aspect to include is a reset function that triggers when a shift is detected and corrected. The reset feature is needed because the CUSUM works as an error accumulator, and an error state will not stop showing until the error is corrected, and the error accumulator reset.

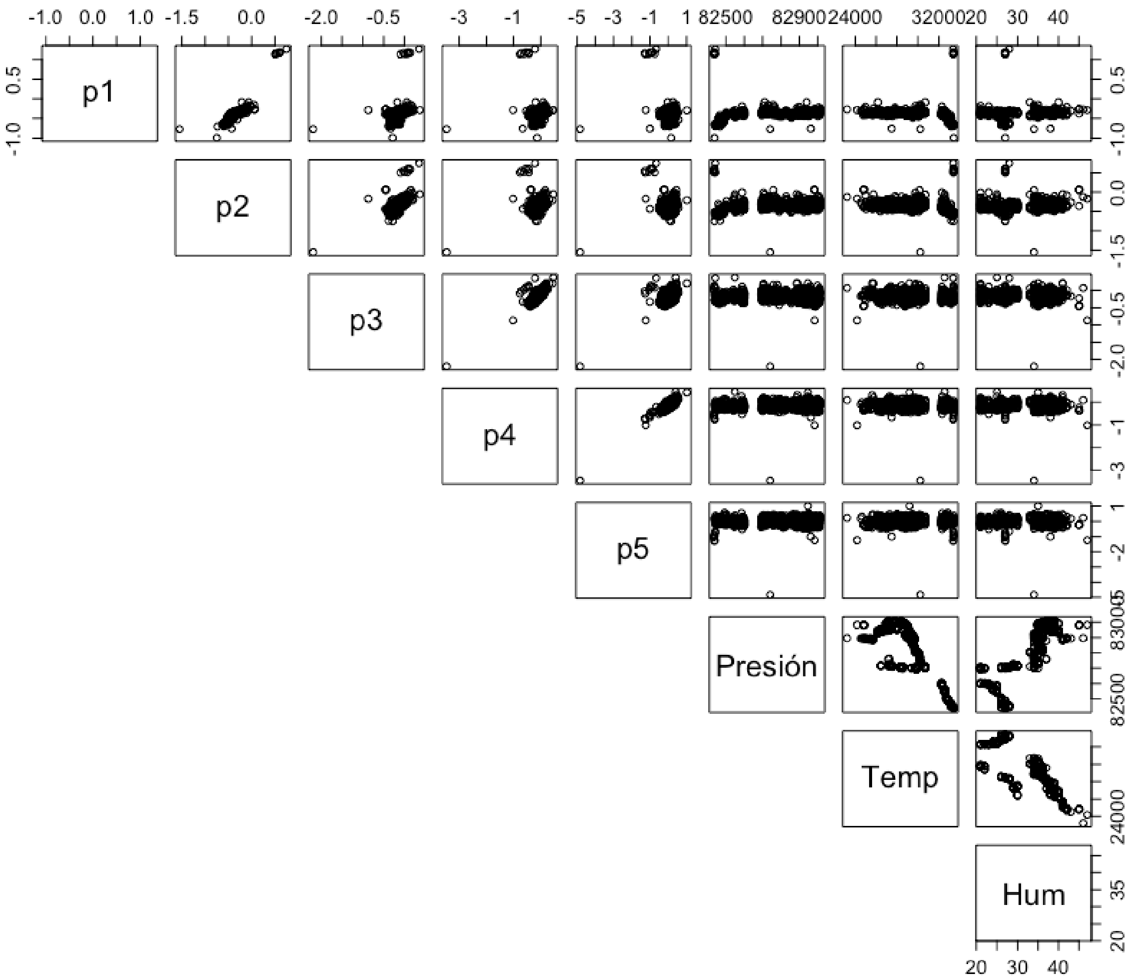
As for the dashboard, a useful feature would be to implement alarms messages that can be sent to a mobile or appear on the screen to alert workers about a possible problem.

The same approach for developing the dashboard can be used to create a digital twin. By enhancing the API functionality, a closed-loop system can be developed for feeding the online data to a simulation program to forecast and compare the machine’s performance.

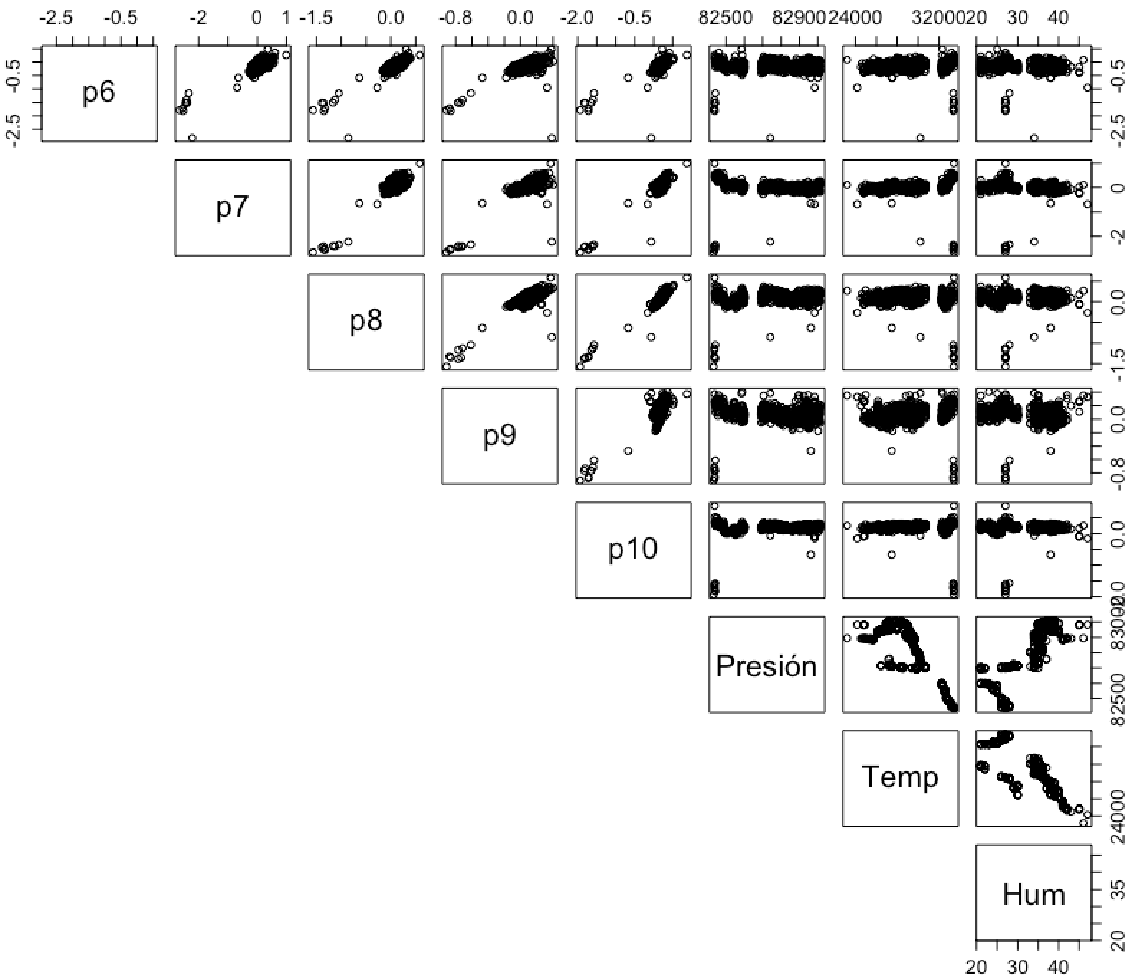
Appendixes

Appendix A

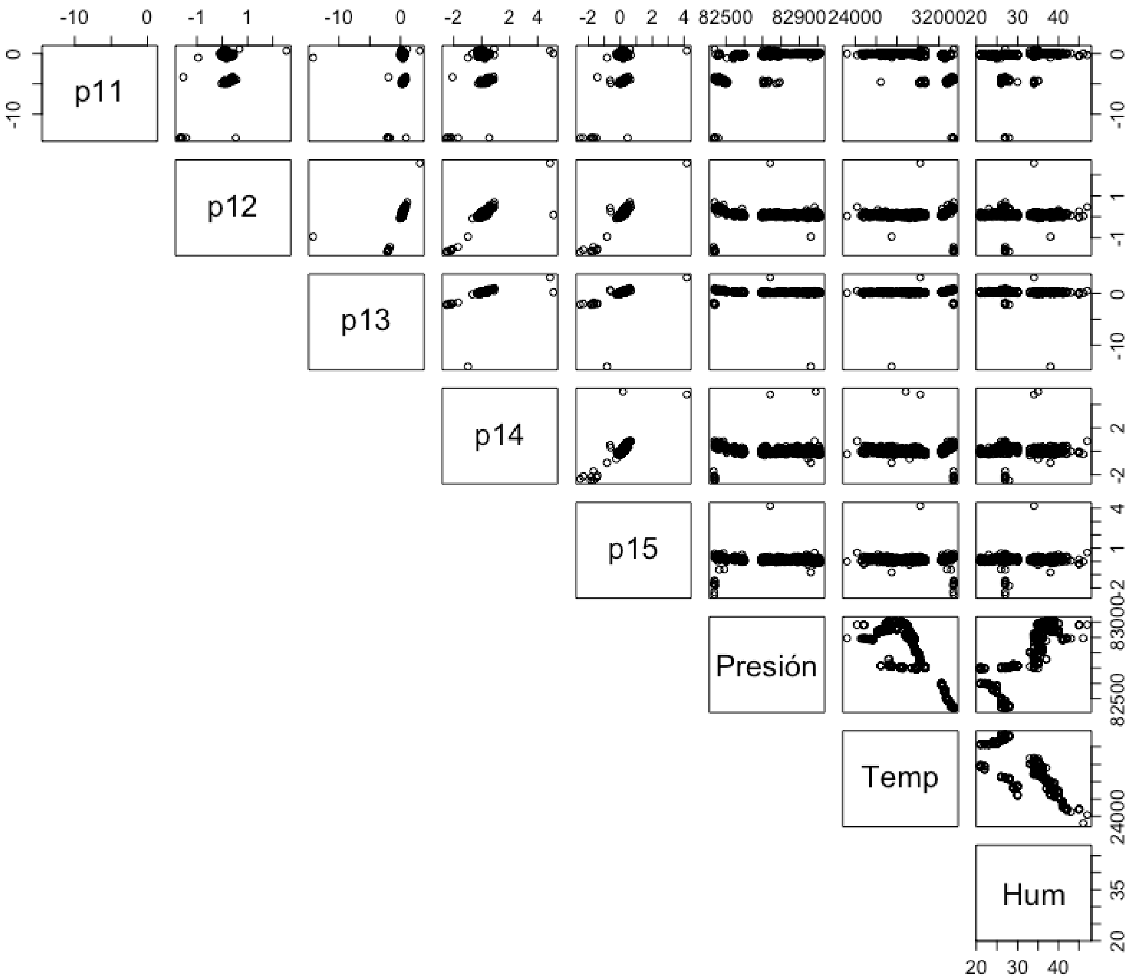
This appendix contains the correlation analysis between the nineteen variables and environmental parameters.



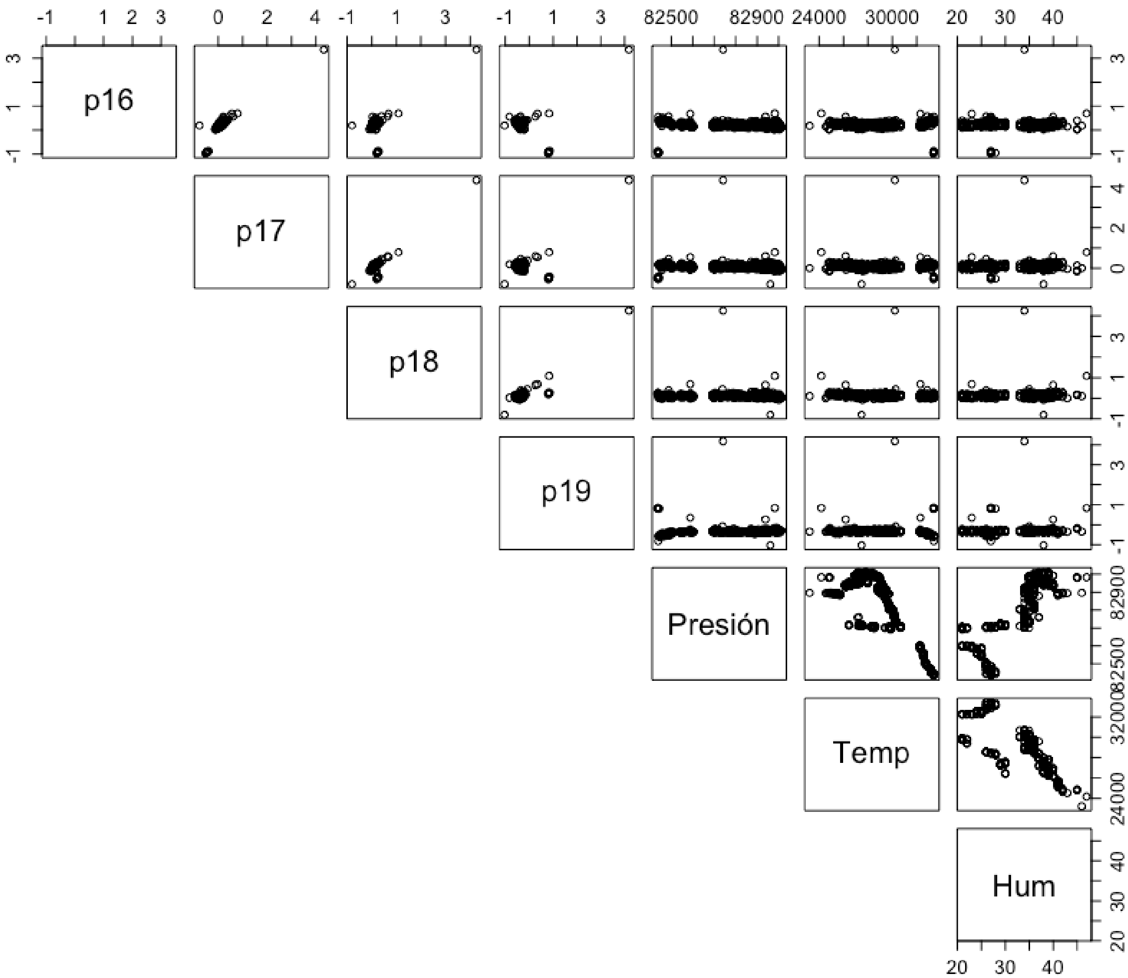
Appendix A. 1. Correlation matrix plot



Appendix A. 2. Correlation matrix plot (cont.)



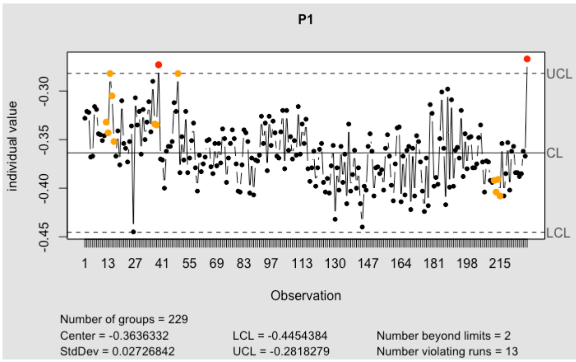
Appendix A. 3. Correlation matrix plot (cont.)

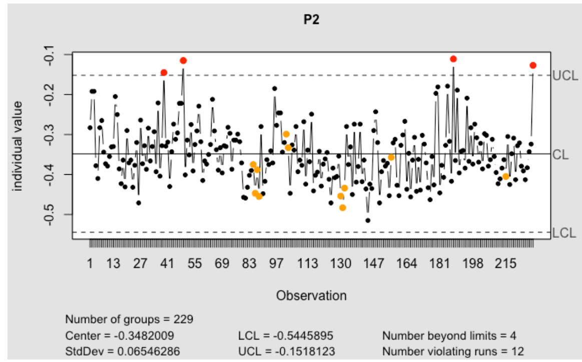


Appendix A. 4. Correlation matrix plot (cont.)

Appendix B

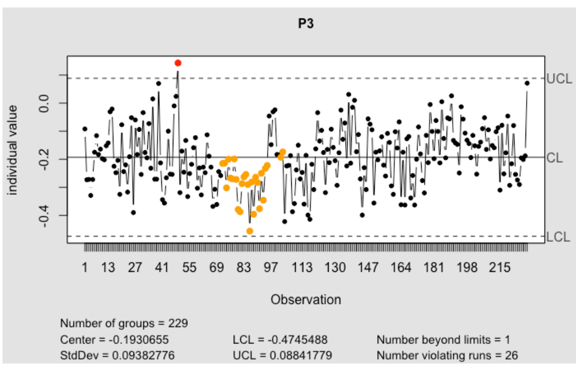
This appendix presents the univariate Shewhart control chart of the IC dataset for the nineteen variables using the data gathered from the first version of the machine.

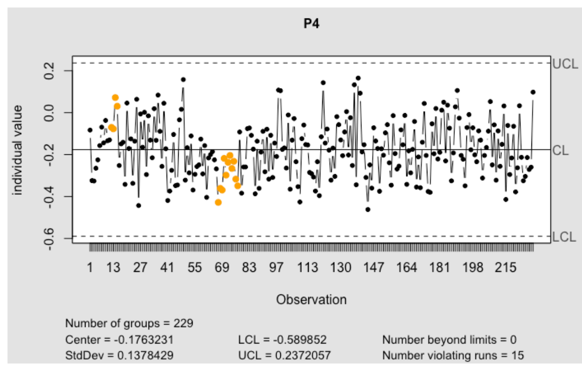




Appendix B. 5. Shewhart chart for P2 (1st version)

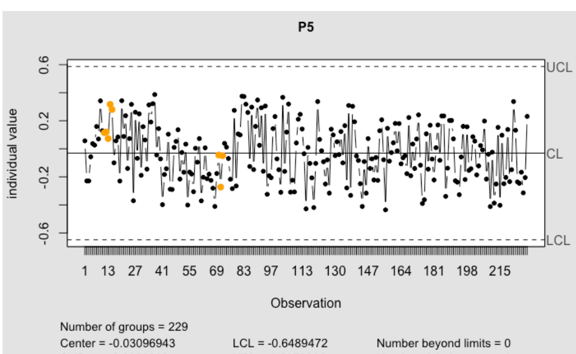
Appendix B. 6. Shewhart chart for P1 (1st version))

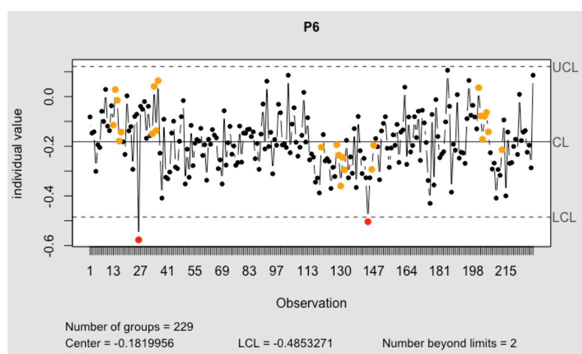




Appendix B. 7. Shewhart chart for P3 (1st version))

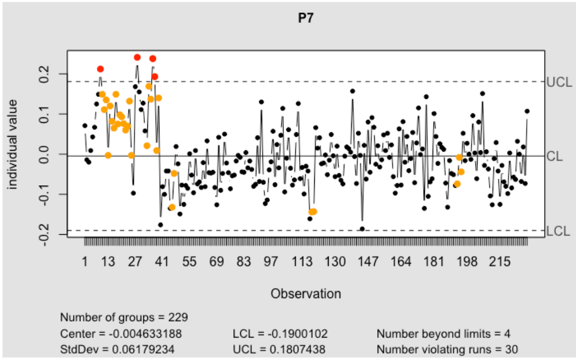
Appendix B. 8. Shewhart chart for P4 (1st version))

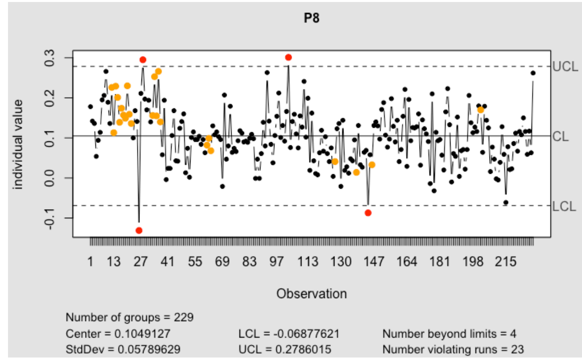




Appendix B. 9. Shewhart chart for P5 (1st version)

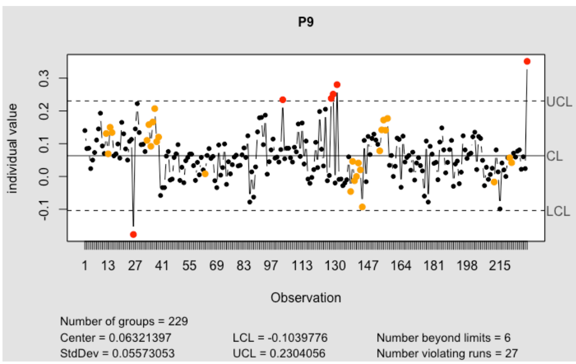
Appendix B. 10. Shewhart chart for P6 (1st version)

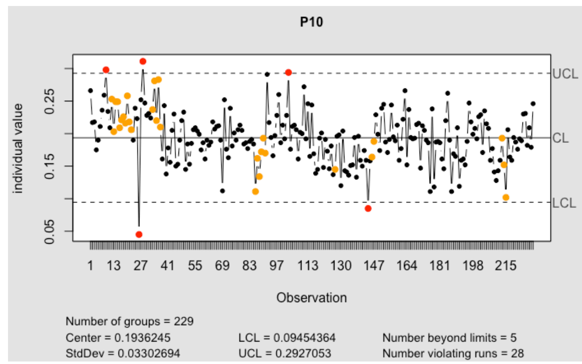




Appendix B. 11. Shewhart chart for P7 (1st version)

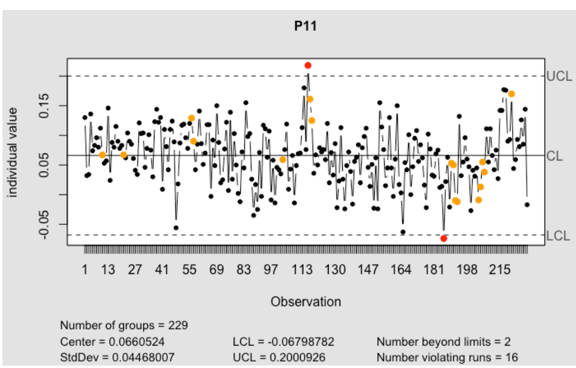
Appendix B. 12. Shewhart chart for P8 (1st version)

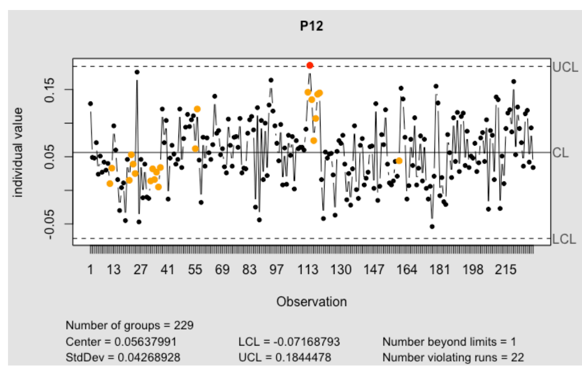




Appendix B. 13. Shewhart chart for P9 (1st version)

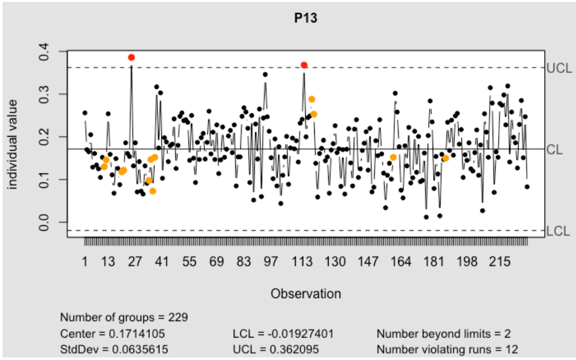
Appendix B. 14. Shewhart chart for P10 (1st version)

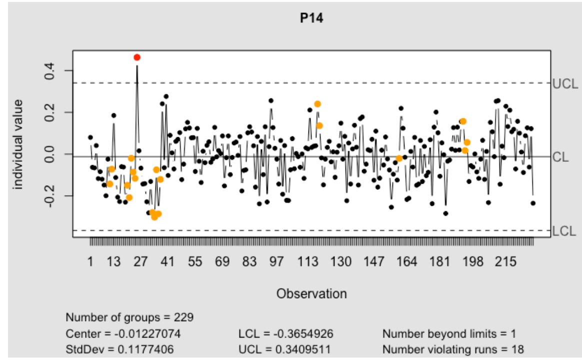




Appendix B. 15. Shewhart chart for P11 (1st version)

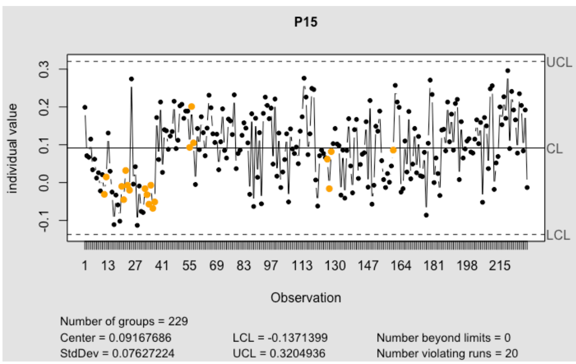
Appendix B. 16. Shewhart chart for P12 (1st version)

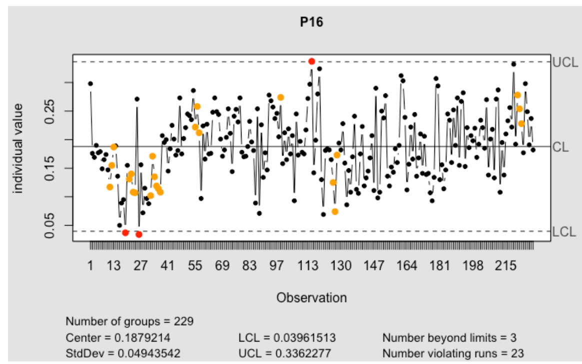




Appendix B. 17. Shewhart chart for P13 (1st version)

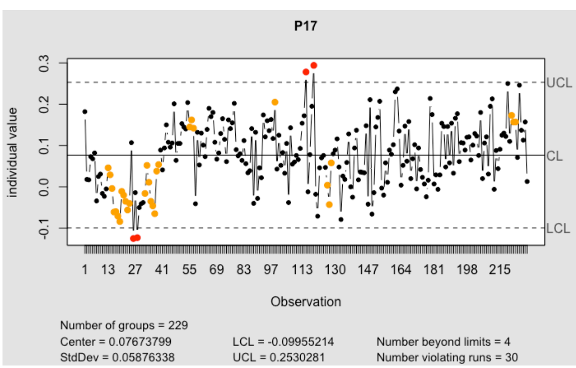
Appendix B. 18. Shewhart chart for P14 (1st version)

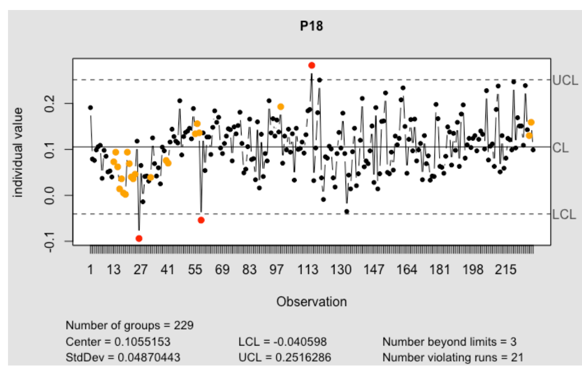




Appendix B. 19. Shewhart chart for P15 (March 2019)

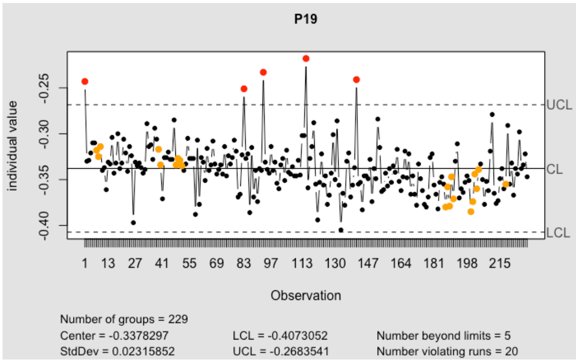
Appendix B. 20. Shewhart chart for P16 (March 2019)





Appendix B. 21. Shewhart chart for P17 (1st version)

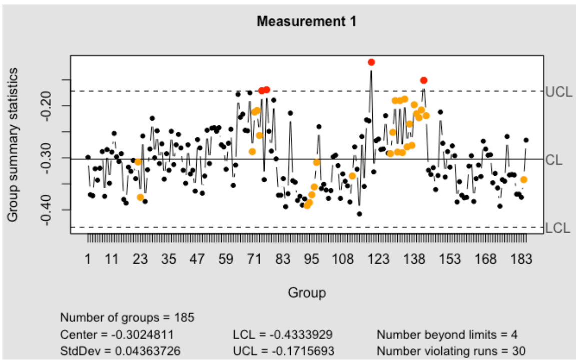
Appendix B. 22. Shewhart chart for P18 (1st version)



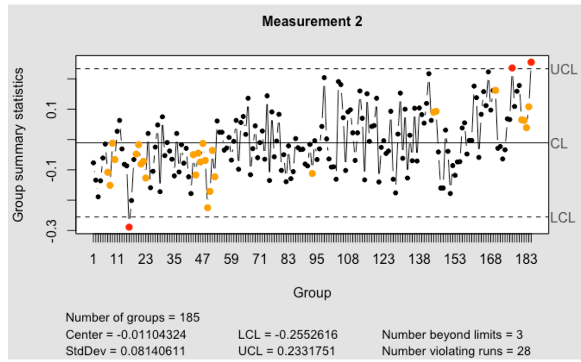
Appendix B. 23. Shewhart chart for P19 (1st version)

Appendix C

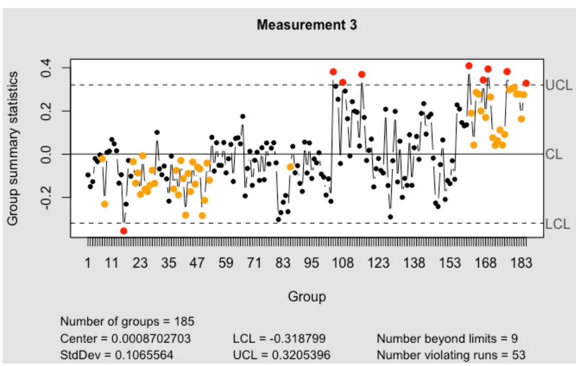
This appendix presents the univariate Shewhart control chart of the IC dataset for the nineteen variables using the data gathered from the second version of the machine.



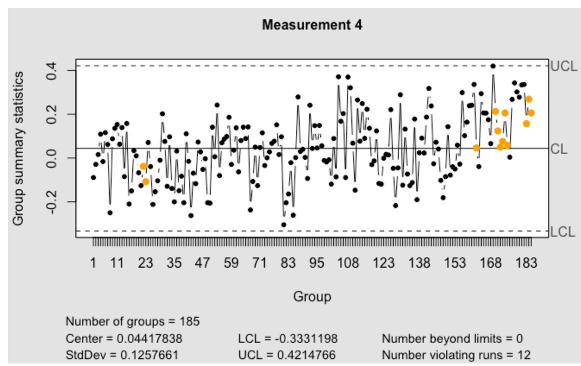
Appendix C. 1. Shewhart chart for P1 (2nd version)



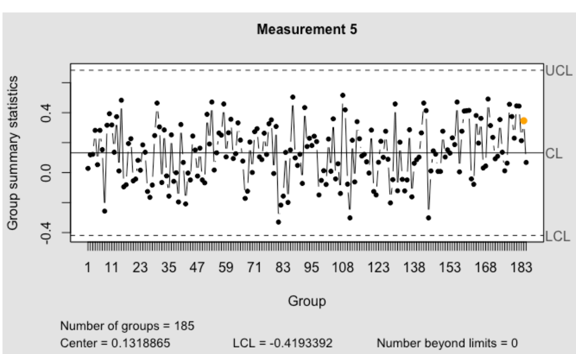
Appendix C. 2. Shewhart chart for P2 (2nd version)



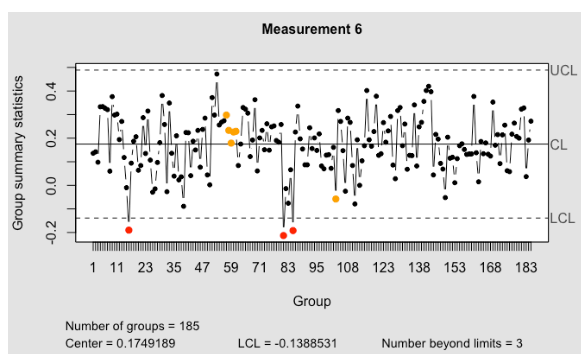
Appendix C. 3. Shewhart chart for P3 (2nd version)



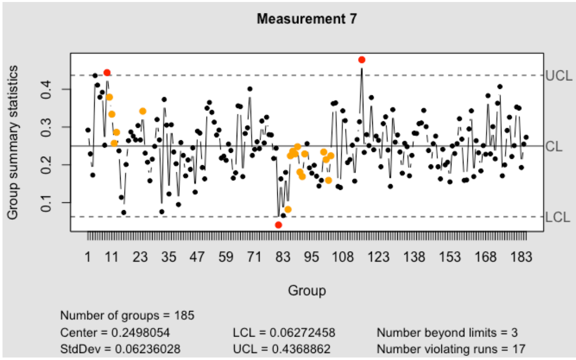
Appendix C. 4. Shewhart chart for P4 (2nd version)



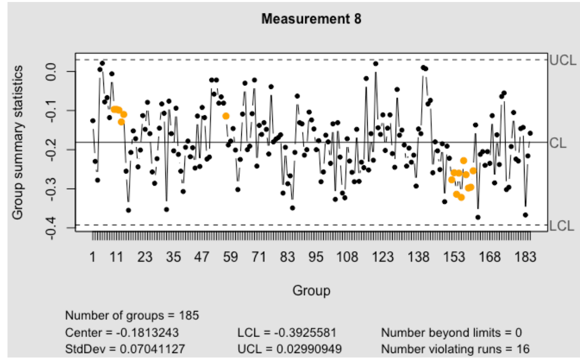
Appendix C. 5. Shewhart chart for P5 (2nd version)



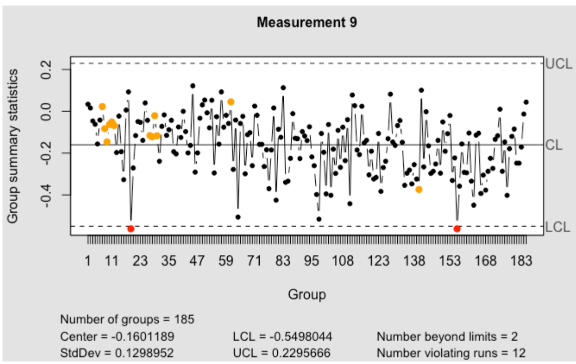
Appendix C. 6. Shewhart chart for P6 (2nd version)



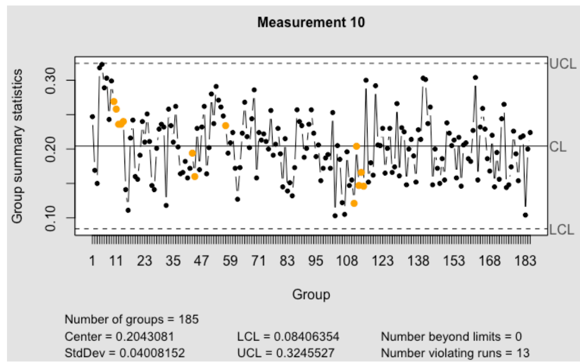
Appendix C. 7. Shewhart chart for P7 (2nd version)



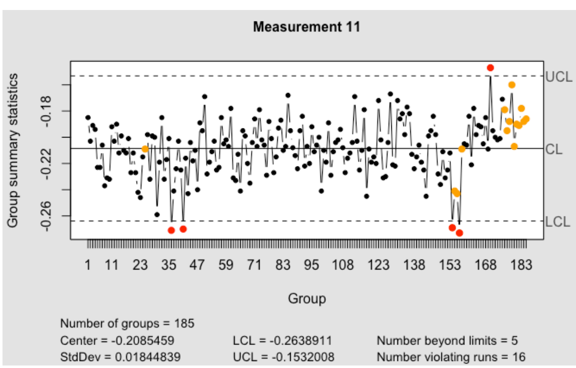
Appendix C. 8. Shewhart chart for P8 (2nd version)



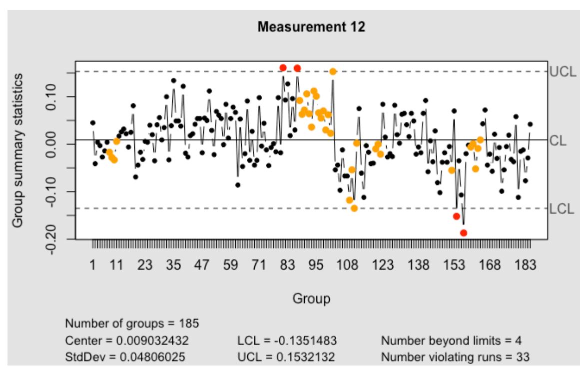
Appendix C. 9. Shewhart chart for P9 (2nd version)



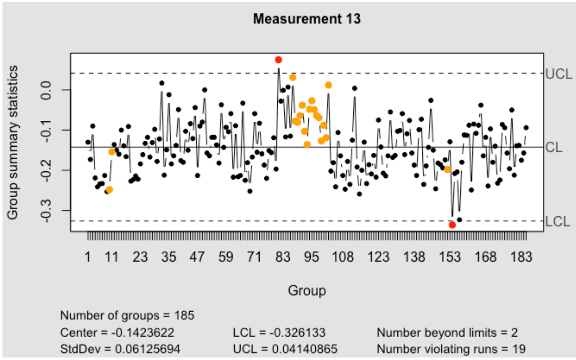
Appendix C. 10. Shewhart chart for P10 (2nd version)



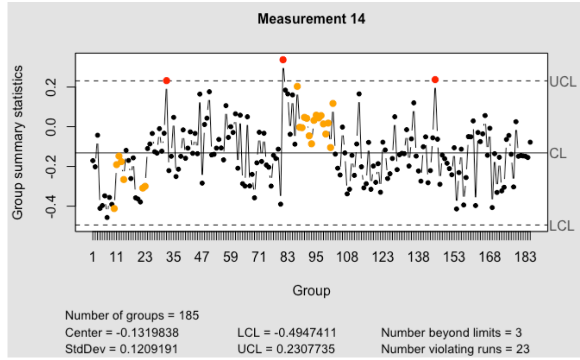
Appendix C. 11. Shewhart chart for P11 (2nd version)



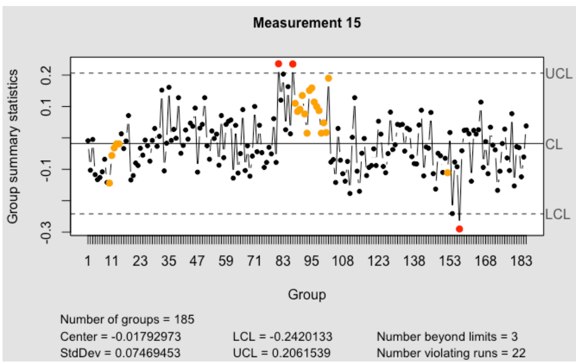
Appendix C. 12. Shewhart chart for P12 (2nd version)



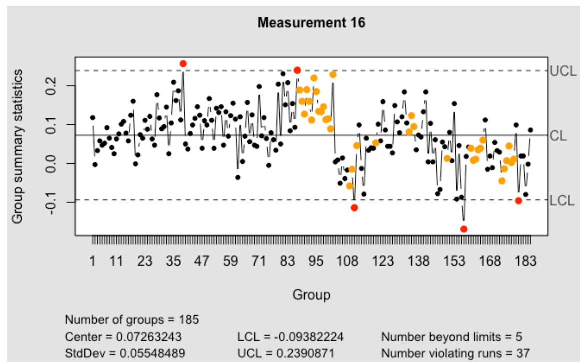
Appendix C. 13. Shewhart chart for P13 (2nd version)



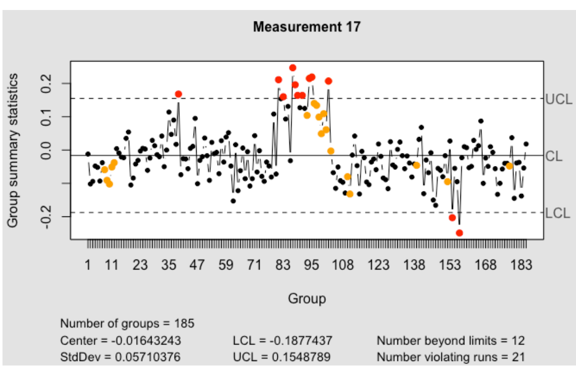
Appendix C. 14. Shewhart chart for P14 (2nd version)



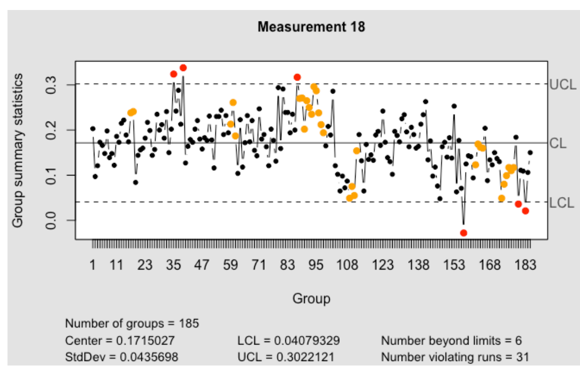
Appendix C. 15. Shewhart chart for P15 (2nd version)



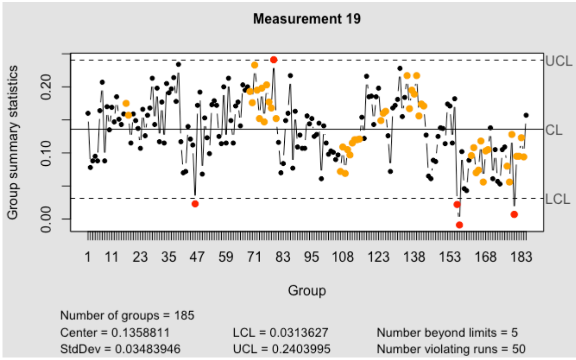
Appendix C. 16. Shewhart chart for P16 (2nd version)



Appendix C. 17. Shewhart chart for P17 (2nd version)



Appendix C. 18. Shewhart chart for P18 (2nd version)



Appendix C. 19. Shewhart chart for P19 (2nd version)

References

[1] Q. Peihua, Introduction to Statistical Process Control, Boca Ratón, Florida: CRC Press, 2014.

[2] D. C. Montgomery, Introduction to Statistical Quality Control, Wiley, 2009.

[3] A. P. Castro, *Design of the technological infrastructure for data acquisition of an in-line measuring Industry 4.0 compatible machine,* Monterrey: ITESM, 2017.

[4] D. Guamán, *Mechatronic Design of a Fast-Non-Contact Measurement System for Inspection of Castings Parts in Production Line,* Monterret: ITESM, 2018.

[5] J. D. Healy, "Multivariate CUSUM Procedures," *Technometrics,* vol. 29, no. 4, pp. 409-412, 1987.

[6] J. Qin, Y. Liu and R. Grosvenor, "A Categorical Framework of Manufacturing for Industry 4.0 and Beyond," *Procedia CIRP 52,* pp. 173-178, 2016.

[7] L. S. Dalenogare, G. B. Benitez, N. F. Ayala and A. G. Frank, "The expected contribution of Industry 4.0 technologies for industrial performance.," *International Journal of Production Economics,* vol. 204, pp. 383-394, 2018.

[8] J. Lee, H.-A. Kao and S. Yang, "Service innovation and smart analytics for Industry 4.0 and big data environment," *Procedio CIRP,* vol. 16, pp. 3-8, 2014.

[9] M. Bortolini, E. Ferrari, M. Gamberi, F. Pilati and M. Faccio, "Assembly system design in the Industry 4.0 era: a general framework," *IFAC PapersOnLine,* pp. 5700-5705, 2017.

[10] S. Kaisler, F. Armour, J. A. Espinosa and W. Money, "Big data: Issues and challenges moving forward," in *46th Hawaii International Conference on System Sciences*, 2013.

[11] P. Russom, "Big data analytics. TDWI best practices report," fourth quarter, 2011.

[12] S. Tiwari, H. Wee and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: Insights to industries," *Computers & Industrial Engineering,* no. 115, pp. 219-330, 2018.

[13] G. Q. Huang, R. Y. Zhong and K. L. Tsui, "Special issue on 'Big data for service and manufacturing supply chain management'," *Int. J. Production Economics,* no. 165, pp. 172-173, 2015.

[14] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International journal of Information Mangement,* no. 35, pp. 137-144, 2015.

[15] R. Séguin, J.-Y. Potvin, M. Gendreau, T. G. Crainic and P. Marcotte, "Real-time decision problems: An operational research perspective," *Journal of the Operational Research Society,* 1997.

[16] G. Michael, "Digital Twin: Manufacturing Excellence through Virtual Factory Replication," pp. 1-7, 2015.

[17] F. Tao, M. Zhang and J. Cheng, "Digital twin workshop: A new paradigm for future workshop," *Computer Integrated Manufacturing Systems,* vol. 23, no. 1, pp. 1-9, 2017.

[18] F. Tao and M. Zhang, "Digital Twin Shop-Floor: A new Shop-Floor PAradigm Towards Smart Manufacturing," *IEEE Acess 5,* pp. 20418-20427, 2017.

[19] R. Noorossana and S. J. M. Vaghefi, "Effect of Autocorrelation on Performance of the MCUSUM Control Chart," *Quality and Reliability Engineering International,* no. 22, pp. 191-197, 2006.

[20] E. S. Page, "Continuous Inspection Schemes," *Biometrika Trust,* vol. 41, pp. 100-115, 1954.

[21] S. W. Roberts, "Control charts tests based on geometric moving averages," *Technometrics,* no. 1, pp. 239-250, 1959.

[22] P. Castagliola, G. Celano and S. Fichera, "A new CUSUM-S2 control chart for monitoring the process variance," *Journal of Quality in Maintenance Engineering,* vol. 15, no. 4, pp. 344-357, 2009.

[23] O. Bodnar and W. Schmid, "CUSUM charts for monitoring the mean of a multivariate Gaussian process," *Journal of Statistical Planning and Inference,* no. 141, pp. 2055-2070, 2011.

[24] B. Zaman, M. H. Lee, M. Riaz and M. R. Abujiya, "An adaptive EWMA chart with CUSUM accumulate error-based shift estimator for efficient process dispersion monitoring," *Computers & Industrial Engineering,* no. 135, pp. 236-253, 2019.

[25] B. Miekley, E. Stamer, I. Traulsen and J. Krieter, "Implementation of multivariate cumulative sum control charts in mastitis and lameness monitoring," *American Dairy Science Association,* no. 96, pp. 5723-5733, 2013.

[26] A. Haq, T. Munir and M. B. Khoo, "Dual multivariate CUSUM mean charts," *Computers & Industrial Engineering,* no. 137, 2019.

[27] N. Tracy, J. Young and R. Mason, "Multivariate Control Charts for Individual Observations," *Journal of Quality Technology,* no. 24, 1992.

[28] D. M. Hawkins and D. H. Olwell, Cumulative Sum Charts and Charting for Quality Improvement, Springer Science & Business Media, 2012.

[29] D. M. Hawkins, "Robustification of Cumulative Sum Charts by Winsorization," *Journal of Quality Technology,* vol. 25, no. 4, pp. 248-261, 1993.

[30] J. M. Lucas and R. B. Crosier, "Robust cusum: a robustness study for cusum quality control schemes," *Communications in Statistics - Theory and Methods,* vol. 11, no. 23, pp. 2669-2687, 1982.

[31] D. M. Hawkins, "Self-Starting Cusum Charts for Location and Scale," *Journal of the Royal Statistical Society. Series D (The Statistician),* vol. 36, no. 4, pp. 299-316, 1987.

[32] J. J. Pignatiello and G. C. Runger, "Comparisons of Multivariate CUSUM Charts," *Journal of Quality Technology,* vol. 22, no. 3, pp. 173-186, 1990.

[33] H. Tokola, C. Gröger, E. Järvenpää and E. Niemi, "Designing manufacturing dashboards on the basis of a Key Performance Indicator survey," *Procedia CIRP 57,* pp. 619-624, 2016.

[34] D. Nadoveza and D. Kiritsis, "Concept for Context-Aware Manufacturing Dashboard Applications," *IFAC Proceedings Volumes,* vol. 46, no. 9, pp. 204-209, 2013.

[35] S. Mazumdar, A. Varga, V. Lanfranchi, D. Petrelli and F. Ciravegna, "A Knowledge Dashboard for Manufacturing Industries," *Extended Semantic Web Conference,* pp. 112-124, 2011.

[36] W. W. Eckerson, Performance dashboards: measuring, monitoring, and managing your business, John Wiley & Sons, 2010.

[37] R. G. Staudte and S. J. Sheather, Robust Estimation & Testing, Toronto: John Wiley & Sons, 2011.

[38] P. Mell and Timothy Grance, "The NIST Definition of Cloud Computing," National Institute of Standards and Technology. Special Publication 800-145, 2011.

This document was typed in using Microsoft Word by María Fernanda Márquez Alderete.