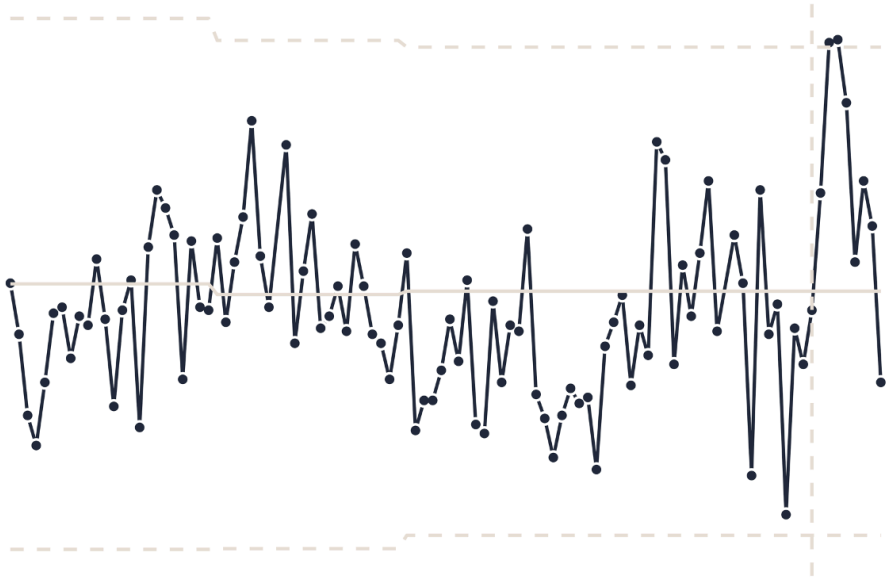
STATISTICAL AND PREDICTIVE

PROCESS MONITORING

LEO C.E. HUBERTS



MONITORING COMPLEX PROCESSES
IN THE AGE OF BIG DATA

Statistical and Predictive Process Monitoring

Monitoring Complex Processes in the Age of Big Data

Leo C.E. Huberts

Statistical and Predictive Process Monitoring

Monitoring Complex Processes in the Age of Big Data

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen
op donderdag 22 april 2021, te 13.00 uur

door

Leonardus Clemens Elisabeth Huberts
geboren te Amsterdam

Promotiecommissie

Promotor

Prof. dr. R.J.M.M. Does	Universiteit van Amsterdam

Co-promotor

Dr. M. Schoonhoven	Universiteit van Amsterdam

Overige leden

Prof. dr. H.P. Boswijk	Universiteit van Amsterdam
Prof. dr. ir. D. den Hertog	Universiteit van Amsterdam
Prof. dr. C.G.H. Diks	Universiteit van Amsterdam
Prof. dr. J. A. Dos Santos Gromicho	Universiteit van Amsterdam
Prof. dr. M. Salomon	Universiteit van Amsterdam
Prof. dr. G.G. Vining	Virginia Polytechnic Institute and State University
Dr. I.M. Zwetsloot	City University of Hong Kong

Faculteit Economie en Bedrijfskunde

To my parents Carla & Leo and my brother Krystan

Contents

1 Introduction **1**

1.1 What is SPM? 2

1.1.1 Control Chart Design 3

1.1.2 Control Chart Performance 5

1.2 What is PPM? 6

1.3 Outline and Scientific Contribution 7

I Statistical Process Monitoring **11**

2 Big Data and the Central Limit Theorem **13**

2.1 Motivation 13

2.2 The Classical Shewhart Control Chart 14

2.3 The Distribution of the Sample Mean 15

2.3.1 The Convolutions 16

2.3.2 Accuracy of the Approximated Distributions 17

2.4 Evaluation of the Central Limit Theorem 18

2.5 Control Chart Performance 18

2.5.1 Simulation Procedure 18

2.5.2 Unconditional Performance 19

2.5.3 Conditional Performance 20

2.6 Concluding Remarks 21

3 Continuously Updating Control Charts **27**

3.1 Motivation . . . 27
3.2 Control Chart Updating . . . 28
3.3 Simulation Scenarios . . . 29
3.4 Simulation Procedure . . . 31
3.5 Performance During Updating . . . 33
3.5.1 Shewhart . . . 33
3.5.2 CUSUM and EWMA . . . 34
3.6 Performance after Updating . . . 34
3.6.1 Shewhart . . . 34
3.6.2 CUSUM and EWMA . . . 37
3.7 Concluding Remarks . . . 37

4 Delayed Updating of Control Charts **43**

4.1 Motivation . . . 43
4.2 Updating the Control Chart Limits . . . 44
4.2.1 Unconditional Expectation . . . 46
4.2.2 Conditional Expectation . . . 46
4.2.3 The Updating Parameters . . . 47
4.3 Performance . . . 48
4.4 Improvements . . . 52
4.4.1 Signal Behavior . . . 54
4.5 Case Study Using COVID-19 Data . . . 54
4.6 Concluding Remarks . . . 56
4.7 Appendices . . . 58
4.7.A Expectation - Unconditional . . . 58
4.7.B Expectation of Sum - Conditional . . . 60

II Predictive Process Monitoring **65**

5 Boosted Predictive Process Monitoring in Mental Health **67**

5.1 Motivation . . . 67

5.2 Problem Description 69
5.2.1 The Mental Healthcare System 70
5.2.2 Data Description 70
5.2.3 The Definition of Crisis 71
5.3 Predictive Model 72
5.3.1 Regression 73
5.3.2 Machine Learning 75
5.3.3 Estimation 78
5.3.4 Results 78
5.4 Monitoring 79
5.4.1 Tuning Procedure 80
5.4.2 Results 82
5.5 Concluding Remarks 84

6 Multilevel Predictive Process Monitoring in Education 87
6.1 Motivation 87
6.1.1 Statistical Process Monitoring 88
6.1.2 Predictive Monitoring 89
6.2 Problem Description 90
6.2.1 Student Performance Literature 90
6.2.2 The Dutch High School System 93
6.2.3 Data Set 94
6.2.4 Determinants of Student Performance 95
6.3 Hierarchical Model 96
6.3.1 The Model 96
6.3.2 Estimation 98
6.3.3 Recurrent Neural Network 99
6.3.4 Results 100
6.4 Monitoring Student Performance 101
6.4.1 Statistical Process Monitoring 102
6.4.2 Predictive Monitoring 106
6.5 Concluding Remarks 111

6.5.1 What Determines Student Performance? 111
6.5.2 Statistical Process Monitoring 113
6.5.3 Predictive Monitoring 113
6.6 Appendix 114
6.6.A Predictive Distribution 114
6.6.B Prior Distributions 115
6.6.C Full Conditional Distributions 116

7 Summary **119**

Bibliography **121**

Acknowledgements **135**

About the Author **138**

Chapter 1

Introduction

The size and frequency of available data in organizations have greatly increased in the past decades. Concurrently, the available computing power has expanded exponentially, which has enabled significant leaps in artificial intelligence and machine learning. These advances are changing the requirements of (statistical) process monitoring. Small sample sizes are less common and a higher frequency of data collection requires flexibility in monitoring procedures.

Process monitoring tracks process data streams and signals in case of a high probability of an unwanted outcome. In Statistical Process Monitoring (SPM), this amounts to discerning special-cause from common-cause variation. Common-cause variation consists of process fluctuation inherent to the design of the process. A process affected by only common-cause variation produces stable outputs. Special-cause variation (temporarily) changes the distribution of the output. Examples of special-cause variation are a ruptured fuel pipe that increases the basic fuel consumption of a car, a traumatic event that impacts a child's performance at school, and unexpected unemployment impacting an individual's mental health. In SPM, the variation of process indicators is monitored using control charts.

Also, the future state of a process can be monitored. In this case, the control chart monitors predictions that can be based on a variety of data sources and novel statistical and machine learning techniques. The control chart gives a signal if the predicted value exceeds a predetermined threshold. For example, monitoring the

probability of success for high school students or monitoring the probability of having an imminent mental health crisis.

In this thesis, we investigate the possibilities offered by the increase in computing power and the size and frequency of data for both statistical and predictive process monitoring. The first part of this thesis will focus on SPM. We will investigate the use of the Central Limit Theorem (CLT) in monitoring subsample means. The tail behavior of the process quality statistic is most important when using a control chart to monitor. We thus investigate the tail behavior when applying the CLT and how the performance of the most commonly-used control chart is affected. We then consider updating the initial control chart parameter estimates using data from the monitoring phase. This can improve the monitoring performance and flexibility for high-frequency processes. In the second part of the thesis, we focus on Predictive Process Monitoring (PPM). Motivated by a real-world application in education, we introduce the use of hierarchical Bayesian regression models in PPM. Furthermore, machine learning methods for prediction are investigated and we integrate gradient boosting in a process monitoring framework which we apply using a large and unique data set on mental health.

Signaling as early as possible can be imperative in taking preventive measures in sectors such as healthcare, education, manufacturing, maintenance, and more. It can improve the quality of products and services. The next two sections will introduce the two parts of this thesis, SPM and PPM respectively.

1.1 What is SPM?

SPM provides techniques to monitor a process in real time. One of these techniques, the control chart, is used to distinguish common-cause from special-cause variation in a process. A process that solely exhibits common-cause variation is called in control. Special-cause variation causes an out-of-control process. The control chart is designed to detect such out-of-control situations.

A wide range of charts has been developed. The Shewhart, Cumulative Sum (CUSUM), and Exponentially Weighted Moving Average (EWMA) control charts, introduced by Shewhart (1926), Page (1954) and Roberts (1959), respectively, are the

most commonly-used charts in practice. These three charts were developed to detect changes in the underlying process, often called assignable or special-cause variation. The Shewhart chart is simple to interpret and implement and is capable of quickly detecting large shifts in the mean of the process quality indicator. The CUSUM and EWMA charts are a bit harder to interpret, as both incorporate previous observations in their test statistic. These two charts are generally better at detecting small shifts in the mean (see Vera do Carmo et al., 2004, for a comparison).

All three mentioned charts have parameters that need to be estimated in practice. This causes uncertainty in the charts' performances. The effects of this uncertainty have been widely researched in recent years and solutions have been proposed to deal with this uncertainty. Jensen et al. (2006) and Psarakis et al. (2014) conducted literature reviews on the effects of parameter estimation on control chart performance and identified directions for future research. Recently, several researchers have proposed adjusted control chart designs based on guaranteed in-control performance (see e.g. Gandy & Kvaløy, 2013; Saleh et al., 2015, 2016; Goedhart et al., 2017a,b; Zwetsloot & Ajadi, 2019; Diko et al., 2019).

1.1.1 Control Chart Design

This section gives the relevant control chart designs. Let X_{ij} denote the j-th observation in sample i ($i = 1, 2, ...$ and $j = 1, 2, ...n$ with n the sample size), and let X_i denote the vector containing the n observations of sample i. Further, let m_I represent the number of Phase I samples for initial parameter estimation, and X_I equal the total Phase I data. We assume that the m_I samples in Phase I are in control (in this stage the practitioner should determine if the process is in control). Further, we assume that the observations X_{ij} in the first m_I samples are independent and identically $N(\mu, \sigma^2)$ distributed.

For each control chart type, μ and σ have to be estimated, and we use the same estimators for each chart.

The parameter μ is estimated by

$$\overline{\overline{X}} = \frac{1}{m_I} \sum_{i=1}^{m_I} \left(\frac{1}{n} \sum_{j=1}^{n} X_{ij} \right). \tag{1.1}$$

Further, σ is estimated by

$$\tilde{S} = \left(\frac{1}{m_I} \sum_{i=1}^{m_I} S_i^2 \right)^{1/2} / c_4(m_I(n-1)+1), \tag{1.2}$$

where S_i is the i-th sample standard deviation defined by

$$S_i = \left(\frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2 \right)^{1/2}$$

and

$$c_4(x) = \left(\frac{2}{x-1} \right)^{1/2} \frac{\Gamma(x/2)}{\Gamma((x-1)/2)},$$

where $\Gamma()$ is the Gamma function. The choice of the estimator of the standard deviation of the sample means is based on Cryer & Ryan (1990). The following sections outline the Shewhart, CUSUM, and EWMA control chart designs.

The estimated Shewhart control limits based on the samples in X_I, used for monitoring are given by

$$\begin{aligned} \widehat{UCL} &= \overline{\overline{X}} + L_s \tilde{S}/\sqrt{n}, \\ \widehat{LCL} &= \overline{\overline{X}} - L_s \tilde{S}/\sqrt{n}, \end{aligned} \tag{1.3}$$

where $\overline{\overline{X}}$ is given by (1.1) and $\tilde{S}$ by (1.2), while L_s is a positive constant for the Shewhart control chart, depending on m_I, n, the expected false alarm probability and distribution of the estimates. The Shewhart control chart signals if $\bar{X}_i$ for $i = m_I + 1, m_I + 2, ...$, is larger than $\widehat{UCL}$ or smaller than $\widehat{LCL}$.

The two-sided CUSUM control chart uses the cumulative sum of observations to monitor the process. The upper and lower statistics are calculated by

$$C_i^+ = max(0, C_{i-1}^+ + \frac{\bar{X}_i - \overline{\overline{X}}}{\tilde{S}} - k) \tag{1.4}$$

$$C_i^- = min(0, C_{i-1}^- + \frac{\bar{X}_i - \overline{\overline{X}}}{\tilde{S}} + k), \tag{1.5}$$

with the chart parameter $k \geqslant 0$ and $C_0^+ = C_0^- = 0$. This CUSUM chart signals if either $C_i^- < -L_c$ or $C_i^+ > L_c$ for $i = m_I + 1, m_I + 2, ...$, where the critical value L_c is a positive constant for the CUSUM control chart depending on k, the desired expected false alarm probability and distribution of the estimates.

The EWMA control chart is an extension of the CUSUM chart and a generalization of the Shewhart chart, by adding weights to the cumulative sum of observations. The EWMA statistic is defined as

$$Z_i = \lambda \bar{X}_i + (1-\lambda) Z_{i-1} \tag{1.6}$$

for $i = m_I + 1, m_I + 2, ...$, where $0 < \lambda \leqslant 1$ and Z_{m_I} equals the mean estimate $\overline{\overline{X}}$. For $\lambda = 1$ the EWMA control chart is equal to the Shewhart control chart. The EWMA control limits for monitoring the process at time $i = m_I + 1, m_I + 2, ...$ are

$$\begin{aligned} \widehat{UCL}_{i-1} &= \overline{\overline{X}} + L_e \frac{\tilde{S}}{\sqrt{n}} \sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2(i-m_I)}]} \\ \widehat{LCL}_{i-1} &= \overline{\overline{X}} - L_e \frac{\tilde{S}}{\sqrt{n}} \sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2(i-m_I)}]}, \end{aligned} \tag{1.7}$$

where λ and L_e determine the expected false alarm probability. When Z_i falls above (below) $\widehat{UCL}_i$ ($\widehat{LCL}_i$) the process is considered out of control.

1.1.2 Control Chart Performance

The performance of control charts can be studied when the process is in or out of control. The performance of a control chart is generally considered in terms of the in-control False Alarm Rate (FAR) or Average Run Length (ARL). The FAR is the probability of an incorrect control chart signal. The ARL is defined as the average number of observations before the chart signals. A recent development in SPM is to evaluate control chart design on the variation of the in-control ARLs of the individually estimated, also called conditional, control charts. This calls for the use of the Conditional FAR ($CFAR$) and Conditional ARL ($CARL$) performance metrics.

The performance of the control charts depends heavily on the choice of the coefficient L (L_s for the Shewhart, L_c for the CUSUM, and L_e for the EWMA control chart). Classical control chart design would suggest using a value of L that delivers the desired *unconditional* FAR for known parameters. To achieve a desired *conditional* performance, L is constructed to guarantee a certain probability that the in-control FAR will be *at most* the desired FAR value, as proposed by Gandy & Kvaløy (2013) and others (see for example Jones & Steiner, 2012; Saleh et al., 2015, 2016; Goedhart et al., 2017a,b). In that setting, the value L is determined such that

$P(CFAR > FAR_0) = P(CARL < ARL_0) = 1 - \beta$, where $CFAR$ is the in-control conditional FAR, FAR_0 is the desired FAR and β is the accepted (small) probability that the $CFAR$ will be larger than FAR_0. The value of L, given FAR_0 and β, can be determined using analytical or numerical/Monte Carlo procedures.

1.2 What is PPM?

In the second part of this manuscript we discuss PPM. Where SPM and the first part of this thesis considers the current state of a process, PPM is a promising research area that focuses on forecasting potential problems during process execution before they occur (Metzger et al., 2015). Applications have been developed in a wide range of domains, such as manufacturing (Spiewak et al., 2000; Zhou et al., 2005), healthcare (Reifman et al., 2007; Clifton et al., 2013; Luo, 2020), networking (Ali et al., 2012) and business processes (Tax et al., 2017).

Increasingly comprehensive data collection provides more process visibility. Advances in machine learning make use of this increase in detail and frequency of data. Data-driven techniques in this area can be used to improve process quality control by forecasting and monitoring potential process problems. Prediction methods include regression techniques, support vector machines, decision trees, random forest, elastic nets, neural networks, and gradient boosting. For an overview of machine learning methods see Hastie et al. (2009).

Predictive monitoring starts with defining an (unwanted) process outcome. Subsequently, a model is specified for the process. The parameters of the model are then estimated using the available data. When monitoring commences, the estimated parameters are used to generate process predictions. The probability of the defined process outcome is then calculated. If the probability exceeds a predetermined threshold, the procedure signals. The parameters can then be re-estimated and the monitoring continues. Note that machine learning techniques require less formal modeling, but more data and computing power to perform predictions. Furthermore, as with SPM, the threshold to signal will determine the expected FAR.

The performance of a PPM procedure can be evaluated using the precision and

recall metrics. The precision is given by

$$\text{Precision}(C) = \frac{tp(C)}{tp(C) + fp(C)},$$

with $tp(C)$ equal to the number of true positives for threshold C and $fp(C)$ the number of false positives for threshold C. The recall is defined as

$$\text{Recall}(C) = \frac{tp(C)}{tp(C) + fn(C)},$$

where $fn(C)$ equals the number of false negatives for threshold C (Powers, 2011).

1.3 Outline and Scientific Contribution

The first part of this thesis considers SPM and consists of three chapters. In Chapter 2 of this thesis, we investigate the use of large sample sizes to eliminate the distributional assumptions of the process indicators through the CLT. In theory, if the sample is large enough and all the observations have the same distribution with mean μ and finite variance σ^2, for sample mean $\bar{X}_i$, the distribution of $\sqrt{n}(\bar{X}_i - \mu)$ converges to a normal distribution $N(0, \sigma^2)$ (cf. Billingsley, 1995). As process monitoring is concerned with the extremes of the process data, we study the tails of the distribution of the process mean. The distributions of the convolutions of common known non-normal distributions are analyzed. Furthermore, the effect of using the CLT for large samples on the performance of the most used control chart (Shewhart $\bar{X}$) is studied. This chapter has been published under the title "The performance of $\bar{X}$ control charts for large non-normally distributed datasets" in *Quality and Reliability Engineering International*. This paper, Huberts et al. (2018), was joint work with dr. M. Schoonhoven, dr. R. Goedhart, dr. M. D. Diko, and prof. dr. R.J.M.M. Does, in which M. Schoonhoven and R.J.M.M. Does initiated the collaboration and I took the lead in the analyses and writing.

Chapter 3 of this thesis is concerned with the updating of parameters during process monitoring. SPM generally estimates the in-control process distribution in a pre-monitoring phase, which is usually referred to as Phase I. The in-control parameters remain fixed during the monitoring phase. Conversely, the data collected during the monitoring phase could be used to update the parameter estimates. We study how

this affects the performance of control charts for 16 different in- and out-of-control monitoring scenarios. This chapter has been published under the title "The effect of continuously updating control chart limits on control chart performance" in *Quality and Reliability Engineering International*. This article, Huberts et al. (2019), was joint work with dr. M. Schoonhoven and prof. dr. R.J.M.M. Does, in which I took the lead in the analyses and writing.

Furthermore, introducing a delay in updating can prevent out-of-control samples to be included in parameter updates. In Chapter 4, a procedure to introduce such a delay is discussed and we propose some improvements. The method is applied to a COVID-19 related data set. This chapter has been conditionally accepted under the title "Improved control chart performance using cautious parameter learning" in *Computers and Industrial Engineering*. In this study, Huberts et al. (2020b), I took the lead in the analyses, dr. R. Goedhart assisted in the writing and prof. dr. R.J.M.M. Does provided supervision.

The second part of this thesis, introduced in Section 1.2, considers PPM in two chapters including applications. Chapter 5 introduces machine learning for PPM. Many recently developed machine learning techniques can be used for prediction. We introduce a procedure to tune the probability threshold towards a desired FAR in monitoring. Using a unique non-public data set on mental health, we investigate the predictive accuracy of machine learning techniques. The Extreme Gradient Boosting (XGBoost) algorithm is subsequently used to monitor the risk of relapse in people diagnosed with schizophrenia. The procedure can aid healthcare workers in identifying people that are likely to need preventive care. This chapter has been submitted under the title "Predictive monitoring using machine learning algorithms and a real-life example on schizophrenia" to *Quality and Reliability Engineering International*. This study, Huberts et al. (2020a), was combined work with prof. dr. R.J.M.M. Does, dr. B. Ravesteijn, and dr. J. Lokkerbol in which dr. B. Ravesteijn provided access to the data and feedback, dr. J. Lokkerbol and prof. dr. R.J.M.M. Does assisted with the writing and I took the lead.

The final chapter of this thesis introduces multilevel process monitoring. Process data often have some hierarchical structure. Modeling this structure can improve parameter estimates and predictions. Furthermore, using a multilevel model allows

monitoring at the different levels in the hierarchy. We illustrate this approach using high school data. Bayesian hierarchical modeling is combined with SPM techniques and used in a predictive monitoring procedure. The procedure allows early warnings for students that have 'exceptional' performance. Exceptional students can be failing students or students with exceptionally good grades. Based on the predictive monitoring procedure, the school can intervene and offer tutoring to the former and more challenging course work to the latter group. This assists schools in personalizing education and controlling quality. This chapter has been published under the title "Multilevel process monitoring: A case study to predict student success or failure" in the *Journal of Quality Technology*. This paper, Huberts et al. (2020c), was combined work with dr. M. Schoonhoven and prof. dr. R.J.M.M. Does and a Dutch high school in which I took the initiative.

The conclusion will summarize the thesis and offer some views on the current state of SPM and PPM as well as future directions for research in the field.

Part I

Statistical Process Monitoring

Chapter 2

Big Data and the Central Limit Theorem

2.1 Motivation

Due to digitalization, many organizations possess large datasets. Furthermore, measurement data are often not normally distributed. However, when samples are sufficiently large and particular conditions met, the Central Limit Theorem (CLT) dictates that the distribution of the sample means will converge to a normal distribution. In this chapter, we evaluate the tail behavior of the CLT for various distributions and sample sizes, as well as its effects on the performance of a Shewhart control chart for these large non-normally distributed datasets.

Shewhart control charts are commonly used to monitor process data. Typically, the performance of such control charts is heavily dependent on the assumption of normally distributed data. In practice, this assumption is often violated. For example, Alwan & Roberts (1995) analyzed 235 real datasets and concluded that most of these datasets do not meet the assumptions underlying the traditional control charts.

Since recent advances have led to an increase in the amount of available information, one way to work around the violation of the normality assumptions is to gather larger datasets and use subgroup averages instead of individual observations. Because

averages are approximately normally distributed under certain conditions, according to the CLT, this should largely resolve the issue of non-normally distributed data (cf. Billingsley, 1995).

While the approach of using averages instead of individual observations is suitable for many statistical techniques, the major difference with many other statistical techniques is that in Statistical Process Monitoring (SPM) we are interested in the long tail behavior of the distribution. This means that, even when the statistic is almost normally distributed, small deviations at the long tails can lead to bad control chart performance in terms of the FAR and the ARL. In this chapter, we therefore investigate the performance of Shewhart type $\bar{X}$ control charts for large, non-normally distributed datasets using the convolutions of the distributions. To the best of our knowledge, the performance of Shewhart $\bar{X}$ control charts in this setting has not been investigated thus far. The results in this chapter, which are based on Huberts et al. (2018), indicate that the $\bar{X}$ control chart should be applied with caution, even with large sample sizes.

This chapter is structured as follows. In the next section, we briefly describe the model and control charts considered in this chapter. Subsequently, in Section 2.3, the CLT is summarized, followed by the convolutions of various probability distributions. In Section 2.4, we investigate the differences between the normal and non-normal convolutions. Next, Section 2.5 describes the performance of the Shewhart control chart based on large non-normally distributed datasets. Finally, Section 2.6 provides some concluding remarks.

2.2 The Classical Shewhart Control Chart

Due to the increase in data supply and storage, nowadays organizations often possess large datasets. As the CLT states that under certain conditions the sample means are normally distributed when the samples are sufficiently large, we could treat the sample means as individual observations and use a Shewhart control chart as described in Section 1.1.1 for individual observations under normal theory.

In this chapter, we study both the unconditional and conditional performance of the control chart constructed with (1.3) including the newly developed factors, for the

cases where the data are non-normally distributed and various sample sizes (n = 5, 30, 50, 100, 250, 1000). With this model, we can investigate whether the CLT works well and whether the newly developed correction factors apply to large non-normal datasets as well. We consider the normal distribution, the standard uniform distribution, heavy-tailed symmetrical distributions (Student's t_4 and t_{10} and the logistic distribution), and skewed distributions (the lognormal, Gamma$(5, 1)$, Gamma$(\frac{5}{2}, 2)$, which is identical to χ^2_5 and χ^2_{20} distributions).

The distribution of the sample means for any one of these non-normal distributions can be found using the convolution of that non-normal distribution, i.e.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}C_n,$$

where C_n is the convolution of n i.i.d. random variables with distribution F. In the next section we produce the distribution of C_n for the considered non-normal distributions.

2.3 The Distribution of the Sample Mean

Let $X_1, X_2, ..., X_n$ be n i.i.d. observations drawn from F, with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then as n tends to infinity, the random variables $\sqrt{n}(\bar{X} - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$ (cf. Billingsley, 1995), i.e.

$$\sqrt{n}\left(\left(\frac{1}{n}C_n\right) - \mu\right) \xrightarrow{d} N(0, \sigma^2).$$

Hence the asymptotic distribution of the sample means is normal under the above restrictions. The exact distribution for finite values of n can be obtained by evaluating the convolution. To assess the performance of the Shewhart control chart for sample means of non-normally distributed samples, we need the distributional properties of the convolution of these samples: $C_n = \sum_{i=1}^{n} X_i$. The convolutions will allow an investigation of the distribution of the sample means of non-normal distributions and a comparison with the asymptotic normal distribution according to the CLT.

The convolutions are given below; further details on the derivations and approximations are given in Huberts et al. (2018).

2.3.1 The Convolutions

The Normal Distribution

The convolution of i.i.d. normal random variables is just a normal distribution, with mean $n\mu$ and variance $n\sigma^2$

$$C_n \sim N(n\mu, n\sigma^2).$$

The Uniform Distribution

The convolution of i.i.d. standard uniform random variables has an Irwin-Hall (IH) distribution, which has a piecewise polynomial probability density function with parameter n (see Hall, 1927)

$$C_n \sim IH(n).$$

The Student's t_v Distribution with ν Degrees of Freedom

For $\nu = 1$, t_1 is equal to a standard Cauchy distribution and its convolution C_n will have a Cauchy distribution as well (see Blyth, 1986)

$$C_n \sim Cauchy(0, n),$$

where 0 and n denote the location and scale parameters of the Cauchy distribution respectively. Note that the conditions needed to apply the CLT do not hold for this case, as the Cauchy distribution has no finite mean and variance. For $\nu > 1$, we use an approximation based on the numerical inversion of the characteristic function.

The Logistic Distribution

The standardized version of the sum of i.i.d. logistically distributed random variables with $\mu = 0$ and $s = 1$ can be approximated by a Student's t_ν distributed random variable with $\nu = 5n + 4$ degrees of freedom (George & Mudholkar, 1983)

$$C_n \dot{\sim} t_{5n+4}.$$

The Lognormal Distribution

The distribution of the convolution C_n of the lognormal distribution can be approximated using two methods: the Fenton-Wilkinson approximation by Fenton (1960)

or the Pearson IV approximation by Nie & Chen (2007). The performance of the Pearson IV approximation turns out to be more accurate than the Fenton-Wilkinson approximation as it matches two more moments (see Section 2.3.2). In the following, we will use the Pearson IV approximation

$$C_n \dot{\sim} Pearson_{IV}(\lambda, \alpha, m, \nu),$$

with location parameter λ, scale parameter $\alpha > 0$ and shape parameters $m > \frac{1}{2}$, $\nu \neq 0$.

The Gamma $\Gamma(\alpha, \beta)$ Distribution with Parameters α and β

If X_i is Gamma distributed $X_i \sim \Gamma(\alpha, \beta)$, with parameters α and β, then its convolution is Gamma distributed with parameters $n\alpha$ and β

$$C_n \sim \Gamma(n\alpha, \beta).$$

The Chi-squared χ^2_ν Distribution with ν Degrees of Freedom

The convolution distribution of the sum of n i.i.d. chi-squared random variables with ν degrees of freedom is again a chi-squared distribution with $n\nu$ degrees of freedom

$$C_n \sim \chi^2_{n\nu}.$$

2.3.2 Accuracy of the Approximated Distributions

As reported in the previous section, the convolutions of the Student's t_ν with $\nu > 1$, logistic and lognormal distributions have to be approximated. In the graphs in the left column of Figure 2.1, the approximated densities of the convolutions for the t_{10}, t_4, logistic and lognormal distributions are plotted and compared to the empirical distribution based on six million samples. The graphs in the middle and right columns of Figure 2.1 zoom in on the 0.135th and 99.865th percentiles. The graphs show that the approximated t_{10}, t_4 and logistic convolutions are accurate. For the lognormal approximations, we find that the Pearson IV approximation is closer to the empirical distribution than the Fenton-Wilkinson approximation. Thus, we will use that approximation in the following sections.

2.4 Evaluation of the Central Limit Theorem

To investigate the differences between the actual distribution of the sample mean and the appropriate normal distribution, we have plotted both distributions and the tail behaviors. In Figure 2.2 we have used $n = 30$ and $\alpha = 0.0027$ to investigate the tail behaviors (see Huberts et al., 2018, for $n = 5$ and $n = 250$). The graphs on the left give the densities, while the graphs in the middle and on the right zoom in on the 0.135th and 99.865th percentiles.

The graphs show that, for a sample size of $n = 30$, the convolutions of the uniform, t_{10}, and logistic distributions do not deviate much from the normal distribution. The distribution of the t_4 convolution, however, clearly has wider tails than the normal distribution.

The overall distribution of the Gamma convolution is quite close to normal, with Gamma(5,1) closer to normal than Gamma($\frac{5}{2}, 2$) $\sim \chi^2_5$. When we zoom in on the tail behavior, the Gamma distributions show skewed tail behavior with narrower tails on the left and wider tails on the right than the normal distribution.

The χ^2_{20} convolution deviates a little from the normal distribution, but less so than the χ^2_5 convolution.

The lognormal convolution shows the largest difference with the normal distribution. The distribution of the lognormal convolution is still strongly skewed for large values of n ($n = 250$).

2.5 Control Chart Performance

2.5.1 Simulation Procedure

To evaluate the control chart performance, we conduct 10,000 simulation runs for each parameter combination. For each simulation run:

1. A dataset consisting of m_I samples of size n is generated. Based on these data, μ is estimated by $\overline{\overline{X}}$ and σ is estimated by $\tilde{S}$, using (1.1)-(1.2). Next, $\widehat{UCL}$ and $\widehat{LCL}$ can be determined using (1.3). We use L_s as the first control limit coefficient that ensures that the in-control ARL in expectation ($EARL$) is equal to a specified value (ARL_0) (see Goedhart et al., 2016). The second conditional control limit coefficient L_s^c is

based on Goedhart et al. (2017b) and ensures that the probability (P_E) that a design delivers an estimated control chart with an in-control Conditional ARL ($CARL$) lower than a specified value (ARL_0) is at most a specified probability (p).

2. For each dataset, the Conditional False Alarm Rate ($CFAR$) is calculated as $CFAR = 1 - P(\widehat{LCL} < \bar{X} < \widehat{UCL}) = 1 - P(n\widehat{LCL} < C_n < n\widehat{UCL})$ using the convolutions of Section 2.3.1. The $CARL$ is given by $1/CFAR$.

When we perform the above procedure, we end up with 10,000 $CARL$s of individually estimated control charts. When L_s is used, the $EARL$ is estimated by averaging the 10,000 $CARL$s of the simulated control charts. When L_s^c is used, the exceedance probability (P_E) is obtained by determining the percentage of $CARL$s lower than ARL_0. Both the unconditional and conditional results were verified using the empirical distribution of the non-normal distributions.

We expect that the higher ARL_0, the larger the sample size should be to ensure that the performance of the control charts is as desired. This is because the higher these values are, the more our interest moves towards the long tail of the distribution of the sample means, where minor deviations from the normal approximation have more impact on the performance. For this reason, we consider various values for ARL_0, namely 1,000, 370.4, and 100.

Finally, as we expect that the correction factors are more accurate when the sample size (n) is larger, we consider a broad range of values, namely $n = 5, 30, 50, 100, 250, 1000$. For the amount of samples m_I, we take values $m_I = 30, 50, 100, 200$.

2.5.2 Unconditional Performance

In this Section, we present the simulation results of the control charts based on (1.3) and L_s as defined in Goedhart et al. (2016). Table 2.1 presents the results for an ARL_0 equal to 370.4 (see Huberts et al., 2018, for $ARL_0 = 100$ and $ARL_0 = 1000$). The table presents the $EARL$ and 5th, 50th and 95th percentiles of the $CARL$ distribution.

Table 2.1 shows that the larger the sample size (n), the closer the $EARL$ is to its desired value ARL_0 and so the more applicable is the correction factor. Increasing the number of samples (m_I) also reduces the deviation in performance with respect

to the case of normally distributed data, but the impact of m_I is less strong than the impact of n, as was to be expected. Also, the value of ARL_0 is of influence: the higher ARL_0, the larger the sample size should be to obtain a performance that resembles the performance under normality (Huberts et al., 2018). This can be explained as the relative difference between the distributions of the means based on the non-normal and normal distributions is the largest in the tails of the distributions. To give an example, for the case $ARL_0 = 1000$, the t_{10} and logistic distributions require a sample size of 100 or larger to obtain a correct in-control performance with the use of the given correction factors while, for the case $ARL_0 = 100$, a sample size of 30 is sufficient to obtain the correct $EARL$.

As discussed in Section 2.4, the uniform distribution is the only distribution that has a convolution distribution with thinner tails than the normal distribution on both sides. This produces extremely large $EARL$ values for small n. Furthermore, as the uniform distribution is bounded by an interval, conditional control limits have been generated that produce a $CFAR$ of zero for small values of n giving an infinite $CARL$. Table 2.1 shows the number of infinite values we found for the uniform distribution within the second parentheses.

In Section 2.4, we already indicated a large difference between the normal distribution and the distribution of the lognormal convolution and small deviations compared to the t_4, logistic, Gamma$(5, 1)$, Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$, and χ^2_{20} convolutions. The $EARL$ results confirm these hypotheses, as for all values of n and m_I the lognormal $EARL$ values are consistently far below the desired ARL_0, indicating the strong skewness as observed in the analysis of the convolutions.

2.5.3 Conditional Performance

In this section, we present the results of the control charts based on (1.3) with L^c_s such that the probability of having an in-control $CARL$ lower ARL_0 is equal to p (cf. Goedhart et al., 2017b). We set $p = 10\%$. Table 2.2 presents the realized exceedance probabilities P_E for a specified ARL_0 of 1,000, 370.4 and 100. The presents the results for various sample sizes ($n = 5, 30, 50, 100, 250, 1000$), various numbers of samples ($m_I = 30, 50, 100, 200$) and various distributions (normal, uniform, t_{10}, t_4, logistic with $\mu = 0$ and $s = 1$, lognormal with $\mu = 0$ and $\sigma = 1$, Gamma$(5, 1)$,

Gamma($\frac{5}{2}, 2$) $\sim \chi^2_5$ and χ^2_{20}).

As for the unconditional case, the tables show that the larger the sample size (n), the closer P_E is to its desired value p, and so the better the applicability of the control charts. Contrary to the unconditional case, increasing the number of samples (m_I) does not reduce deviation in conditional performance. Also, the value for ARL_0 has an impact: the lower the ARL_0, the closer the control chart performance is to the desired performance. This can be explained by the increase in relative differences further in the tails of the distributions.

The normal approximation is worst in the case of the lognormal distribution, as we see that the deviation of P_E for $p = 10\%$ is the largest. A very large sample size (n) is needed to guarantee a desired conditional performance. In the case of $ARL_0 = 100$, a sample size of 1,000 gives reasonable P_E values, also for the lognormal distribution, while for $ARL_0 = 1,000$ and 370.4 even a sample size of 1,000 is not large enough to ensure the right exceedance probabilities.

Interestingly, increasing m_I actually increases P_E for the non-normal distributions in most situations. For example, the t_4 distribution for $ARL_0 = 370.4$ and $n = 50$ has a P_E of 17.2% for $m_I = 30$. With m_I increased to 200, for t_4 now 40.3% of the $CARL$s are below the desired $ARL_0 = 370.4$. This can be explained by a decrease in parameter estimate variation and thus a decrease in the constant k_c, causing tighter control limits.

2.6 Concluding Remarks

In this chapter, we have studied the applicability of the CLT to large non-normal datasets. According to the CLT, sufficiently large samples should lead to normally distributed sample averages. However, since SPM is concerned with the far tail of the distribution, it was unclear whether the convergence to normality would be sufficient.

In this research, we have thus investigated whether the charting constants that are designed for normally distributed data can also be applied to large, non-normal datasets. In particular, we have applied the Shewhart control chart for individual observations to monitor the sample means of non-normally distributed datasets.

The study demonstrates that the appropriateness of the control charting con-

stants, also for non-normally distributed data, depends on various factors. These factors include the sample size (n), the number of samples (m_I), the specified desired performance of the control chart, and the degree of the deviation from normality. When the deviation from normality is moderate (as is the case for the uniform, t_{10}, logistic, Gamma$(5, 1)$, Gamma$(\frac{5}{2}) \sim \chi^2_5$ and χ^2_{20} distributions), a sample size of 100 is large enough in order to ensure appropriate use of the correction factors.

However, when the deviation from normality is substantial due to heavy tails (t_4) or substantial skewness (lognormal), the correction factors are not applicable even when the sample size (n) is 1000. The implications are especially relevant within the field of SPM, where the estimation of accurate tail behavior is important. The results indicate that the $\bar{X}$ control chart should be used with caution when the data are not normally distributed, even with relatively large datasets.

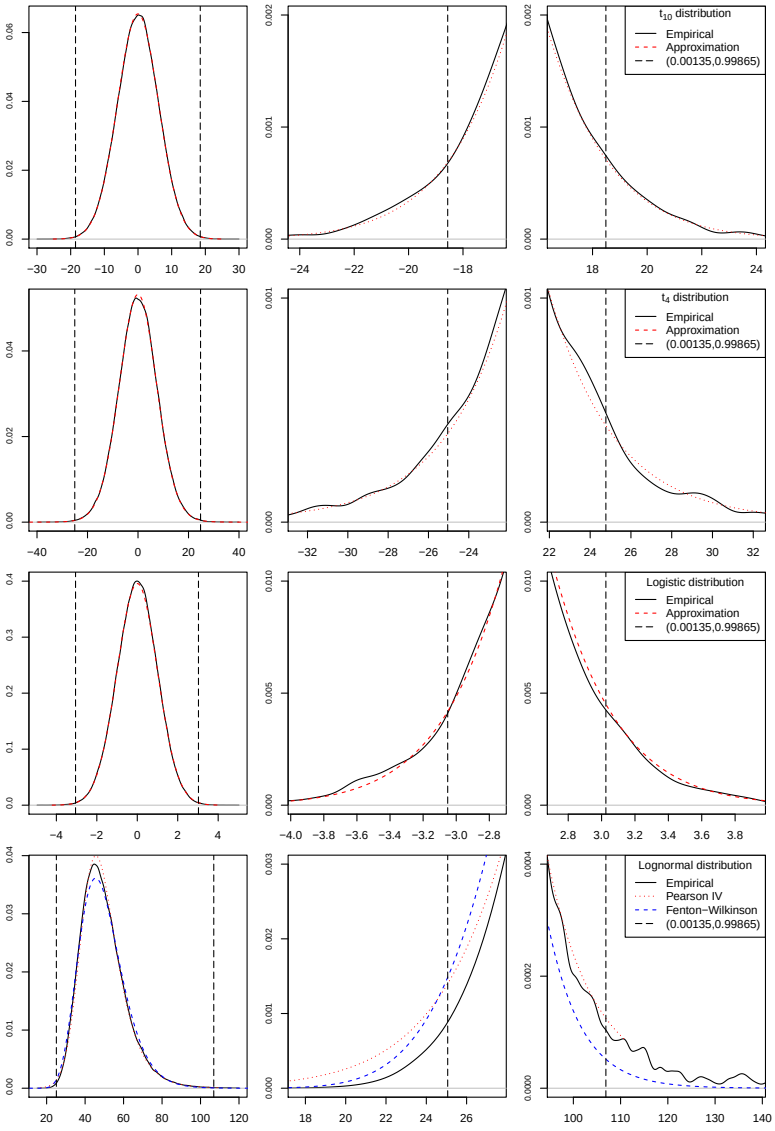


Figure 2.1 – Approximated versus empirical densities for $n = 30$

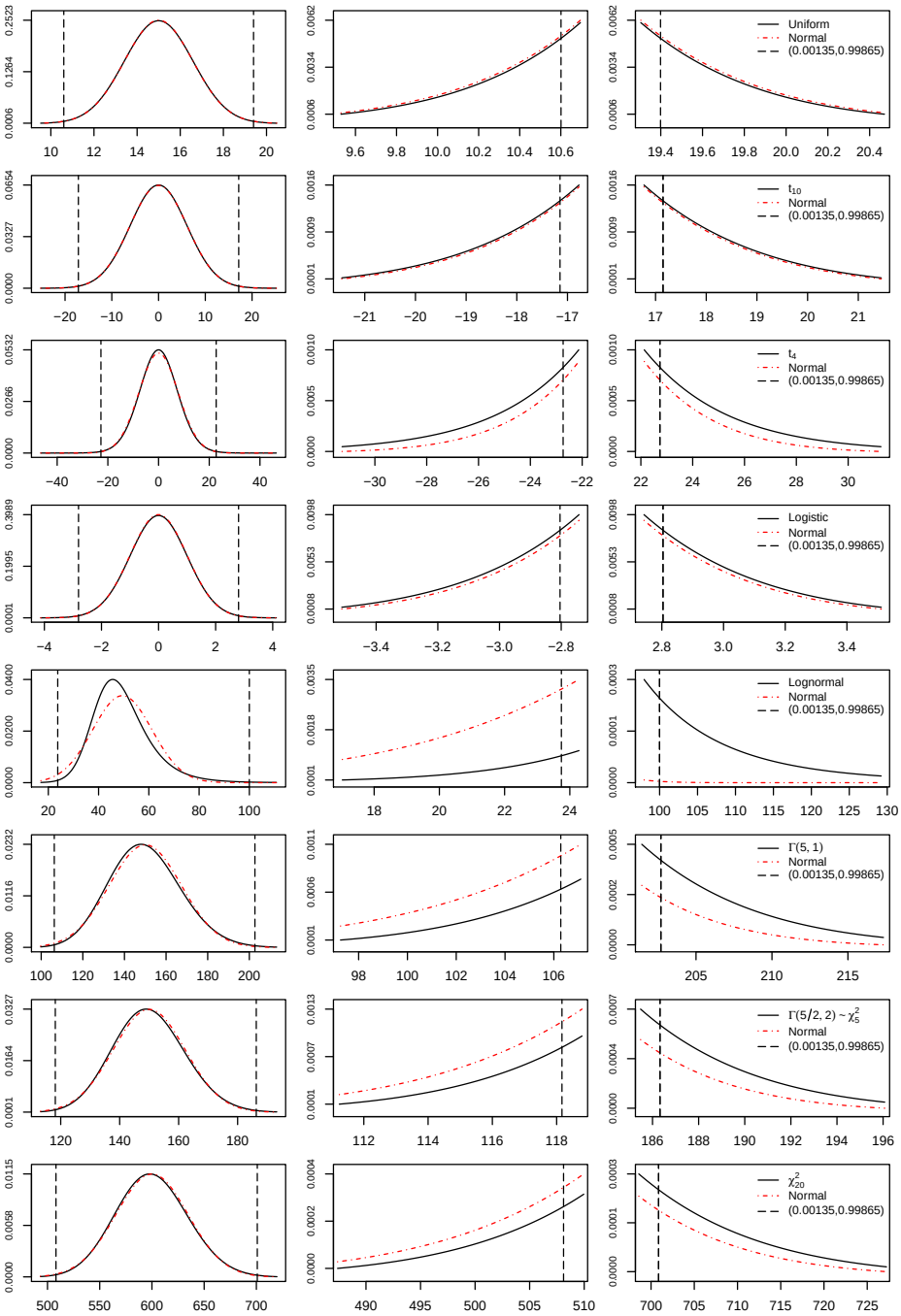


Figure 2.2 – Densities of non-normal convolutions versus normal distributions for $n = 30$ and $\alpha = 0.0027$

Table 2.1 – $EARL$ (5th, 50th, 95th percentile of $CARL$) with $ARL_0 = 370.4$

n	Distribution	$m_I = 30$	$m_I = 50$	$m_I = 100$	$m_I = 200$
5	Normal	378 (32,155,1224)	365 (56,217,1142)	375 (103,283,945)	368 (148,321,737)
	Uniform†	271351 (36,251,7283)(2202)	15844 (74,430,6530)(679)	3061 (157,663,5127)(32)	1292 (279,827,3571)(0)
	t_{10}	224 (28,121,727)	245 (49,165,693)	246 (81,202,561)	252 (118,228,469)
	t_4	128 (22,70,276)	113 (34,85,258)	119 (51,102,226)	121 (66,109,197)
	Logistic	211 (28,117,683)	225 (47,155,623)	230 (80,190,506)	238 (113,217,434)
	Lognormal	75 (14,39,177)	67 (19,45,169)	64 (26,51,132)	63 (33,55,114)
	Gamma$(5,1)$	236 (29,123,786)	229 (47,156,642)	224 (77,185,498)	224 (107,204,410)
	Gamma$(\frac{5}{2},2) \sim \chi^2_5$	171 (26,101,523)	174 (42,124,462)	170 (63,143,363)	167 (84,154,294)
	χ^2_{20}	292 (30,138,966)	284 (51,177,854)	278 (86,222,656)	278 (126,249,529)
30	Normal	344 (32,155,1167)	366 (57,219,1123)	366 (101,282,907)	368 (150,322,746)
	Uniform	406 (31,167,1413)(0)	430 (58,239,1355)(0)	417 (109,310,1073)(0)	418 (166,357,874)(0)
	t_{10}	324 (31,149,1085)	340 (54,206,1047)	340 (94,264,836)	340 (145,299,675)
	t_4	193 (27,111,543)	197 (48,144,510)	209 (75,174,440)	212 (106,193,367)
	Logistic	338 (31,145,1052)	323 (55,204,986)	336 (96,263,817)	338 (145,297,669)
	Lognormal	89 (20,59,230)	92 (29,70,210)	89 (40,77,170)	90 (51,83,149)
	Gamma$(5,1)$	318 (30,147,1059)	328 (56,202,958)	331 (95,262,793)	331 (141,294,645)
	Gamma$(\frac{5}{2},2) \sim \chi^2_5$	302 (31,144,1042)	304 (54,193,906)	302 (90,241,715)	301 (132,266,586)
	χ^2_{20}	349 (31,151,1166)	344 (56,208,1035)	352 (99,271,871)	347 (146,304,690)
50	Normal	336 (31,152,1112)	370 (56,220,1127)	366 (101,282,905)	368 (151,322,744)
	Uniform	373 (31,156,1244)(0)	407 (58,226,1237)(0)	389 (105,295,984)(0)	390 (157,339,788)(0)
	t_{10}	366 (30,148,1096)	352 (57,213,1065)	349 (98,273,867)	349 (148,306,694)
	t_4	213 (29,120,639)	225 (49,159,591)	236 (80,195,500)	243 (116,220,434)
	Logistic	327 (31,149,1127)	344 (55,210,1051)	344 (97,269,835)	352 (149,308,695)
	Lognormal	130 (21,67,272)	104 (32,79,240)	105 (46,91,204)	104 (58,97,172)
	Gamma$(5,1)$	350 (31,147,1187)	342 (56,206,1043)	345 (100,267,841)	348 (145,309,691)
	Gamma$(\frac{5}{2},2) \sim \chi^2_5$	334 (31,145,1119)	324 (55,204,980)	329 (96,256,785)	323 (137,286,630)
	χ^2_{20}	358 (32,152,1197)	352 (55,212,1083)	356 (99,274,876)	358 (147,314,719)
100	Normal	371 (31,153,1233)	367 (56,215,1127)	369 (102,282,927)	367 (149,323,739)
	Uniform	373 (32,156,1247)(0)	371 (56,221,1152)(0)	382 (105,290,948)(0)	384 (156,334,774)(0)
	t_{10}	349 (30,154,1245)	356 (56,215,1057)	364 (101,279,897)	358 (149,312,711)
	t_4	253 (29,129,758)	271 (50,182,755)	279 (88,225,623)	282 (128,253,529)
	Logistic	334 (31,153,1157)	346 (56,209,1073)	361 (101,275,907)	359 (147,317,716)
	Lognormal	147 (24,83,377)	134 (36,98,331)	131 (54,114,261)	133 (72,123,222)
	Gamma$(5,1)$	372 (31,156,1275)	361 (55,211,1067)	359 (102,277,896)	360 (149,317,713)
	Gamma$(\frac{5}{2},2) \sim \chi^2_5$	338 (31,149,1153)	356 (55,212,1088)	348 (98,269,856)	347 (146,304,694)
	χ^2_{20}	366 (31,152,1164)	364 (56,216,1134)	367 (102,281,909)	362 (148,315,726)
250	Normal	372 (32,154,1227)	371 (55,220,1152)	370 (103,283,924)	371 (153,325,740)
	Uniform	364 (31,154,1314)(0)	368 (56,216,1119)(0)	370 (103,286,935)(0)	374 (152,326,759)(0)
	t_{10}	361 (31,154,1232)	357 (56,214,1089)	373 (102,286,918)	365 (151,321,731)
	t_4	295 (30,141,956)	308 (54,197,904)	313 (94,253,735)	317 (137,282,607)
	Logistic	362 (33,154,1201)	357 (54,217,1084)	366 (101,279,923)	365 (150,319,730)
	Lognormal	206 (27,106,580)	187 (44,134,491)	191 (70,159,409)	191 (95,175,334)
	Gamma$(5,1)$	365 (31,160,1258)	362 (58,213,1122)	367 (100,282,905)	367 (150,320,747)
	Gamma$(\frac{5}{2},2) \sim \chi^2_5$	359 (31,157,1220)	364 (55,215,1092)	361 (100,280,889)	359 (149,314,721)
	χ^2_{20}	360 (32,156,1220)	360 (56,217,1120)	364 (98,278,905)	368 (153,320,746)
1000	Normal	376 (32,151,1235)	358 (56,215,1100)	372 (102,285,947)	367 (153,322,730)
	Uniform	356 (32,152,1199)(0)	368 (57,219,1199)(0)	373 (103,287,926)(0)	369 (151,323,745)(0)
	t_{10}	367 (32,154,1220)	368 (56,218,1141)	365 (101,278,914)	369 (152,321,744)
	t_4	323 (32,150,1150)	339 (56,210,1013)	357 (100,276,876)	353 (148,310,706)
	Logistic	347 (31,153,1155)	363 (56,215,1125)	371 (100,286,904)	365 (150,319,735)
	Lognormal	282 (29,135,926)	290 (52,184,845)	288 (88,230,671)	287 (127,258,544)
	Gamma$(5,1)$	367 (31,154,1189)	369 (58,217,1165)	373 (103,281,923)	367 (151,321,747)
	Gamma$(\frac{5}{2},2) \sim \chi^2_5$	341 (31,153,1190)	371 (57,222,1149)	369 (101,284,911)	369 (150,322,750)
	χ^2_{20}	357 (31,155,1225)	361 (56,213,1110)	374 (100,280,959)	367 (149,319,736)

† The amount of infinite $CARL$ values we found is indicated within the second parentheses.

Table 2.2 – P_E with $ARL_0 = 1000$, 370.4, 1000 and $p = 10\%$

		$ARL_0 = 1000$ and $p = 10\%$				$ARL_0 = 370.4$ and $p = 10\%$				$ARL_0 = 100$ and $p = 10\%$			
n	Distribution	$m_I = 30$	$m_I = 50$	$m_I = 100$	$m_I = 200$	$m_I = 30$	$m_I = 50$	$m_I = 100$	$m_I = 200$	$m_I = 30$	$m_I = 50$	$m_I = 100$	$m_I = 200$
5	Normal	8.9	9.5	9.4	9.3	8.9	9.6	9.2	9.8	9.3	9.8	9.2	9.7
	Uniform	2.9	1.6	0.7	0	4	2.6	1.3	0.3	5.9	4.4	3	1.5
	t_{10}	18.8	22.4	31.7	45.2	15.5	18.6	24.8	34.1	13.6	14.3	17.9	22.9
	t_4	83.7	93.5	98.7	99.6	62.1	78	91.6	97.9	35.6	45.3	60.9	76.5
	Logistic	20.2	25.2	35.8	52.6	16.9	21.7	27.6	39.6	14.5	16.7	19.6	26.5
	Lognormal	97.6	99	99.7	99.9	93.4	96.7	98.7	99.7	77	83.6	91.5	95.9
	Gamma$(5, 1)$	28.3	36.8	53.4	73.6	22.6	27.9	39.4	54.4	15.9	17.1	20.3	25.4
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	44	57.9	78.2	94.3	34.6	44.3	63.6	82.3	22.8	25.6	34.6	44.6
	χ^2_{20}	18.6	23.7	31.8	45.5	15.8	18.6	23.6	30.3	12	12.5	14.2	16.1
30	Normal	9	9.1	9.7	9.5	9.3	9.4	9.7	9.6	9.1	9.2	9.5	10.4
	Uniform	8.9	7.9	6.5	5.7	9.3	8.2	7.1	6.4	9.7	8.9	7.8	7.6
	t_{10}	10.6	11.3	12.5	13.3	10	10.3	11.5	12.6	9.4	9.9	11	11.2
	t_4	31.2	42.5	61.8	82.2	22.4	28.1	40.6	57.3	15.5	17.9	22.5	29.5
	Logistic	10.9	11.1	12.1	14.9	10.1	11.1	11.5	14.2	9.9	10.1	10.8	12
	Lognormal	92.3	97.2	99.6	100	80.9	90.7	97.7	99.6	52	63.1	77.8	91.1
	Gamma$(5, 1)$	12.8	14.2	16	19.5	11.5	12	13.4	15.5	9.9	10.2	11.2	11.5
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	15.3	18.2	24.2	32.3	13.1	14.5	19.4	22.2	10.9	12.1	12.6	13.6
	χ^2_{20}	11	11.3	12.5	14.3	10	10.9	11.2	12.3	9.5	9.8	10.2	10.8
50	Normal	9	9.4	9	9.4	8.8	9.5	9.7	9.7	9.2	9.2	10	9.6
	Uniform	9.3	8.9	8.2	7.8	9.5	9.2	8.8	8.5	9.8	9.7	9.5	9
	t_{10}	9.8	10.3	10.9	11.8	9.6	10.4	10.8	10.6	9.3	9.8	10.5	10.4
	t_4	21.7	30	42.6	62.3	17.2	21.3	28.3	40.3	13.8	15.1	17.8	22.1
	Logistic	10.1	10.5	11.5	12.2	9.9	10.3	10.6	11.2	10.1	10	9.8	11.5
	Lognormal	85.8	94.3	99.3	99.9	71	83.8	95.7	99.4	41.3	53	68.6	84.1
	Gamma$(5, 1)$	11.3	11.8	13.4	15.5	9.9	11	12.4	13.6	9.2	9.5	10.1	11.1
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	12.6	14.8	17.9	22.2	11.7	12.6	15.3	16.9	10.4	11.3	11.3	12
	χ^2_{20}	10	10.2	12.1	12.5	10.3	10.1	11	11.2	9.9	9.9	9.5	9.9
100	Normal	9.4	9.2	9.5	9.6	9.1	9.3	9.7	10	9.1	9.2	9.5	9.3
	Uniform	9.2	9.5	9	8.4	9.3	9.7	9.2	8.7	9.5	9.9	9.4	9.1
	t_{10}	9.1	9.7	10.1	10.9	9.1	9.6	10.3	10.8	9.1	9.7	10	9.8
	t_4	16.1	19.4	26.7	37	13.2	15.6	19.6	25.9	11.8	11.5	14.2	16.4
	Logistic	9.5	9.6	10.2	10.9	9.2	9.8	10.5	10.2	9.7	9.8	10.7	9.8
	Lognormal	67.5	82.5	95.9	99.8	52.2	66.4	84.8	96.3	29.2	36.9	49.6	66.5
	Gamma$(5, 1)$	9.5	10.3	11.7	12	9.5	10.3	10	11.1	9.5	9.6	9.7	10.4
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	11	11.5	13.6	16.1	9.9	11.4	12	12.9	9.5	9.9	9.9	10.5
	χ^2_{20}	9.7	9.9	10.5	11.4	9.1	9.4	10.4	9.8	9.7	9.5	9.5	9.8
250	Normal	9.5	9.8	8.8	9.4	8.8	9.5	9.8	9.5	9	9.5	9.3	9.5
	Uniform	9.8	10.1	9.5	9.5	9.9	10.1	9.6	9.7	10	10.2	9.8	9.9
	t_{10}	8.9	9.2	9.4	10.2	9.3	9.3	9.3	10.1	9.2	9.5	9.4	10
	t_4	12.1	13.3	16.2	20	11.6	12.4	13.8	15.8	10.5	10.8	12	13.4
	Logistic	9	9.4	9.5	10	9	9.9	10.1	9.5	9	9	9.3	9.5
	Lognormal	37.4	49.9	70.8	89.2	28	37.2	52.8	71	18	21.4	28.1	35.9
	Gamma$(5, 1)$	9.1	10	10.6	11	9.1	9.6	10	10	9.3	9.9	9.8	9.8
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	9.8	10.6	9.8	11.8	9.6	9.6	10.3	11	8.9	9.5	9.5	10
	χ^2_{20}	9.6	9.8	9.8	10	9.4	9.1	9.9	10.1	9.3	9.4	9.3	10
1000	Normal	8.8	9.4	9.3	9.6	9.7	9	9.5	10.1	9.4	10	9.3	9.4
	Uniform	10.5	8.8	9.1	10.3	10.5	8.9	9.1	"10.3 "	10.6	8.9	9.1	"10.3
"	t_{10}	9.2	9.5	9.7	9.6	8.9	8.9	9.5	9.4	9.2	10	9.3	10.3
	t_4	10	10.5	11.5	12.8	9.2	10.1	10.7	11.5	9.5	10.1	10.3	10.3
	Logistic	9.2	9.2	9.6	10	9.3	9.1	9.3	9.8	8.7	9.4	9.1	10.2
	Lognormal	15.5	19.2	25.1	34.5	13.7	16.6	19.3	24.4	11.6	12.6	14.3	16.2
	Gamma$(5, 1)$	9.2	9.4	9.7	9	9.5	9.2	9.6	10	9	8.8	9.7	9.8
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	8.6	9.4	9.3	10.1	9.3	9.5	9.5	9.8	9	9.8	9.2	9.7
	χ^2_{20}	9.6	9.3	8.7	9.8	8.9	9.2	9.3	9.7	9.4	9.3	9.7	9.3

Chapter 3

Continuously Updating Control Charts

3.1 Motivation

Jensen et al. (2006) surveyed the literature on estimated control charts and identified 13 issues for future research. Issue 6 of Jensen et al. (2006) suggests re-estimating the limits during the monitoring period: *"Related to the previous research question is the effect on control chart properties when the control limits are updated in some future time that is not necessarily during a start-up period. If the process is in control, it would be reasonable to use the data to update control limits during Phase II and not continue to use the original limits indefinitely. It is not clear how control chart performance is impacted, but it seems that making use of earlier Phase II data would lead to better control charts."* This issue is the subject under study in the present chapter. In this context, we should also mention the self-starting Cumulative Sum (CUSUM) and Shewhart control chart designs proposed by Hawkins (1987) and Quesenberry (1991) respectively. These designs can already be used when just a few samples are available. The performance of these charts was studied by Keefe et al. (2015) using a simulation procedure. However, they did not study the effect of out-of-control data and reset the parameters to the initial Phase I estimates when the conditional control chart gives a signal.

In this chapter we study the effect of updating for different scenarios: the updating

data may be in or out of control, signals may or may not be correctly classified (depending on the scenario), and when the control chart signals the parameters are re-estimated, and the updating continues.

The chapter is structured as follows. The next section gives the relevant control chart designs. The following section describes the scenarios that will be included in the simulation procedure, which is described in the subsequent section. Then, the results during and after updating are given and the last section offers conclusions and recommendations. This chapter is based on Huberts et al. (2019).

3.2 Control Chart Updating

This section gives the continuously updating Shewhart, CUSUM, and Exponentially Weighted Moving Average (EWMA) control chart designs. Let m_u be the number of samples within the updating period. The m_u monitoring/updating samples may or may not be out of control. We assume that the observations X_{ij} in the first m_I samples are independent and identically $N(\mu, \sigma^2)$ distributed and that the observations in the next m_u samples are independent and identically $N(\mu + \delta, \sigma^2)$ distributed with probability P (and $N(\mu, \sigma^2)$ distributed with probability $1 - P$), where the values of δ and P depend on the scenario. Let i be the time stamp during monitoring ($i = m_I + 1, m_I + 2, ..., m_I + m_u$) and let X_{i-1}^{ic} denote the samples that are classified as in control up to and including time $i - 1$. The number of samples within X_{i-1}^{ic} is denoted by m_{i-1} (so $m_I \leqslant m_{i-1} \leqslant m_I + m_u$). Similar to (1.1), μ is estimated by

$$\overline{\overline{X}}_{i-1} = \frac{1}{m_{i-1}} \sum_{r=1}^{i-1} \left(\frac{1}{n} \sum_{j=1}^{n} X_{rj} \right) \mathbb{1}_{X_i \in X_{i-1}^{ic}}, \tag{3.1}$$

with $\mathbb{1}$ the indicator function. Further, similar to 1.2, σ is estimated by

$$\tilde{S}_{i-1} = \left(\frac{1}{m_{i-1}} \sum_{r=1}^{i-1} S_r^2 \mathbb{1}_{X_i \in X_{i-1}^{ic}} \right)^{1/2} / c_4(m_{i-1}(n-1)+1), \tag{3.2}$$

where S_i is the i-th sample standard deviation.

For the Shewhart chart, we use $L_s = 3$ which 3 is the traditional constant used for known parameters to ensure that the False Alarm Rate (FAR) of the chart is equal to 0.0027 and the Average Run Length (ARL) of the chart is equal to 370. The reason why we use the traditional constant is that we want to make a comparison with the

CUSUM and EWMA control charts with traditional constants. We have, however, also studied the performance of the updated Shewhart chart with the constants for estimated parameters of Goedhart et al. (2016, 2017b) but the results are similar.

As the Crosier CUSUM control chart outperforms the classical CUSUM chart (see Crosier, 1986), we will apply the Crosier chart design of Section 1.1.1. The Crosier CUSUM control chart is a one-sided version of the classical CUSUM chart (see Crosier, 1986). It is defined as

$$C_i = |V_{i-1} + \widehat{Q_i}|, \tag{3.3}$$

where $\widehat{Q_i} = \frac{\overline{X}_i - \overline{\overline{X}}}{S/\sqrt{n}}$ and $V_{m_I} = 0$. The monitoring statistic then becomes

$$V_i = \begin{cases} 0 & \text{if } C_i \leqslant k \\ (V_{i-1} + \widehat{Q_i})(1 - \frac{k}{C_i}) & \text{if } C_i > k, \end{cases} \tag{3.4}$$

where k is the reference value that determines the point at which V_i accumulates deviations from the target value $\overline{\overline{X}}$ and is commonly set at $k = \frac{1}{2}$. The chart signals if $|V_i| > H$, where H is the CUSUM control chart limit whose value depends on desired chart performance. To achieve an in-control ARL of 370 for $k = \frac{1}{2}$ and known parameters, H should be set at 4.3904 (see also Crosier, 1986).

For the EWMA control chart of Section 1.1.1 we set $\lambda = 0.1$ and $L = 2.703$ because for these settings the in-control ARL is 370 when the distributional parameters are known and the chart is able to detect small mean shifts quickly (cf. Lucas & Saccucci, 1990).

3.3 Simulation Scenarios

In this section we describe the various scenarios included in the simulation. For each scenario and control chart type, we simulate 100,000 conditional control charts. The scenarios that we consider are presented in Table 3.1. The case $m_u = 0$ is used to investigate how control chart performance improves when updating ($m_u > 0$) compared to not updating ($m_u = 0$). We consider three reference scenarios: $m_I = 5, 200, 2000$ all with $m_u = 0$. The reference scenario with $m_I = 2000$ and $m_u = 0$ allows us to determine the difference in performance by starting with a small sample

and then updating (using $m_I + m_u = 2000$ samples in total) compared to directly constructing a control chart based on $m_I = 2000$ samples.

In scenarios 1-4 the monitoring/updating data are in control and m_I is 5 or 200 and m_u is 1995 (for $m_I = 5$) or 1800 (for $m_I = 200$). In scenarios 5-8, 9-12, and 13-16 the monitoring/updating data may be out of control, with varying levels of contamination. The contamination is modeled as follows: each sample has a probability P that the observations are drawn from a $N(\mu + \delta, \sigma^2)$ distribution. We assume that when the process is out of control, it remains out of control until the control chart gives a signal.

Table 3.1 – Scenarios

	m_I	m_u	Phase I: IC or OOC?	Updating: IC or OOC?	Signal: Reason known?
Reference 1	5	0	IC	N/A	N/A
Reference 2	200	0	IC	N/A	N/A
Reference 3	2000	0	IC	N/A	N/A
Scenario 1	5	1995	IC	IC	No
Scenario 2	200	1800	IC	IC	No
Scenario 3	5	1995	IC	IC	Yes
Scenario 4	200	1800	IC	IC	Yes
Scenario 5	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	No
Scenario 6	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	No
Scenario 7	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	Yes
Scenario 8	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	Yes
Scenario 9	5	1995	IC	OOC: $P = 0.01; \delta = 2$	No
Scenario 10	200	1800	IC	OOC: $P = 0.01; \delta = 2$	No
Scenario 11	5	1995	IC	OOC: $P = 0.01; \delta = 2$	Yes
Scenario 12	200	1800	IC	OOC: $P = 0.01; \delta = 2$	Yes
Scenario 13	5	1995	IC	OOC: $P = 0.1; \delta = 3$	No
Scenario 14	200	1800	IC	OOC: $P = 0.1; \delta = 3$	No
Scenario 15	5	1995	IC	OOC: $P = 0.1; \delta = 3$	Yes
Scenario 16	200	1800	IC	OOC: $P = 0.1; \delta = 3$	Yes

Moreover, we investigate the effect of the incorrect classification of a signal. In scenarios 1, 2, 5, 6, 9, 10, 13, and 14 it is assumed that the operator can not classify a signal correctly and therefore relies on the control chart. Thus, samples that give no signal are considered as in control and used for re-estimating the control limits, and data that give a signal are classified as out of control and not used for re-estimation. In these scenarios, it is thus not possible to trace back the start of the out-of-control situation and filter out previous out-of-control samples. On the other hand, in scenarios 3, 4, 7, 8, 11, 12, 15, and 16, we assume that the operator is able to identify the

cause of a signal; so data samples that give a false signal are included and we assume that in this case, it is possible to retrace the start of the out-of-control situation and exclude previous out-of-control samples from estimation.

3.4 Simulation Procedure

Below, we describe the simulation procedure.

Step 1: Generate Conditional Control Chart

In this step we construct the initial conditional control chart based on the Phase I dataset of m_I (5, 200) samples of size 5 ($n = 5$). We estimate μ and σ with the estimators $\overline{\overline{X}}_{m_I}$ and $\tilde{S}_{m_I}$ given by (1.1) and (1.2) respectively, and determine the Shewhart or EWMA control limits using (1.3) or (1.7) respectively. The limit for the CUSUM control chart is 4.3904.

Step 2: Use Conditional Control Chart for Monitoring and Update the Chart

For each $i = m_I + 1, m_I + 2, ..., m_I + m_u$: for scenarios 1-4 draw a sample X_i from $N(\mu, \sigma^2)$ and for scenarios 5-16, when the process is in control, draw a sample with probability $1 - P$ from $N(\mu, \sigma^2)$ and otherwise from $N(\mu + \delta, \sigma^2)$, with the values of P and δ depending on the scenario (see Table 3.1). Calculate the test statistics ($\bar{X}_i$, V_i or Z_i).

When the test statistic falls between $\widehat{UCL}_{i-1}$ and $\widehat{LCL}_{i-1}$, the dataset of classified in-control samples, X_{i-1}^{ic}, is augmented with X_i and denoted by X_i^{ic}. The process parameters are recalculated using (1.1) and (1.2), and the control limits $\widehat{UCL}_i$ and $\widehat{LCL}_i$ are determined with the new parameter estimates.

When the test statistic falls outside the control limits $\widehat{UCL}_{i-1}$ and $\widehat{LCL}_{i-1}$, the next step depends on the scenario. For scenarios 1, 2, 5, 6, 9, 10, 13 and 14 (the *reason unknown* scenarios) the current data sample X_i is considered out of control and therefore not added to the dataset of classified in-control samples. The limits are not recalculated. Thus, $\widehat{UCL}_i = \widehat{UCL}_{i-1}$, $\widehat{LCL}_i = \widehat{LCL}_{i-1}$, $\overline{\overline{X}}_i = \overline{\overline{X}}_{i-1}$ and $\tilde{S}_i = \tilde{S}_{i-1}$. For the CUSUM and EWMA control charts, we set $V_i = 0$ and $Z_i = \overline{\overline{X}}_{m_I}$.

In contrast, when the test statistic falls outside the control limits, but the process

is **in control** and the operator is able to determine the cause of the signal – as is the case in scenarios 3, 4, 7, 8, 11, 12, 15 and 16 – X_i will be used for updating the control chart. The dataset of classified in-control samples, X_{i-1}^{ic}, is augmented with X_i and is denoted by X_i^{ic}. The process parameters are recalculated using (1.1) and (1.2), and the control limits $\widehat{UCL}_i$ and $\widehat{LCL}_i$ are determined with the new parameter estimates.

When the test statistic falls outside the control limits, the process is **out of control** and the operator is able to determine the cause of the signal – as is the case in scenarios 3, 4, 7, 8, 11, 12, 15 and 16 – X_i will **not** be used for updating the control chart. Moreover, previous real out-of-control samples, the number denoted by N with $N \leqslant m_u$, are excluded from X_{i-1}^{ic}. Thus, $X_i^{ic} = X_{i-N}^{ic}$, $\widehat{UCL}_i = \widehat{UCL}_{i-N}$, $\widehat{LCL}_i = \widehat{LCL}_{i-N}$, $\overline{\overline{X}}_i = \overline{\overline{X}}_{i-N}$ and $\tilde{S}_i = \tilde{S}_{i-N}$. For the CUSUM and EWMA control charts, we set $V_i = 0$ and $Z_i = \overline{\overline{X}}_{i-N}$.

To assess the performance *during* updating, we determine for each conditional control chart the true alarm percentage ($CTAP$) and the false alarm percentage ($CFAP$) within each simulation run. Related measures were presented by Fraker et al. (2008), Chakraborti et al. (2009) and Frisén (2009). To this end, for 100,000 simulation runs we count the number of correct signals, out-of-control samples, false alarms and in-control samples.

Step 3: Assess the Conditional Control Chart Performance

To assess the performance of a conditional control chart *during* updating we determine the $CTAP$ and $CFAP$ as follows

$$CTAP = \frac{\#\text{correct signals}}{\#\text{out-of-control samples}},$$

and

$$CFAP = \frac{\#\text{false signals}}{\#\text{in-control samples}}.$$

To assess the conditional performance of a conditional control chart *after* updating we determine the Conditional FAR ($CFAR$) and the Conditional ARL ($CARL$). For the Shewhart chart these values can be obtained by

$$CFAR = \Phi\left(\frac{\widehat{LCL}_i - \mu}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(\frac{\widehat{UCL}_i - \mu}{\sigma/\sqrt{n}}\right), \tag{3.5}$$

and

$$CARL = 1/CFAR. \tag{3.6}$$

For the CUSUM and EWMA control charts, $CFAR$ is assessed by determining the number of false alarms on an interval of 100,000 samples and $CARL$ is determined by the Markov chain approaches given by Hany & Mahmoud (2016) and Saleh et al. (2013).

Step 4: Assess the Overall Control Chart Performance

To assess the unconditional control chart performance *during* updating, we determine the average true alarm percentage ($ATAP$) and the average false alarm percentage ($AFAP$) by averaging the $CTAP$ and $CFAP$ values for the R simulation runs.

The Expected ARL after updating ($EARL$) and the expected FAR after updating ($EFAR$) are determined by averaging the corresponding conditional values obtained in the R simulation runs. Moreover, we determine the 10th and 90th percentiles of the $CARL$ and $CFAR$ values, which are indicated by $CARL_{10}$, $CARL_{90}$, $CFAR_{10}$ and $CFAR_{90}$.

The next two sections present the performance results *during* and *after* updating.

3.5 Performance During Updating

In this section we consider the chart performance during updating for either $m_u = 1995$ or $m_u = 1800$ updates. The $ATAP$ and $AFAP$ values for the three charts are reported in Table 3.2.

3.5.1 Shewhart

The in-control behavior of scenarios 1-4 for the Shewhart chart is as expected. For the small mean deviations ($\delta = 0.5$) of scenarios 5-8, in the reason unknown scenarios (5-6), the $ATAP$ values are smaller than the known parameter detection probability (0.03) and the $AFAP$ values are larger as well. In the reason known scenarios (7-8) the $ATAP$ and $AFAP$ values are very close to the known detection probabilities for the Shewhart chart. For larger deviations $\delta = 2, 3$ in scenarios 9-16 the detection percentages of Table 3.2 show almost perfect detection, as expected for the Shewhart chart when large mean shifts occur.

3.5.2 CUSUM and EWMA

As the results for the CUSUM and EWMA are comparable, we will consider them together in this section. The in-control $AFAP$ performance is around 0.0027. As expected, the CUSUM and EWMA charts are more capable of detecting small mean shifts ($\delta = 0.5$) in scenarios 5-8 than the Shewhart chart. For $\delta = 2$, the $ATAP$ values for both charts are around 0.5-0.6, detecting more than half of the out-of-control observations. Furthermore, the $AFAP$ values for both charts in Table 3.2 are acceptably close to 0.0027 in scenarios 9-12. However, in scenarios 13-16 for large and frequent deviations ($\delta = 3, P = 0.1$), the $ATAP$ values indicate that the CUSUM and EWMA charts still miss a significant portion of the out-of-control observations. The CUSUM (EWMA) chart detects during updating only about 90% (70% to 82%) of the out-of-control observations.

The $ATAP$ values are affected by the effects of (contaminated) parameter estimates. Consider the mean estimates during updating for scenario 13 in Figure 3.1 for the three charts. For the first 100 updates, all three charts show a very inaccurate estimation of the mean due to (mostly) the effects of an inaccurate Phase I mean estimate. Over time, the Shewhart chart improves to a very accurate mean estimate of around 0 as it perfectly detects all out-of-control observations. The CUSUM and EWMA chart converge to higher estimates of the mean, as due to their lower detection probability they include contaminated samples in the mean estimate.

3.6 Performance after Updating

This section presents the performance results of the control charts *after* updating. The tables that correspond to this section are 3.3 (Shewhart), 3.4 (CUSUM) and 3.5 (EWMA).

3.6.1 Shewhart

As we can see in Table 3.3, updating the chart after a limited initial Phase I dataset (scenarios 1-4 and 7-16) can result in a control chart performance that is similar to

Table 3.2 – Performance *during* updating

	m_I	m_u	Phase I: IC or OOC?	Phase II: IC or OOC?	Signal: Reason known?	$ATAP$ ($AFAP$) Shewhart	CUSUM	EWMA
Reference 1	5	0	IC	N/A	N/A	NA (NA)	NA (NA)	NA (NA)
Reference 2	200	0	IC	N/A	N/A	NA (NA)	NA (NA)	NA (NA)
Reference 3	2000	0	IC	N/A	N/A	NA (NA)	NA (NA)	NA (NA)
Scenario 1	5	1995	IC	IC	No	NA (0.0027)	NA (0.0028)	NA (0.0025)
Scenario 2	200	1800	IC	IC	No	NA (0.0027)	NA (0.0027)	NA (0.0026)
Scenario 3	5	1995	IC	IC	Yes	NA (0.0027)	NA (0.0028)	NA (0.0025)
Scenario 4	200	1800	IC	IC	Yes	NA (0.0027)	NA (0.0027)	NA (0.0026)
Scenario 5	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	No	0.0107 (0.0086)	0.1106 (0.0054)	0.0962 (0.0091)
Scenario 6	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	No	0.0167 (0.0044)	0.1199 (0.0029)	0.113 (0.0028)
Scenario 7	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	Yes	0.0225 (0.0027)	0.1263 (0.0026)	0.1143 (0.0024)
Scenario 8	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	Yes	0.0277 (0.0027)	0.13 (0.0026)	0.1245 (0.0024)
Scenario 9	5	1995	IC	OOC: $P = 0.01; \delta = 2$	No	0.9296 (0.0033)	0.5901 (0.0027)	0.5477 (0.0027)
Scenario 10	200	1800	IC	OOC: $P = 0.01; \delta = 2$	No	0.932 (0.0027)	0.593 (0.0027)	0.5564 (0.0025)
Scenario 11	5	1995	IC	OOC: $P = 0.01; \delta = 2$	Yes	0.9297 (0.0027)	0.5993 (0.0026)	0.5638 (0.0024)
Scenario 12	200	1800	IC	OOC: $P = 0.01; \delta = 2$	Yes	0.9309 (0.0027)	0.5995 (0.0026)	0.5681 (0.0024)
Scenario 13	5	1995	IC	OOC: $P = 0.1; \delta = 3$	No	0.9999 (0.0027)	0.8803 (0.0024)	0.6971 (0.0074)
Scenario 14	200	1800	IC	OOC: $P = 0.1; \delta = 3$	No	0.9999 (0.0027)	0.8935 (0.0019)	0.745 (0.003)
Scenario 15	5	1995	IC	OOC: $P = 0.1; \delta = 3$	Yes	0.9999 (0.0027)	0.9095 (0.0018)	0.8191 (0.0015)
Scenario 16	200	1800	IC	OOC: $P = 0.1; \delta = 3$	Yes	0.9999 (0.0027)	0.91 (0.0017)	0.8176 (0.0015)

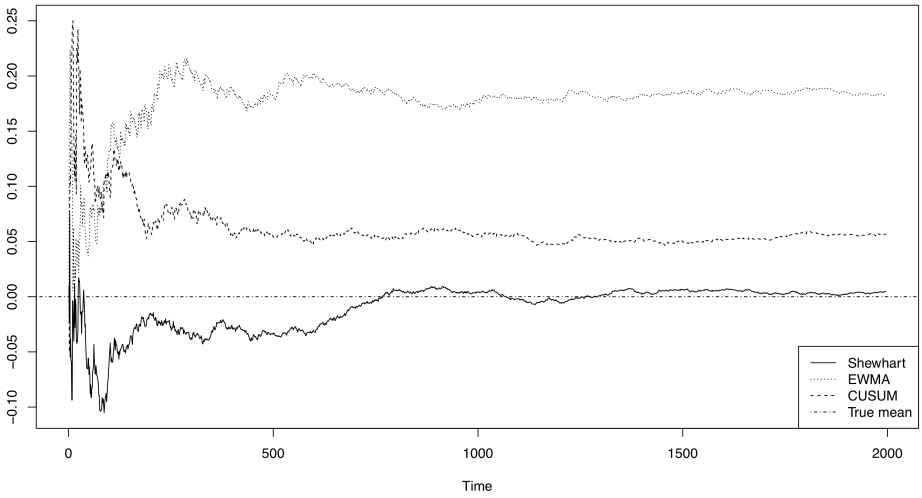


Figure 3.1 – Mean estimates during updates for scenario 13

directly estimating limits from a very large Phase I dataset (reference scenario 3). More specifically, for these scenarios, a chart estimated on $m_I = 5$ or $m_I = 200$ samples, which is then used for monitoring and updating during $m_u = 1995$ or $m_u = 1800$ samples respectively, results in a chart that has the same performance as charts estimated directly from a large Phase I dataset ($m_I = 2000$ and $m_u = 0$). This means that in these scenarios monitoring can start quickly on a limited dataset. It should also be noted that a chart that is initially estimated on a small dataset ($m_I = 5$), followed by monitoring and updating, leads to similar performance as of a chart for which a larger Phase I dataset is used ($m_I = 200$) before the monitoring and updating starts.

We can also conclude that for a more serious level of contamination (scenarios 9-16) it does not matter whether the practitioner classifies signals correctly (scenarios 9, 10, 13, 14) or incorrectly (scenarios 11, 12, 15, 16). The reason is that these out-of-control conditions are quickly detected so that there are fewer out-of-control samples in the estimation dataset. Moreover, excluding from the estimation dataset any samples that give a false alarm, does not result in significantly worse performance, as we have also seen in scenarios 1-4.

When there is a little contamination (scenarios 5-8) performance depends on the operator's ability to classify out-of-control signals correctly. When it is possible to determine the cause of the signal and trace back the start of the out-of-control situation, the resulting control chart performs well. On the other hand, when it is not possible to identify the cause and retrace the start of the signal then many out-of-control samples will be left in the estimation sample (because smaller out-of-control levels are not detected quickly by the Shewhart control chart) resulting in less representative final control limits and worse performance.

3.6.2 CUSUM and EWMA

As the conclusions for the CUSUM (Table 3.4) and EWMA (Table 3.5) control charts are very similar we describe them together.

As with the Shewhart control chart, we can see that depending on the scenario, updating generates the same performance as direct estimation from a large Phase I dataset (reference scenario 3: $m_I = 2000$, $m_u = 0$). Mainly in scenarios where the data are in control (scenarios 1-4) or where the operator is able to identify the cause of a signal and trace back the start (scenarios 7, 8, 11, 12, 15, and 16), control chart performance is similar to the performance of charts estimated on a large Phase I dataset.

When the operator is not able to identify the cause of a signal (and filter out previous out-of-control data samples) the performance of the updated control chart depends on the contamination level. Mainly for major contamination (scenarios 13 and 14), out-of-control signals will not be detected quickly so that – until the chart gives a signal – many of the out-of-control samples are included in the estimation dataset, affecting the control chart limits.

3.7 Concluding Remarks

In the present chapter, we have simulated 16 scenarios, differing in the size of Phase I datasets ($m_I = 5, 200$), in the ability to determine the cause of a signal, as well as the status (in- or out-of-control) of the updating datasets. The charts' performances have been analyzed both during and after (at $i = 2000$) the updating period.

Table 3.3 – Performance results Shewhart control chart

	m_I	m_u	Phase I: IC or OOC?	Phase II: IC or OOC?	Signal: Reason known?	Shewhart $EARL$ ($CARL_{10}$; $CARL_{90}$)	Shewhart $EFAR$ ($CFAR_{10}$; $CFAR_{90}$)
Reference 1	5	0	IC	N/A	N/A	1434 (33; 1897)	0.01170 (0.00053; 0.02989)
Reference 2	200	0	IC	N/A	N/A	373 (264; 497)	0.00285 (0.00201; 0.00379)
Reference 3	2000	0	IC	N/A	N/A	371 (335, 409)	0.00271 (0.00245, 0.00299)
Scenario 1	5	1995	IC	IC	No	371 (335; 408)	0.00272 (0.00245; 0.00299)
Scenario 2	200	1800	IC	IC	No	371 (335; 409)	0.00271 (0.00245; 0.00299)
Scenario 3	5	1995	IC	IC	Yes	371 (334; 408)	0.00271 (0.00245; 0.00299)
Scenario 4	200	1800	IC	IC	Yes	370 (334; 409)	0.00272 (0.00244; 0.00299)
Scenario 5	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	No	143 (51; 270)	0.01009 (0.00371; 0.01978)
Scenario 6	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	No	208 (102; 311)	0.00578 (0.00322; 0.00978)
Scenario 7	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	Yes	369 (324; 419)	0.00285 (0.00239; 0.00309)
Scenario 8	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	Yes	371 (329; 415)	0.00272 (0.00241; 0.00304)
Scenario 9	5	1995	IC	OOC: $P = 0.01; \delta = 2$	No	371 (334; 410)	0.00271 (0.00244; 0.00299)
Scenario 10	200	1800	IC	OOC: $P = 0.01; \delta = 2$	No	371 (335; 410)	0.00271 (0.00244; 0.00299)
Scenario 11	5	1995	IC	OOC: $P = 0.01; \delta = 2$	Yes	371 (335; 408)	0.00271 (0.00245; 0.00299)
Scenario 12	200	1800	IC	OOC: $P = 0.01; \delta = 2$	Yes	370 (334; 407)	0.00272 (0.00246; 0.00299)
Scenario 13	5	1995	IC	OOC: $P = 0.1; \delta = 3$	No	371 (333; 410)	0.00272 (0.00244; 0.00300)
Scenario 14	200	1800	IC	OOC: $P = 0.1; \delta = 3$	No	371 (333; 410)	0.00271 (0.00244; 0.00300)
Scenario 15	5	1995	IC	OOC: $P = 0.1; \delta = 3$	Yes	370 (332; 410)	0.00272 (0.00244; 0.00301)
Scenario 16	200	1800	IC	OOC: $P = 0.1; \delta = 3$	Yes	371 (333; 410)	0.00272 (0.00244; 0.00300)

Table 3.4 – Performance results CUSUM control chart

	m_I	m_u	Phase I: IC or OOC?	Phase II: IC or OOC?	Signal: Reason known?	CUSUM $EARL$ ($CARL_{10}$; $CARL_{90}$)	CUSUM $EFAR$ ($CFAR_{10}$; $CFAR_{90}$)
Reference 1	5	0	IC	N/A	N/A	205 (5; 264)	0.02545 (0.00167; 0.06595)
Reference 2	200	0	IC	N/A	N/A	257 (106; 401)	0.00319 (0.00224; 0.00431)
Reference 3	2000	0	IC	N/A	N/A	347(299,389)	0.00281(0.00259,0.0031)
Scenario 1	5	1995	IC	IC	No	346 (298; 389)	0.00275 (0.00244; 0.00306)
Scenario 2	200	1800	IC	IC	No	346 (298; 387)	0.00274 (0.00243; 0.00306)
Scenario 3	5	1995	IC	IC	Yes	346 (297; 388)	0.00274 (0.00243; 0.00306)
Scenario 4	200	1800	IC	IC	Yes	346 (297; 389)	0.00274 (0.00243; 0.00306)
Scenario 5	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	No	189 (89; 306)	0.00500 (0.00271; 0.00433)
Scenario 6	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	No	226 (130; 332)	0.00314 (0.00263; 0.00373)
Scenario 7	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	Yes	344 (294; 389)	0.00276 (0.00242; 0.00308)
Scenario 8	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	Yes	344 (293; 389)	0.00275 (0.00243; 0.00307)
Scenario 9	5	1995	IC	OOC: $P = 0.01; \delta = 2$	No	309 (225; 376)	0.00283 (0.00249; 0.0032)
Scenario 10	200	1800	IC	OOC: $P = 0.01; \delta = 2$	No	316 (238; 379)	0.00281 (0.00247; 0.00316)
Scenario 11	5	1995	IC	OOC: $P = 0.01; \delta = 2$	Yes	346 (296; 389)	0.00274 (0.00244; 0.00307)
Scenario 12	200	1800	IC	OOC: $P = 0.01; \delta = 2$	Yes	345 (297; 388)	0.00275 (0.00244; 0.00306)
Scenario 13	5	1995	IC	OOC: $P = 0.1; \delta = 3$	No	178 (81; 295)	0.00356 (0.00276; 0.00452)
Scenario 14	200	1800	IC	OOC: $P = 0.1; \delta = 3$	No	219 (121; 329)	0.00317 (0.00265; 0.00378)
Scenario 15	5	1995	IC	OOC: $P = 0.1; \delta = 3$	Yes	344 (292; 390)	0.00275 (0.00242; 0.00308)
Scenario 16	200	1800	IC	OOC: $P = 0.1; \delta = 3$	Yes	344 (292; 389)	0.00274 (0.00242; 0.00307)

Table 3.5 – Performance results EWMA control chart

	m_I	m_u	Phase I: IC or OOC?	Phase II: IC or OOC?	Signal: Reason known?	EWMA $EARL$ ($CARL_{10}$; $CARL_{90}$)	EWMA $EFAR$ ($CFAR_{10}$; $CFAR_{90}$)
Reference 1	5	0	IC	N/A	N/A	156 (5; 235)	0.03491 (0.00477; 0.08170)
Reference 2	200	0	IC	N/A	N/A	248 (96; 396)	0.00324 (0.00225; 0.00442)
Reference 3	2000	0	IC	N/A	N/A	346(292,391)	0.00279 (0.00257,0.00300)
Scenario 1	5	1995	IC	IC	No	345 (291; 390)	0.00272 (0.00242; 0.00303)
Scenario 2	200	1800	IC	IC	No	345 (290; 390)	0.00272 (0.00241; 0.00303)
Scenario 3	5	1995	IC	IC	Yes	345 (289; 390)	0.00271 (0.00241; 0.00303)
Scenario 4	200	1800	IC	IC	Yes	346 (293; 390)	0.00271 (0.00241; 0.00301)
Scenario 5	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	No	155 (40; 279)	0.00803 (0.00278; 0.00695)
Scenario 6	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	No	205 (107; 323)	0.00326 0.00266; 0.00398)
Scenario 7	5	1995	IC	OOC: $P = 0.01; \delta = 0.5$	Yes	343 (289; 391)	0.00281 (0.00241; 0.00305)
Scenario 8	200	1800	IC	OOC: $P = 0.01; \delta = 0.5$	Yes	344 (285; 390)	0.00272 (0.00241; 0.00304)
Scenario 9	5	1995	IC	OOC: $P = 0.01; \delta = 2$	No	286 (188; 371)	0.00291 (0.00250; 0.00330)
Scenario 10	200	1800	IC	OOC: $P = 0.01; \delta = 2$	No	300 (206; 375)	0.00283 (0.00247; 0.00322)
Scenario 11	5	1995	IC	OOC: $P = 0.01; \delta = 2$	Yes	346 (291; 392)	0.00271 (0.00241; 0.00303)
Scenario 12	200	1800	IC	OOC: $P = 0.01; \delta = 2$	Yes	345 (289; 390)	0.00271 (0.00241; 0.00303)
Scenario 13	5	1995	IC	OOC: $P = 0.1; \delta = 3$	No	23 (10; 40)	0.01414 (0.00674; 0.02353)
Scenario 14	200	1800	IC	OOC: $P = 0.1; \delta = 3$	No	44 (22; 72)	0.00756 (0.00476; 0.01089)
Scenario 15	5	1995	IC	OOC: $P = 0.1; \delta = 3$	Yes	342 (281; 390)	0.00272 (0.00240; 0.00304)
Scenario 16	200	1800	IC	OOC: $P = 0.1; \delta = 3$	Yes	343 (284; 391)	0.00272 (0.00240; 0.00304)

The results show improved chart behavior for updated limits when the updating dataset is in control, even when the size of the initial dataset is very small ($m_I = 5$). This holds for all three charts and means that excluding samples that give a false alarm from the estimation dataset does not affect control chart performance for these values of m_I and m_u.

For a low level of contamination (e.g. $\delta = 0.5$), the limits can be updated safely for all three charts as long as the signal reason is known and the out-of-control data can be removed. When the signal reason is unknown, so that the origin of the out-of-control situation can not be retraced, the results show a decline in performance for all three charts. The Shewhart chart is especially vulnerable, as it has the least ability to quickly detect small shifts in the process mean.

For higher contamination levels (e.g. $\delta = 2, 3$), it is safe to update the Shewhart chart control limits even if the signal reason is unknown. This is due to the chart's ability to detect all out-of-control signals, preventing contaminated samples to be included in the data set. For the CUSUM and EWMA charts updating is only safe if the signal reason is known and the origin can be retraced. If this is not the case, the data and evolving statistic will get contaminated and the performance of the CUSUM and EWMA charts is quite poor.

In summary, we recommend updating control limits for the Shewhart, EWMA, and CUSUM charts as long as the reason for out-of-control signals is known and the origin can be retraced. If this is not the case, the best strategy depends on the size of the expected mean deviation. For large deviations, the Shewhart chart is safe to use, but the EWMA and CUSUM charts are not. For smaller deviations, the Shewhart chart fails and the performance of the CUSUM and EWMA charts is better. The next chapter will consider updating the control chart limits with a delay.

Chapter 4

Delayed Updating of Control Charts

4.1 Motivation

Parameter estimation is an important topic in Statistical Process Monitoring (SPM), as inaccurate estimates may lead to undesirable control chart performance. Updating the control chart limits during the monitoring period reduces estimation uncertainty. Chapter 3 considered continuously updating control charts. However, when out-of-control situations remain undetected, using the corresponding samples to update the parameter estimates can deteriorate the control chart performance in terms of in-control and out-of-control run lengths. For this reason, updating of parameter estimates should only occur when there is sufficient evidence of an in-control process state.

Control charts are used to monitor quality indicators in industry and services. All three charts of Section 1.1.1 have parameters that need to be estimated in practice. One of the directions that can further improve conditional control chart performance concerns re-estimating the control limits of a control chart. A few studies have appeared, among which Huberts et al. (2019) (Chapter 3) investigating the effects of updating the control limits in various scenarios and Capizzi & Masarotto (2020) proposing a delayed updating procedure for the Shewhart, Cumulative Sum (CUSUM), and Exponentially Weighted Moving Average (EWMA) charts. This chapter builds upon this work as Huberts et al. (2019) showed that updating can improve performance in

certain settings and the proposed delayed updating by Capizzi & Masarotto (2020) is a promising approach.

In this chapter, we evaluate and extend the approach of Capizzi & Masarotto (2020). Depending on the practitioner's needs, important choices have to be made concerning if and when to update. These choices depend on the type of control chart, the sizes of mean deviations deemed important, and the desired False Alarm Rate (FAR_0) and Average Run Length (ARL_0). We propose simple rules for updating parameters that improve the out-of-control performance of the control charts. We show the added value of using these updating rules in practice through a case study using data related to the COVID-19 pandemic.

The chapter is structured as follows. In the next section, we explain the procedure proposed by Capizzi & Masarotto (2020). In the subsequent section, we analyze the performance of this procedure in various settings. Then, the adjustments to the procedure are motivated in Section 4.4 and a practical example is given in Section 4.5. In the last section, we provide some concluding remarks. This chapter has been based on Huberts et al. (2020b).

4.2 Updating the Control Chart Limits

Similar to Capizzi & Masarotto (2020) we consider the Shewhart chart, the two-sided CUSUM chart, and the EWMA control chart designs of section 1.1.1 for individual observations ($n = 1$, so we drop subscript j in X_{ij}). Assume that Phase I runs from $i = 1$ to $i = m_I$. Further, we assume that the Phase I samples are independent and identically $N(\mu, \sigma^2)$ distributed and that the observations in Phase II are independent and identically $N(\mu + \delta\sigma, \sigma^2)$ distributed. Note that Phase II starts at $i = m_I + 1$. As $n = 1$ in contrast to the estimates of Section 1.1.1, the parameter μ up to and including observation i is then estimated by

$$\bar{X}_i = \frac{1}{i} \sum_{r=1}^{i} X_r, \tag{4.1}$$

Further, σ is estimated by

$$S_i = \left(\frac{1}{i-1} \sum_{r=1}^{i} (X_r - \bar{X}_i)^2 \right)^{1/2}. \tag{4.2}$$

In many cases, in-control Phase II data could be used to re-estimate Equations (4.1) and (4.2) and update the control limits. Updating the estimates of the mean and standard deviation could occur after every new observation as with self-starting control charts (Hawkins, 1987; Quesenberry, 1991).

Chapter 3 explored a variety of scenarios and concluded that updating is often a good choice but the type of chart and size of shift (δ) are important. Furthermore, the outcome depends on the ability of practitioners to retrospectively identify out-of-control samples. A potential hazard of an updating scheme is that small shifts may not directly be detected, in which case the corresponding out-of-control observations would be used in the updated in-control parameter estimates. An approach to counter this is to use some form of delay in updating, as was done by Capizzi & Masarotto (2020). The effect of updating the control limits on the control chart performance depends heavily on the parameters, the type of chart, and the choices made by practitioners related to when to update and what data to use for updating.

The approach of Capizzi & Masarotto (2020) was to update Equations (4.1) and (4.2) using a delay. Monitoring begins at time $i = m_I + 1$. The main concept is that at some time $i > m_I$, if it is reasonable to assume the process is in control, the newly collected samples together with the initial m_I Phase I observations can be used to determine new values of Equations (4.1) and (4.2). This reduces the parameter estimation uncertainty. The time at which an update occurs could be fixed beforehand or determined using the collected samples. Capizzi & Masarotto (2020) proposed a solution for the latter option, using the inequality

$$\sum_{h=i-d_i+1}^{i} \left(\frac{X_h - \bar{X}_{i-d_i}}{S_{i-d_i}} \right)^2 < Ad_i - B, \tag{4.3}$$

where d_i counts the number of samples from the last update, $d_{m_I+1} = 1$ and A and B are parameters that need to be set. As long as this inequality doesn't hold, d_i (the updating delay) is increased by one (i.e. $d_{i+1} = d_i + 1$). Thus, the right-hand side of the inequality increases by A every i as long as there is no update. Capizzi & Masarotto (2020) proposed using the values $A = 1.5$ and $B = 50$. As we will show in the following section, this procedure can result in a deterioration of out-of-control chart performance. Improvements can be made to the settings, which we will propose later on in this chapter.

4.2.1 Unconditional Expectation

The unconditional expectation of an individual term in the sum on the left-hand side of Equation (4.3) can be shown to be (cf. Appendix 4.7.A)

$$E\left[\left(\frac{X_h - \bar{X}_{i-d_i}}{S_{i-d_i}}\right)^2\right] = \left(\frac{m_I - 1}{m_I - 3}\right)\left(1 + \delta^2 + \frac{1}{m_I}\right) \tag{4.4}$$

for $m_I > 3$. This shows that, in expectation, the left-hand side of Equation (4.3) increases faster with larger values of δ thus confirming that the procedure is less likely to update when a mean shift has occurred. However, when there is no shift ($\delta = 0$) and $m_I = 50$, the expectation equals 1.0634. Therefore, in expectation, for values $A < 1.0634$ the left-hand side of Equation (4.3) increases faster than the right-hand side, preventing updates. For $\delta = 0.5, m_I = 50$ the expectation in Equation (4.4) equals 1.324. Then for $A = 1.5$ as considered by Capizzi & Masarotto (2020), in expectation, the right-hand side of Equation (4.3) grows more quickly than the left-hand side. The setting for B does not affect the growth rate but does determine the delay. A larger value for B will mean the left-hand side of Equation (4.3) will have more 'catching up' to do. The settings for A and B will prove to be very important for chart performance.

4.2.2 Conditional Expectation

We will now consider the conditional expectation of the sum on the left-hand side of Equation (4.3). We only consider the time until the first update, such that $d_i = i - m_I$ until the update is done, so that $\bar{X}_{i-d_i} = \bar{X}_{m_I}$ and $S_{i-d_i} = S_{m_I}$, and such that the inequality (4.3) becomes

$$\sum_{h=m_I+1}^{i}\left(\frac{X_h - \bar{X}_{m_I}}{S_{m_I}}\right)^2 < A(i - m_I) - B. \tag{4.5}$$

In Appendix 4.7.B we show that the expectation of the left-hand side of inequality (4.5), conditional on $\bar{X}_{m_I}$ and S_{m_I}, is equal to

$$E\left[\sum_{h=m_I+1}^{i}\left(\frac{X_h - \bar{X}_{m_I}}{S_{m_I}}\right)^2 \,\middle|\, \bar{X}_{m_I}, S_{m_I}\right] = (i - m_I)\left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta\right)^2\right)\frac{\sigma^2}{S_{m_I}^2}. \tag{4.6}$$

Next, we replace the sum in the left-hand side of inequality (4.5) by its expectation, so that we obtain the inequality

$$(i - m_I)\left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta\right)^2\right)\frac{\sigma^2}{S^2_{m_I}} < A(i - m_I) - B. \tag{4.7}$$

We use this inequality to provide an estimate of the expected time to the first update (ETFU). Since B should be a positive number in this method, note that this inequality will never hold if $\left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta\right)^2\right)\frac{\sigma^2}{S^2_{m_I}} \geqslant A$. If $\left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta\right)^2\right)\frac{\sigma^2}{S^2_{m_I}} < A$, then we can solve the inequality for i and find that

$$i \geqslant m_I + \frac{B}{A - \left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta\right)^2\right)\frac{\sigma^2}{S^2_{m_I}}}. \tag{4.8}$$

Thus, our estimate of ETFU, conditional on $\bar{X}_{m_I}$ and S_{m_I}, is equal to

$$ETFU|\bar{X}_{m_I}, S_{m_I} = \left\lceil \frac{B}{A - \left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta\right)^2\right)\frac{\sigma^2}{S^2_{m_I}}} \right\rceil, \tag{4.9}$$

where $\lceil . \rceil$ represents the ceiling function. The $ETFU$ shows that B and A are important, as well as the shift size δ and of course the parameter estimation error. Although the $ETFU$ itself is an approximation, it provides some useful insight towards the essence of the problem at hand. For example, given $\delta = 0.5, \bar{X}_0 = \mu, S^2_{m_I} = \sigma^2, A = 1.5$, and $B = 50$, the expected first update will occur at $ETFU = 200$. For the Shewhart chart with $ARL_0 = 500$ the Expected ARL ($EARL$) for $\delta = 0.5$ equals 202. This means that, in expectation, this Shewhart chart will update the parameter estimates with out-of-control observations before it is able to signal them.

4.2.3 The Updating Parameters

To analyze the impact of A, B, and δ on the time to update and on the ARL performance we use a Monte Carlo simulation because analytical expressions for the three control charts are infeasible. We do this by determining the probability that a control chart will update the parameter estimates before it produces an out-of-control signal.

We simulate for the Shewhart chart, but the same principle applies to the EWMA and CUSUM charts. We apply the following procedure for $ARL_0 = 200, 370$ and 500:

For $u = 1, 2, ..., 6{,}000$:

1. Simulate a $N(0, 1)$ Phase I sample of size $m_I = 50$ and calculate $X_{m_I,u}$ and $S_{m_I,u}$, which are the the Phase I estimates according to Equations (4.1) and (4.2), respectively, for the u-th simulated sample.

2. Initialize Shewhart control chart u using $A \in \{1, 1.25, 1.5\}$, $B \in \{50, 100, 200\}$, $m = 50$ and the CautiousLearning R-package of Capizzi & Masarotto (2020).

3. Simulate a $N(\delta, 1)$ Phase II sample of size 1,000,000 for a wide range of δ $(0.0, 0.25, 0.5, ..., 2.0)$ and calculate the first update FU_u and first signal FS_u given $\bar{X}_{m_I,u}$ and $S_{m_I,u}$.

4. If $u < 6{,}000$ increase u with 1 and go back to step 1

5. Calculate the percentage of charts that have a first update before first signal as $\frac{\sum_{u=1}^{6,000} I(FU_u < FS_u)}{6{,}000}$ with I the indicator function.

The results of the simulation procedure for $ARL_0 = 370$ and various combinations of δ, A, and B are shown in Figure 4.1. For $ARL_0 = 200$ the percentages of charts that update before signaling are slightly lower and for $ARL_0 = 500$ slightly higher.

Figure 4.1 shows that for values of δ smaller than 1.5, the charts often update using out-of-control observations. For example, given $ARL_0 = 370, A = 1.5, B = 50$ and $\delta = 0.5$ the percentage of Shewhart charts that update before signalling is larger than 60%. This means that there is a substantial risk of using out-of-control observations to update in-control parameter estimates, which may negatively affect control chart performance, as we show in the next section.

4.3 Performance

The previous section has shown that there is a large likelihood of updating control limits using out-of-control samples. The effects on chart performance in terms of in-control and out-of-control $EARL$ are studied in this section.

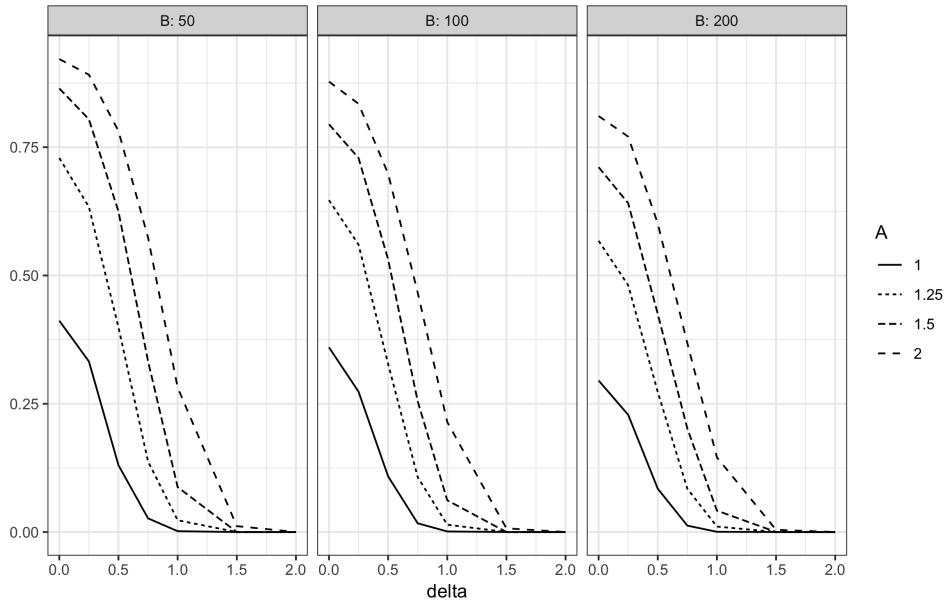


Figure 4.1 – Percentage of control charts with a first update before a signal for $ARL_0 = 370$

We perform a Monte Carlo simulation to assess the effects of the updating parameters (A, B) and the shift size (δ) on the control chart performance. For the Shewhart, EWMA, and CUSUM charts we set $ARL_0 = 370$. We have also analyzed $ARL_0 = 200$ and $ARL_0 = 500$ for which the results were very similar. We let δ vary from 0 to 2 in steps of 0.25, $A = 1, 1.5, 2$, $B = 50, 100, 150, 200$ and include the reference without-updating $EARL$ values ($A = 0$, $B = 0$). For each combination of δ, A, and B we simulate 6,000 Shewhart, EWMA and CUSUM charts using the CautiousLearning R-package (Capizzi & Masarotto, 2020) and calculate the $EARL$ as the average of the run lengths of these 6,000 charts.

The results for $m_I = 50$ are reported in Table 4.1. Since the control charts are designed to provide a guaranteed in-control ($\delta = 0$) performance when parameters are estimated, we focus on the out-of-control ($\delta > 0$) performance here. Note that in the out-of-control situation smaller $EARL$ values are preferred. Therefore, for each out-of-control column, smaller $EARL$ values are indicated by increased green shading and higher $EARL$ values by increased orange shading. The lowest value is printed in bold.

We can evaluate the performance of the different combinations of A and B in the various scenarios. A first observation is that the values chosen by Capizzi & Masarotto (2020), $A = 1.5$ and $B = 50$, are sub-optimal for all cases. The Shewhart chart performs best for larger values of A regardless of δ. It appears that updating the parameter estimates is very important for the Shewhart control chart performance in this situation. For the EWMA and CUSUM charts the optimal parameters are more mixed but for small values of $\delta = (0.25, 0.5)$ it is best to update quickly using a high value for A and small for B. For larger values of δ, better results are achieved for smaller values of A and higher values of B.

Table 4.1 – *EARL*s for 6,000 simulated control charts and $m_I = 50$ in-control Phase I samples

	m_I	50																						
	δ	0	0.25	0.5	0.75	1	1.5	2	0	0.25	0.5	0.75	1	1.5	2	0	0.25	0.5	0.75	1	1.5	2
A	**B**	Shewhart							EWMA							CUSUM						
0	0	7388	6321	3242	1178	527	103	28	9528	4018	535	78	25	8	5	12572	6878	1341	205	43	9	5
1	50	1255	1197	1160	827	458	98	31	1420	858	276	73	24	8	5	1431	1020	457	139	39	9	5
	100	1417	1358	1197	828	480	103	26	1598	996	295	**69**	24	8	5	1623	1136	478	137	39	9	5
	200	1692	1633	1387	863	451	110	29	2017	1145	308	71	24	8	5	2000	1366	528	136	38	9	5
1.25	50	726	698	654	553	394	101	28	781	583	278	88	26	8	5	802	680	405	148	42	9	5
	100	835	793	736	622	409	100	27	888	633	273	76	24	8	5	884	730	401	139	39	9	5
	200	981	968	856	672	437	105	29	1060	735	271	76	25	8	5	1081	853	425	**135**	39	9	5
1.5	50	686	659	594	469	320	106	28	677	565	296	97	25	8	5	729	651	437	186	50	9	5
	100	726	704	625	491	338	103	27	740	577	293	82	25	8	5	796	645	436	158	43	9	5
	200	798	782	662	536	347	101	27	856	658	**264**	76	24	8	5	887	759	411	150	**38**	9	5
2	50	585	**569**	**548**	466	300	**70**	**20**	*561*	**487**	269	102	23	**7**	**5**	*591*	**551**	417	213	68	**8**	**4**
	100	615	611	557	**458**	289	77	21	634	511	273	81	23	8	5	657	588	**396**	180	49	8	4
	200	690	676	589	464	**284**	87	25	713	574	268	71	**22**	8	5	753	650	406	155	40	9	5

We conclude that updating the parameter estimates using contaminated samples can have a positive effect on performance. This surprising finding is due to the large parameter estimation uncertainty when $m_I = 50$. To illustrate this, we calculate the unconditional expected time to first update using Equation (4.9). We then compare the estimated upper control limit values using only the m_I samples of Phase I to the estimated upper control limit when using $m_I + ETFU_{\delta,m_I}$ samples. The latter updates the control limits using contaminated Phase II data.

Consider the simulated Shewhart Control Limits in Figure 4.2. In green the control limits resulting from a Phase I sample size $m_I = 50$ are shown. In orange, the control limits are depicted that result from the updating procedure using contaminated Phase II data with $\delta = 0.25$. The expected time to first update for $\delta = 0.25$ and $m_I = 50$ equals $ETFU_{\delta=0.25,m_I=50} = 115$ samples together (e.g., 165 in total). The green histogram thus represents control limits based on $m_I = 50$ in-control samples.

The orange histogram depicts the limits calculated using $m_I = 50$ in-control and $ETFU_{\delta=0.25,m_I=50} = 115$ out-of-control Phase II samples. The updated distribution of control limits in orange is more narrow due the updated parameter estimates. A small bias has been introduced, as Phase II samples with mean deviation $\delta = 0.25$ have been included in the parameter estimates. However, the updated limits are on average still more accurate than the original Phase I control limits. The reduction in parameter uncertainty outweighs the small bias that is introduced. This is because the value of L_s (cf. Section 1.1.1) required to guarantee a minimum in-control performance will be smaller when more observations are available. In particular, for the non-updated limits we have $L_s = 3.61$ for $m_I = 50$, while for the updated limits we have $L_s = 3.26$ when using estimates based on 165 observations (cf. the CautiousLearning R-package by Capizzi & Masarotto, 2020). As a consequence, even though a positive bias is introduced in the estimate of the mean, the estimated control limits will move closer towards $\bar{X}_i$ in this situation.

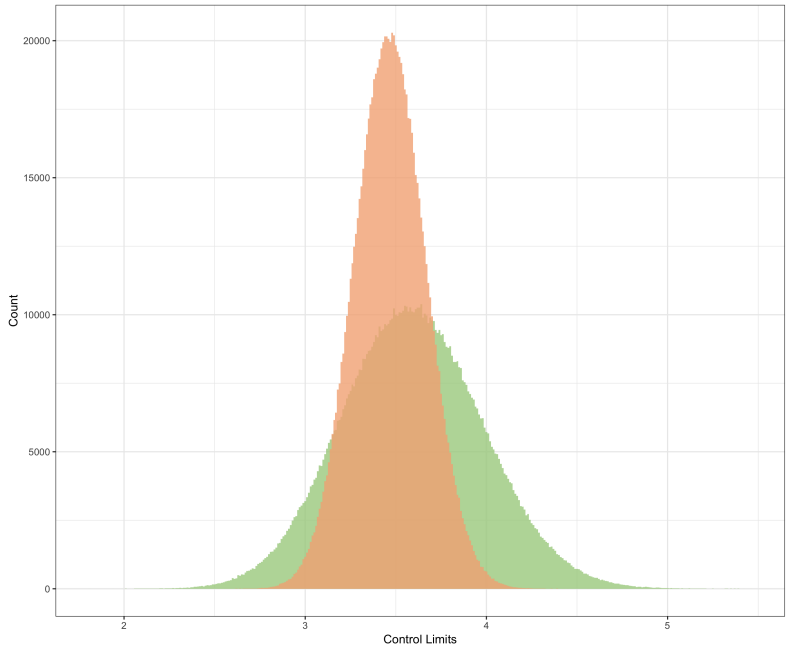


Figure 4.2 – Histograms of 10 million simulated Shewhart control limits based on $m_I = 50$ (green) and updated control limits using using an additional 115 contaminated observations (orange).

We have repeated the Monte Carlo simulation of Table 4.1 for larger Phase I sample sizes $m_I = (250, 500)$. The results for $m_I = 250$ are reported in Table 4.2, and for $m_I = 500$ in Table 4.3. Consider Table 4.2 with $m_I = 250$. Compared to Table 4.1, the parameter estimation error is smaller. For the smallest $\delta = 0.25$, the Shewhart chart should still update quickly using parameters $A = 2$ and $B = 50$. For values of $\delta > 0.5$ this is not the case, $A = 1$ and $B = 200$ give good results. The EWMA and CUSUM charts show a similar pattern for low A values. The CUSUM does require a lower value of B for small δ. Table 4.3 shows the results when $m_I = 500$. In this case the Phase I sample size is larger still and hence parameter estimation is more accurate. The table clearly shows that $A = 1$ or $A = 1.25$ generally performs well. This means updating very slowly or not at all. For the Shewhart chart with $\delta = 0.75$ the best performing chart is the non-updating chart $A = 0, B = 0$. Note that for large shifts ($\delta = 1.5, 2$), for almost all charts and all m_I, setting $A = 2$ and $B = 50$ achieves the optimal $EARL$.

Table 4.2 – $EARL$s for 6,000 simulated control charts and $m_I = 250$ in-control Phase I samples

	m_I	250																						
	δ	0	0.25	0.5	0.75	1	1.5	2	0	0.25	0.5	0.75	1	1.5	2	0	0.25	0.5	0.75	1	1.5	2
A	**B**	Shewhart							EWMA							CUSUM						
0	0	895	636	330	163	84	25	9	821	252	58	23	12	6	4	957	428	121	41	18	6	4
1	50	659	541	324	163	81	24	9	624	247	60	22	12	6	4	668	**396**	126	40	18	6	4
	100	715	572	331	162	84	25	9	675	252	**57**	22	12	6	4	740	401	**120**	41	17	6	4
	200	779	603	**323**	**160**	81	25	9	717	**247**	58	22	12	6	4	789	430	123	40	18	6	4
1.25	50	569	516	358	170	84	25	9	538	266	60	22	12	6	4	583	415	142	42	17	6	4
	100	582	526	339	166	81	24	9	569	255	58	22	12	6	4	602	403	132	40	18	6	4
	200	649	547	327	163	84	26	10	619	251	57	22	12	6	4	659	400	122	41	17	6	4
1.5	50	533	508	389	190	82	24	9	507	277	64	21	12	6	4	537	436	176	44	17	6	4
	100	566	519	371	170	**78**	24	9	538	274	58	22	12	6	4	572	420	146	40	**17**	6	4
	200	590	531	338	163	84	24	9	568	250	58	22	12	6	4	599	398	132	40	17	6	4
2	50	528	**492**	453	317	128	**23**	9	*488*	288	71	**21**	**11**	**6**	4	*526*	429	208	54	17	6	4
	100	547	503	420	262	98	24	**9**	500	273	62	21	11	6	**4**	541	418	180	46	17	**6**	**4**
	200	554	510	390	205	82	24	9	539	271	59	21	12	6	4	563	418	158	**39**	17	6	4

4.4 Improvements

In this section, we discuss the optimal settings when (cautiously) updating the Shewhart, EWMA, and CUSUM charts. As shown in the previous section these settings depend on the number of Phase I samples m_I, the desired ARL_0, and the mean shift δ.

The first general result is that the EWMA chart given the chosen parameter settings obtains the smallest out-of-control $EARL$ values for all combinations of δ and

Table 4.3 – *EARL*s for 6,000 simulated control charts and $m_I = 500$ in-control Phase I samples

	m_I	500																				
	δ	0	0.25	0.5	0.75	1	1.5	2	0	0.25	0.5	0.75	1	1.5	2	0	0.25	0.5	0.75	1	1.5	2
A	**B**	Shewhart							EWMA							CUSUM						
0	0	627	474	247	**117**	65	20	8	572	183	47	19	11	6	4	640	322	97	35	15	6	4
1	50	572	449	254	130	65	21	8	542	180	49	20	11	6	4	578	313	102	35	16	6	**3**
	100	581	483	252	123	66	20	8	543	181	47	20	11	6	4	594	**310**	96	35	16	6	4
	200	604	469	**244**	125	64	21	8	570	**179**	**46**	20	11	6	4	628	314	**95**	36	15	6	4
1.25	50	523	467	273	124	65	21	8	497	202	47	19	11	6	4	525	346	110	**34**	16	6	4
	100	516	**444**	261	126	66	21	8	518	193	49	20	11	6	4	540	326	101	36	16	6	4
	200	559	459	250	127	64	21	8	528	187	47	19	11	6	4	556	323	99	35	16	6	4
1.5	50	501	459	326	145	65	20	8	489	218	48	19	11	6	4	513	353	121	35	16	6	4
	100	510	468	306	125	**63**	21	8	492	202	47	**19**	11	6	4	531	350	112	35	**15**	6	4
	200	534	446	262	125	66	20	8	512	194	46	19	11	6	4	540	323	100	35	16	6	4
2	50	498	461	354	220	84	20	**8**	*472*	218	51	19	11	6	**4**	*504*	359	136	37	16	6	3
	100	502	464	345	186	73	20	8	473	213	48	20	**11**	**6**	4	507	350	123	36	15	**6**	4
	200	520	452	320	146	65	**20**	8	483	197	47	20	11	6	4	516	345	113	34	16	6	4

m_I. The second general finding is that for large Phase I sample sizes (i.e. $m_I \geqslant 500$), updating the limits often has negative effects on the control chart performance. Thus, when a sufficient number of observations ($m_I \geqslant 500$) are available, we recommend using the EWMA chart for $\delta \leqslant 1$ and do not update Phase I parameter estimates.

The optimal choice of A and B depends on the value of δ that is important to the practitioner, as well as the number of available in-control Phase I samples m_I. Tables 4.1-4.3 give guidance on choosing the optimal A and B. We have translated the findings from these tables into a few very simple rules of thumb.

1. For large numbers of Phase I samples ($m_I \geqslant 500$) consider if updating is still necessary.

2. For detecting moderate to large shifts ($\delta > 1$) set $A = 2, B = 50$.

3. For detecting small shifts ($\delta \leqslant 1$) use the following rules. For the Shewhart chart set A and B as

$$A = max\left(\left\lceil 2 - \frac{1}{2}|\delta| - \frac{m_I - 50}{250}\right\rceil, 0\right) \tag{4.10}$$

$$B = (m_I + 50)|\delta|. \tag{4.11}$$

For the EWMA and CUSUM charts set A and B as

$$A = max\left(\left\lceil 2 - \frac{4}{3}|\delta| - \frac{m_I - 50}{250}\right\rceil, 0\right) \tag{4.12}$$

$$B = 2(m_I + 50)|\delta|. \tag{4.13}$$

These rules will result in the use of the values of A and B that deliver good out-of-control performance and less unnecessary updating when a large number of Phase I samples are available. Note that these rules apply to the specific settings investigated in this chapter and do not (necessarily) generalize to other control chart settings.

4.4.1 Signal Behavior

The main motivation for updating control chart limits during monitoring (Phase II) is a lack of sufficient reliable Phase I data before monitoring is required. Thus any updating monitoring scheme should consider signal behavior. Capizzi & Masarotto (2020) advised to re-run Phase I methods on all data collected so far, and re-estimating the parameters with the remaining representative observations. Huberts et al. (2019) gave examples of scenarios where updating and continued use of the chart after a signal is beneficial. If the practitioner can retrospectively identify out-of-control samples and remove them from the data, the chart can safely be updated even after signals. In situations where this is not possible and there is no way to distinguish a false alarm from a correct out-of-control signal, updating is often inadvisable. This does depend on the values of δ, m_I, and the chart that is used (Huberts et al., 2019).

4.5 Case Study Using COVID-19 Data

In this section, we demonstrate the (cautious) updating procedure on a control chart of mortality data. The data consists of the weekly number of deaths in the Netherlands among 0-65 year-old people. Note that we have chosen this age group to limit the size of the shift. Phase I consists of $m_I = 50$ weeks in total; from week 24 of 2017 to week 21 of 2018. Phase II consists of 100 weeks from week 22 in 2018 to week 19 in 2020. The COVID-19 disease was given the official pandemic label by the World Health Organization on the 11th of March 2020 (week 12 of 2020). At that time there were 503 positive tests for COVID-19 in the Netherlands and 5 reported deaths related to the virus. No nationwide restrictions were in place on that date. Figure 4.3 displays the full 150 weeks of data, where the peak near the end corresponds to the increased death rate due to COVID-19 infections.

We set $ARL_0 = 370$ and $\delta = 0.25$ for the control charts in this case study.

Figure 4.4 displays the Shewhart chart as described in Section 1.1.1 without updating parameter estimates. This chart fails to signal the outbreak of the pandemic. In Figure 4.5 we display the Shewhart chart with the proposed method of Capizzi & Masarotto (2020), using $A = 1.5$ and $B = 50$. Using these settings the chart does not update during monitoring and thus also fails to signal the increased mortality among the age groups considered here. In Figure 4.6 we display the Shewhart chart using the updating rules in Equations (4.10) and (4.11). These rules lead to $A = 2$ and $B = 25$ as input for the updating procedure. As a consequence of these settings, the parameter estimates are updated twice during the monitoring phase. This delivers more accurate parameter estimates and results in a signal from the Shewhart chart.

Figure 4.7 shows the EWMA chart using $\lambda = 0.2$. Similar to the Shewhart chart, there is no signal for the COVID-19 pandemic. The chart does signal for decreasing death rates near the end of 2019, although there is no known assignable cause for this. The same results are obtained when using the values $A = 1.5$ and $B = 50$ as proposed by Capizzi & Masarotto (2020), as is demonstrated in Figure 4.8. In that scenario, the parameters are not updated during the monitoring phase. Figure 4.9 shows the chart using updating rules (4.12) and (4.13) resulting in $A = 2$ and $B = 50$. This time the chart does signal the increased mortality rate at the time of the COVID-19 pandemic.

Finally, the CUSUM chart with $k = 1$ signals both a decrease in death rates at the end of 2019 and the increase in rates due to the COVID 19 pandemic, without updating the limits. Updating the limits in Figure 4.12 leads to slightly narrower limits, but doesn't make a difference in the detection in this case.

Concluding this section, in the case of signaling the pandemic in the death rates for ages 0 to 65, updating the parameters improves the charts' performances. Using the parameters set by Capizzi & Masarotto (2020) does not trigger updates, thus neutralizing the updating procedure. Using the proposed rules in Equations (4.10)-(4.13) does trigger updates of the parameter estimates in the monitoring phase, improving performance and leading to an out-of-control signal at the time of the COVID-19 pandemic for all three charts. This case study is not exhaustive, but it does show a clear example of the benefit of using an updating procedure, especially with adjusted rules for the parameters.

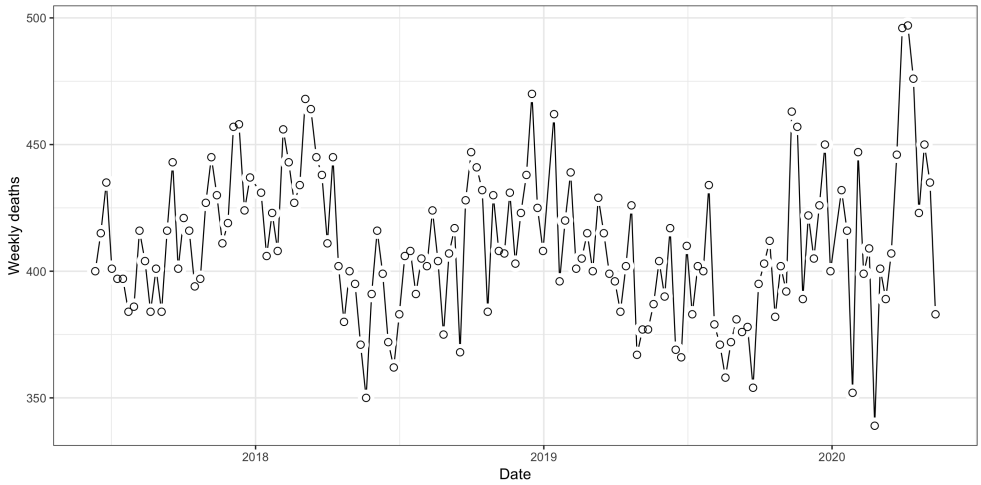


Figure 4.3 – Weekly deaths in the Netherlands for people aged under 65

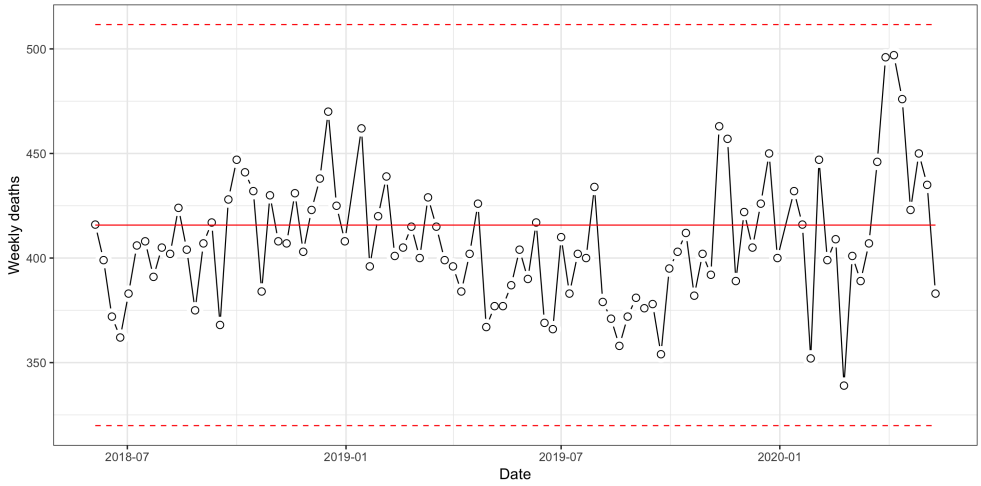


Figure 4.4 – The Shewhart control chart, without updating the limits ($A = 0, B = 0$)

4.6 Concluding Remarks

In this chapter, we investigated the cautious parameter updating approach of Capizzi & Masarotto (2020). Parameter estimation is an issue when determining control limits for the Shewhart, EWMA, and CUSUM control charts, and can have a substantial impact on the control chart performance. One approach to dealing with the estimation

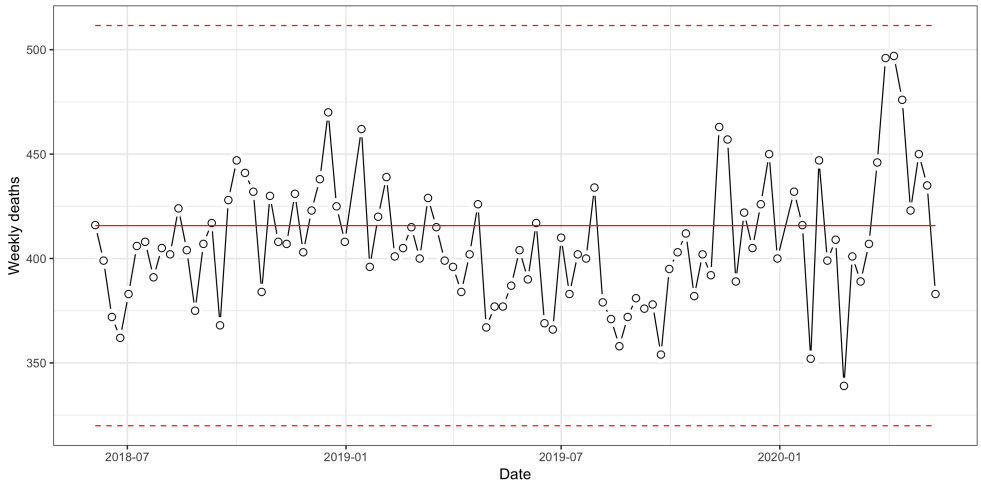


Figure 4.5 – The Shewhart control chart, updating the limits using $A = 1.5, B = 50$

error is to update the parameter estimates during Phase II.

We evaluated the cautious updating approach of Capizzi & Masarotto (2020) and propose adjustments to their procedure. An approximation of the expected time to the first parameter update shows that choosing the appropriate updating parameters is important to prevent incorporating contaminated samples in the parameter estimates. We have shown that the $EARL$s are a result of the mean deviation δ, the number of Phase I samples m_I, and the updating parameters A and B. To ensure optimal Phase II performance, formulas were developed for A and B given the available Phase I data and the value of δ that is important to the practitioner. Using these formulas delivers very promising chart performance.

In a case study using COVID-19 data, we demonstrated the added value of updating the control limits for mortality rates in the Netherlands. The updating procedure works especially well when using Equations (4.10)-(4.13) as rules for updating Equation (4.3).

Updating control chart limits is a logical step towards reducing parameter estimation uncertainty. However, updating using contaminated samples can cause the estimates to spiral out of control. The methods described in this chapter greatly reduce the probability of updating using contaminated samples, while still benefitting from the improved estimation accuracy when possible.

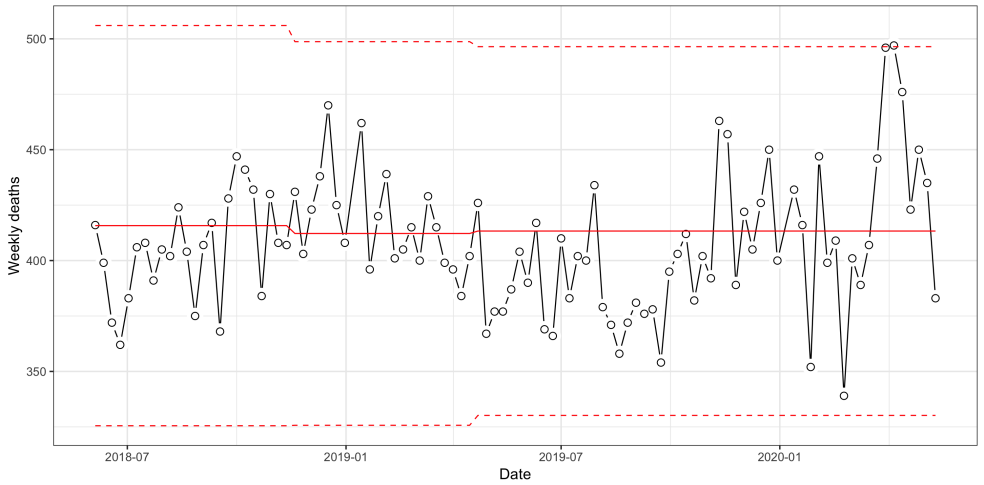


Figure 4.6 – The Shewhart control chart, updating the limits using rules (4.10) and (4.11) resulting in $A = 2$ and $B = 25$

4.7 Appendices

4.7.A Expectation - Unconditional

In this section, we consider the unconditional expectation of Equation (4.3). For the left-hand side of Equation (4.3), it is possible to determine the expectation of an individual term in the sum. First, note that

$$\begin{aligned} X_h &\sim N\left(\mu + \delta\sigma, \sigma^2\right), \\ \bar{X}_{i-d_i} &\sim N\left(\mu, \sigma^2/m_I\right), \\ \frac{(m_I - 1)S^2_{i-d_i}}{\sigma} &\sim \chi^2_{m_I-1}. \end{aligned}$$

Since X_h and $\bar{X}_{i-d_i}$ are independent, we also know that $X_h - \bar{X}_{i-d_i} \sim N\left(\delta\sigma, \sigma^2(1 + 1/m_I)\right)$. Denote $Y = \frac{X_h - \bar{X}_{i-d_i}}{S_{i-d_i}}$. We can then rewrite this as

$$\begin{aligned} Y &= \frac{X_h - \bar{X}_{i-d_i}}{S_{i-d_i}} \\ &= \sqrt{1 + 1/m_I}\ \frac{\left(X_h - \bar{X}_{i-d_i} - \delta\sigma\right) / \left(\sigma\sqrt{1 + 1/m_I}\right) + \delta\sigma / \left(\sigma\sqrt{1 + 1/m_I}\right)}{S_{i-d_i}/\sigma} \qquad (4.14) \\ &= \sqrt{1 + 1/m_I}\ \frac{Z + \delta/\sqrt{1 + 1/m_I}}{\sqrt{V/\nu}}, \end{aligned}$$

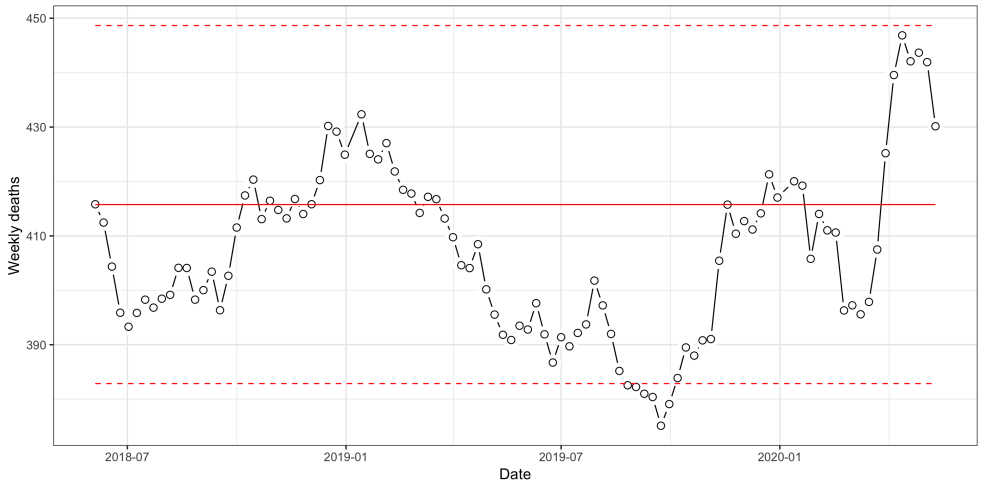


Figure 4.7 – The EWMA control chart, without updating the limits ($A = 0, B = 0$)

where $Z = \frac{X_h - \overline{X}_{i-d_i} - \delta\sigma}{\sigma\sqrt{1+1/m_I}}$ is a standard normal variable, and $V = \frac{(m_I-1)S^2_{i-d_i}}{\sigma^2}$ is a chi-squared variable with $\nu = m_I - 1$ degrees of freedom. Next, note that

$$T = \frac{Z + \delta/\sqrt{1+1/m_I}}{\sqrt{V/\nu}}$$

follows a noncentral t-distribution with $\nu = m_I - 1$ degrees of freedom and noncentrality parameter $\gamma = \delta/\sqrt{1+1/m_I}$. Consequently, $F = T^2$ follows a noncentral F-distribution with $\nu_1 = 1$ numerator degrees of freedom, $\nu_2 = \nu = m_I - 1$ denominator degrees of freedom, and noncentrality parameter $\lambda = \gamma^2 = \delta^2(1 + 1/m_I)$. To get back to (4.3), for $m_I > 3$ the expectation of an individual term in the sum on the left-hand side of the inequality can be calculated to be

$$\begin{aligned} E\left[\left(\frac{X_h - \bar{X}_{i-d_i}}{S_{i-d_i}}\right)^2\right] &= E[Y^2] = \left(1 + \frac{1}{m_I}\right) E[F] \\ &= \left(1 + \frac{1}{m_I}\right) \frac{(m_I - 1)(1 + \delta^2 \frac{m_I}{m_I+1})}{m_I - 3} \\ &= \left(\frac{m_I - 1}{m_I - 3}\right)\left(1 + \delta^2 + \frac{1}{m_I}\right). \end{aligned} \tag{4.15}$$

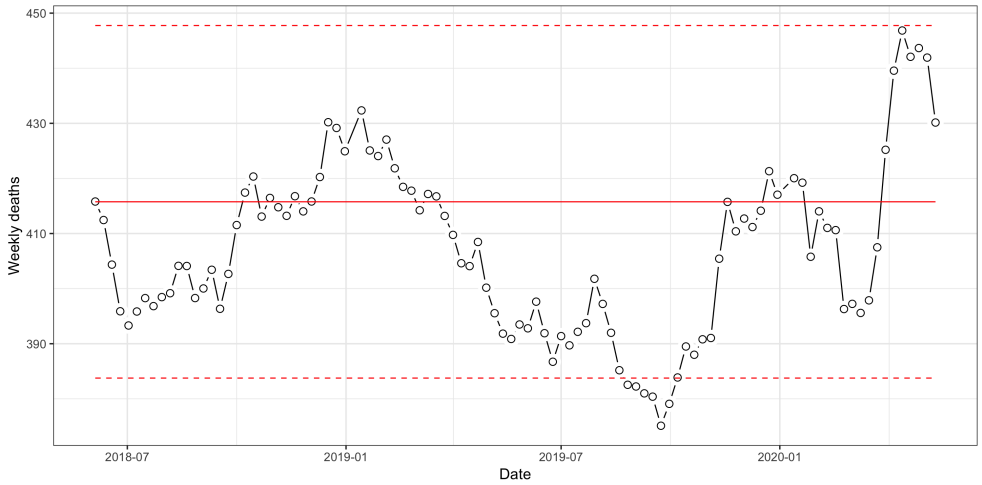


Figure 4.8 – The EWMA control chart, updating the limits using $A = 1.5, B = 5$

4.7.B Expectation of Sum - Conditional

Consider the conditional expectation of the sum on the left-hand side of Equation (4.3). We only consider the time until the first update, such that $d_i = i - m_I$ until the update is done, so that $\bar{X}_{i-d_i} = \bar{X}_{m_I}$ and $S_{i-d_i} = S_{m_I}$, and such that the inequality (4.3) becomes

$$\sum_{h=m_I+1}^{i} \left(\frac{X_h - \bar{X}_{m_I}}{S_{m_I}} \right)^2 < A(i - m_I) - B. \tag{4.16}$$

Consider $Y_h = \frac{X_h - \bar{X}_{m_I}}{S_{m_I}}$. Conditional on $\bar{X}_{m_I}$ and S_{m_I}, we know that

$$Y_h | \bar{X}_{m_I}, S_{m_I} \sim N\left(\frac{\mu - \bar{X}_{m_I}}{S_{m_I}} + \delta \frac{\sigma}{S_{m_I}}, \frac{\sigma^2}{S_{m_I}^2} \right),$$

or equivalently

$$\frac{S_{m_I}}{\sigma} Y_h | \bar{X}_{m_I}, S_{m_I} \sim N\left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta, 1 \right).$$

We then rewrite the left-hand side of inequality (4.5) into

$$\begin{aligned} \sum_{h=m_I+1}^{i} \left(\frac{X_h - \bar{X}_{m_I}}{S_{m_I}} \right)^2 &= \sum_{h=m_I+1}^{i} Y_h^2 \\ &= \frac{\sigma^2}{S_{m_I}^2} D_i, \end{aligned} \tag{4.17}$$

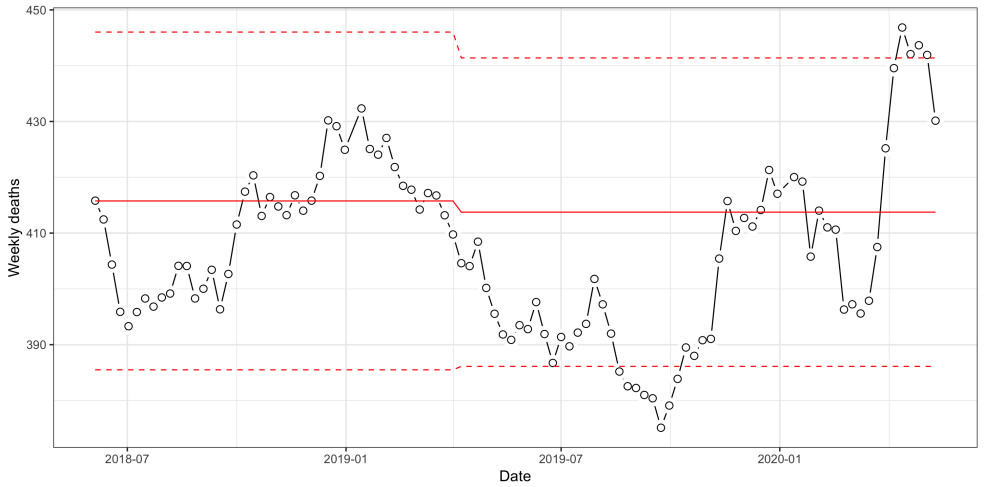


Figure 4.9 – The EWMA control chart, updating the limits using rules (4.12) and (4.13) resulting in $A = 2$ and $B = 50$

where $D_i = \sum_{h=m_I+1}^{i} \left(\frac{S_{m_I}}{\sigma} Y_h \right)^2$. Note that $D_i | \bar{X}_{m_I}, S_{m_I}$ follows a noncentral chi-square distribution with $(i - m_I)$ degrees of freedom and noncentrality parameter $(i - m_I) \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta \right)^2$. From this, we calculate the expectation of the left-hand side of inequality (4.5), conditional on $\bar{X}_{m_I}$ and S_{m_I}, to be

$$\begin{aligned} & E\left[\frac{\sigma^2}{S_{m_I}^2} D_i \,\middle|\, \bar{X}_{m_I}, S_{m_I} \right] \\ & = (i - m_I) \left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta \right)^2 \right) \frac{\sigma^2}{S_{m_I}^2}. \end{aligned} \tag{4.18}$$

Next, we replace the sum in the left-hand side of inequality (4.5) by its expectation, so that we obtain the inequality

$$(i - m_I) \left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta \right)^2 \right) \frac{\sigma^2}{S_{m_I}^2} < A(i - m_I) - B. \tag{4.19}$$

We can use this inequality to provide an estimate of the expected time to the first update (ETFU). Since B should be a positive number in this method, note that this inequality will never hold if $\left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta \right)^2 \right) \frac{\sigma^2}{S_{m_I}^2} \geqslant A$. If $\left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta \right)^2 \right) \frac{\sigma^2}{S_{m_I}^2} <$

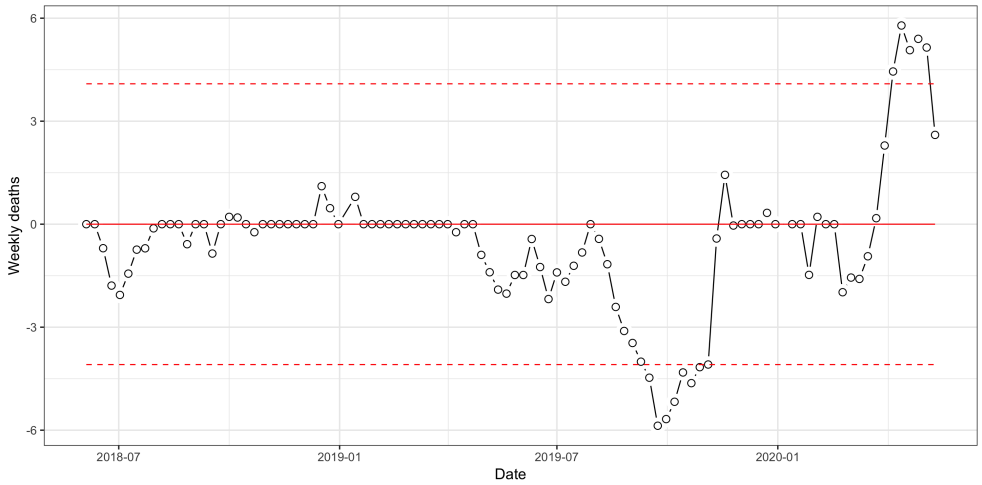


Figure 4.10 – The CUSUM control chart, without updating the limits ($A = 0, B = 0$)

A, then we can solve the inequality for i and find that

$$i \geqslant m_I + \frac{B}{A - \left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta\right)^2\right) \frac{\sigma^2}{S^2_{m_I}}}. \tag{4.20}$$

Thus, our estimate of ETFU, conditional on $\bar{X}_{m_I}$ and S_{m_I}, is equal to

$$ETFU|\bar{X}_{m_I}, S_{m_I} = \left\lceil \frac{B}{A - \left(1 + \left(\frac{\mu - \bar{X}_{m_I}}{\sigma} + \delta\right)^2\right) \frac{\sigma^2}{S^2_{m_I}}} \right\rceil, \tag{4.21}$$

where $\lceil X \rceil$ represents the ceiling function.

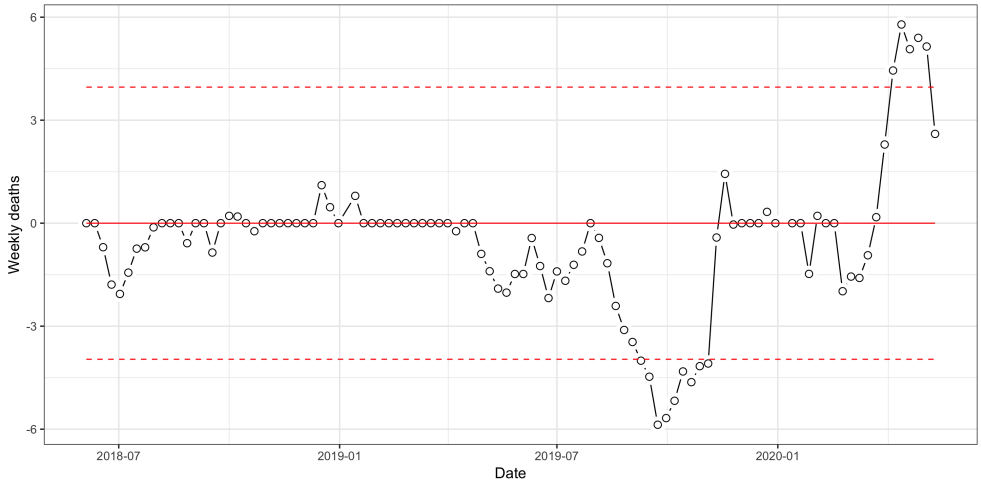


Figure 4.11 – The CUSUM control chart, updating the limits using $A = 1.5, B = 50$

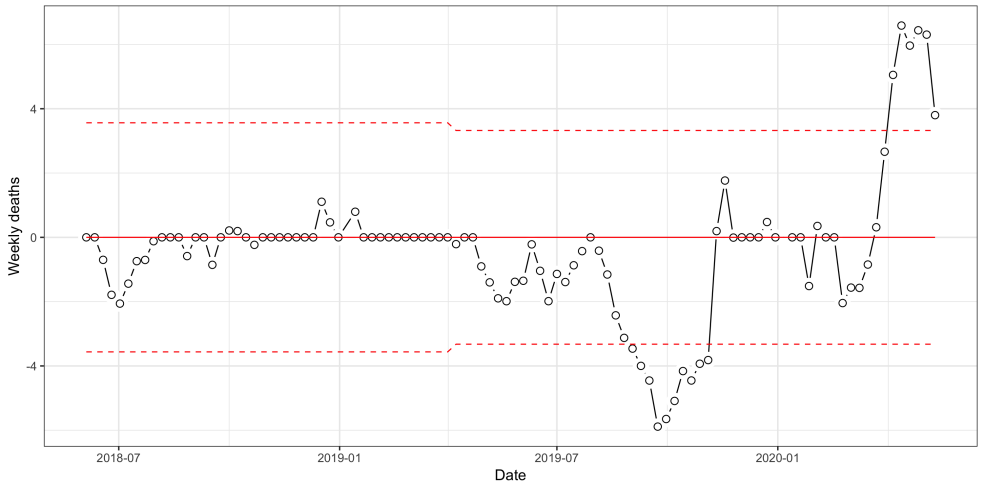


Figure 4.12 – The CUSUM control chart, updating the limits using rules (4.12) and (4.13) resulting in $A = 2$ and $B = 50$

Part II

Predictive Process Monitoring

Chapter 5

Boosted Predictive Process Monitoring in Mental Health

5.1 Motivation

In this chapter, we consider a wide range of techniques for predictive process monitoring (PPM). As described in Section 1.2, PPM aims to produce early warnings of unwanted events. A mental health case study is presented, in which we demonstrate the use of novel gradient boosting methods in predictive monitoring.

Gradient boosting is an important recent development within machine learning for regression and classification. This technique produces an ensemble of decision trees that minimize an appropriate loss function. Such an ensemble often produces better predictions than a single, more comprehensive model. When used for classification, as in most process monitoring applications, the predictions are in the form of probabilities, similar to the output of regression models. In this paper, these predicted probabilities will be used in a process monitoring procedure.

An area that has a real interest in predictive monitoring for quality control is mental health (Hahn et al., 2017). Nineteen percent of adults in the United States have a mental, behavioral, or emotional disorder (Substance Abuse and Mental Health Services Administration, 2018). These disorders pose a heavy burden on the patient and affected families, the healthcare system, and healthcare expenditure. Early intervention is important for many of the severe mental disorders and could prevent the escalation of the disease. However, it is very hard to predict the progress of a disease

and thus determine when to intervene.

One of the most debilitating mental disorders is schizophrenia. According to the World Health Organization (2019), schizophrenia is characterized by distortions in thinking, perception, emotions, language, sense of self, and behavior. Common experiences include hallucinations and delusions. Around 0.75 percent of people suffer from the disease worldwide (Moreno-Küstner et al., 2018). Schizophrenia is one of the top 15 leading causes of disability worldwide (Vos et al., 2017). The all-cause standardized mortality rate is around 3.7 times higher for people diagnosed with schizophrenia compared to the general adult population (Olfson et al., 2015). Suicide is much more frequent among people suffering from schizophrenia. An estimated 4.9% of people with schizophrenia die by suicide compared to 0.013% for the general population (Palmer et al., 2005).

The overwhelming majority of people suffering from schizophrenia will relapse into crisis care over time, even with access to good care (Emsley et al., 2013). Relapse averages are reported between 20% and 40% per year, depending on many factors (Ruetsch et al., 2018). In the United States, the estimated total cost of schizophrenia was $155.7 billion in 2013 (Cloutier et al., 2016). The healthcare costs for people diagnosed with schizophrenia are significantly higher than the national average, where relapse events (i.e. hospital admissions) are the most expensive (Karve et al., 2012). This motivates the need for early identification of patients at high risk of having a mental health crisis, to facilitate preventive measures, and mitigate the high costs associated with these crises.

In a systematic review, Sullivan et al. (2017) investigated models to predict crises and found a lack of high-quality evidence on prediction methods. Paxton et al. (2013) and Amarasingham et al. (2014) highlighted the challenges that predictive modeling based on Electronic Medical Records faces. According to Sullivan et al. (2017), the number of studies with promising results is very limited. One of the exceptions is the study by Vigod et al. (2015), which used Canadian data from 2008-2011 to predict 30-day readmission rates for acute psychiatric units using logistic regression. The study shows moderate discriminative capacity with an area under the receiver operating characteristic curve of 0.630.

In this study, we have access to all mental health treatment records covered by the

Dutch Health Insurance Act, which covers all specialist mental health treatment of schizophrenia for all 17 million residents of the Netherlands between 2010 and 2014. We use this data to predict readmission into crisis care for the 75,000 people diagnosed with schizophrenia in the Netherlands. We compare the predictive power of logistic regression as used by Vigod et al. (2015) to a hierarchical regression model.

Subsequently, a gradient boosting algorithm named Extreme Gradient Boosting (abbreviated by XGBoost) is applied to the data. Teinemaa et al. (2019) reviewed available PPM techniques and concluded that XGBoost is reasonably fast and often the most accurate technique. It can deal with class imbalance and incomplete observations, which are often found in medical data (Paxton et al., 2013). Furthermore, Zhang et al. (2018) used XGBoost for predictive monitoring of faults in wind turbines.

The XGBoost predictions are then used to monitor the set of people diagnosed with schizophrenia during a monitoring phase. We present an algorithm to determine the threshold to signal and consider the monitoring performance. The goal of the procedure is to support healthcare workers in identifying individuals at risk of a crisis.

The article is structured as follows. In the next section, we present the mental health problem context. We describe the predictive models we consider in this study and their performance in Section 5.3. In Section 5.4 we propose an algorithm to determine the monitoring threshold and present the monitoring results when using the XGBoost technique. In the last section, we provide concluding remarks and limitations. This chapter has been based on Huberts et al. (2020a).

5.2 Problem Description

This section describes the setting in which this case study attempts to monitor the risk of mental health crises. First, we give a summary of the mental healthcare system in the Netherlands. We then describe the available data, followed by the definition of a crisis event.

5.2.1 The Mental Healthcare System

The total cost of mental healthcare in the Netherlands was estimated to be 6.5 billion euros in 2017 (Statistics Netherlands, 2017), 416 million euros of which was directly related to schizophrenia (National Institute for Public Health and the Environment, 2017). The details of the healthcare system are out of scope, but this section describes the basics.

Similar to the US system, mental healthcare in the Netherlands is organized through managed competition. In contrast to the US, health insurance is mandatory for all Dutch citizens and covers 99.9% of the population (OECD & European Observatory on Health Systems and Policies, 2019). Insurers are required to accept all applicants and offer community rating. The deductibles are relatively low in the Netherlands and there is risk adjustment among insurers. Coverage includes a broad set of essential health benefits, including out- and in-patient treatment of nearly all disorders in the Diagnostic and Statistical Manual of Mental Disorders 5 (American Psychiatric Association, 2013), except select diagnoses such as adjustment disorder (since 2012). Curative healthcare expenditure is relatively high with per capita spending of $3,791$ euros in 2017 and the long-term care spending is the highest of all EU countries. The system is comparatively effective and the Netherlands reports the lowest rate of unmet medical needs among EU countries (OECD & European Observatory on Health Systems and Policies, 2019).

5.2.2 Data Description

The non-public Microdata used in this paper is provided by Statistics Netherlands. It consists of a wide range of de-identified administrative data sets on all 17 million Dutch citizens. As the data is very sensitive, it is stored on secure servers at Statistics Netherlands and can solely be accessed on their local terminals.

A patient requires a diagnosis for the health insurer to reimburse the mental healthcare treatments. The diagnosis, together with the amount of treatment provided, constitutes the so-called Diagnosis-Treatment-Combination (DBC in Dutch) that determines the amount of reimbursement. As reimbursement is dependent on consistent registration, the system results in a clear timeline of the diagnoses and

activities for a patient.

Mental health data were available for the years 2010 through 2013. The data includes all registered healthcare information, such as detailed mental health diagnoses (1.4 million) and psychological treatments (25 million). Furthermore, individual data on employment, housing, and personal information were available. The data included in the following concerns the subset of 75,000 people diagnosed with schizophrenia. The resulting selection consists of an unbalanced panel data set, with a large variation in the number of registered treatment activities per individual.

The gathered data cannot be used directly for statistical modeling, as the sequences of diagnoses and treatments vary widely in length, frequency, and type. Time series models thus require padding and aggregation to balance the data, which will produce sparse sequences for many of the included individuals. The decision of the level of aggregation was motivated by domain experts and set at the week-level. The week-level aggregation had enough detail to expect predictive value in the data while resulting in an actionable time-frame for intervention.

Diagnoses and treatments were described in detailed labels. These were condensed into broader categories to avoid high-dimensionality and sparsity. The 36 resulting categories of diagnoses are tracked cumulatively. The weekly aggregations of treatments are further aggregated into a short-, medium- and long-term history of 4, 12, and 64 weeks respectively. The 4 weeks represent the past month of data, the 12 weeks represent the last quarter and the past 64 weeks consist of the last year plus one quarter to make sure there is overlap between separate (administrative) years. These three levels result in 264 predictors (74 for each of the three treatment aggregation levels, 36 cumulative diagnoses, and 6 fixed variables) for more than 15 million person-weeks.

5.2.3 The Definition of Crisis

A variable indicating a crisis was not readily available and had to be constructed from the raw data. This section describes how the variable to signal a crisis event was constructed. As described previously, individuals are assigned DBCs. Some of these DBCs are directly defined as 'crisis care', i.e. treating a patient in a mental health institution due to a crisis. In other cases, crisis care is registered to an existing,

non-crisis DBC. Therefore, we define the start of a crisis event as the first moment a crisis DBC was opened or any crisis care was given.

Furthermore, as the goal of the procedure is to support healthcare workers in identifying individuals at risk of a crisis, the model should not focus on individuals that are already known to be in a mental health crisis. The crisis care service in the Netherlands aims to provide a maximum of 12 weeks of crisis care. In the data, following the start of a crisis an individual will be excluded for the following 20 weeks. This 20-week period covers the 12 weeks that a crisis can cover, plus some additional weeks after that, where it could be argued that a patient will be on the radar of the healthcare professional and a signaling mechanism is not needed to achieve this.

This results in a binary dependent variable on a weekly basis, where a TRUE value equals the start of the crisis in that week and a FALSE indicating the subject is not in a crisis. This binary dependent variable is sparse with only 0.285% TRUE values for all person-weeks.

As the long-term aggregation level (64-week aggregations) includes all diagnoses and healthcare activities from the past 64 weeks, the first week we can model is week 65 (i.e. March 2011). Figure 5.1 plots the aggregated number of crises per week. There are far fewer crises that start in 2013 than the other two years due to administrative changes and incomplete data for that year. This is important to consider when evaluating the results.

In total there are just over 28,300 crises that start between March 2011 and December 2013. These crises are divided over 22,600 people with an average of around 0.4 crisis per person in the data. The mean number of weeks between crises (for people with multiple crises in the data) equals 51 weeks. Figure 5.2 gives an overview of the number of crises per individual that had a crisis at least once. A large majority has one crisis during the three years we consider. Table 5.1 gives an overview of the ten mental healthcare activities with the largest weekly mean values in the data.

5.3 Predictive Model

This section describes the logistic regression model (similar to Vigod et al., 2015), the hierarchical regression model, and the XGBoost algorithm that are used to predict

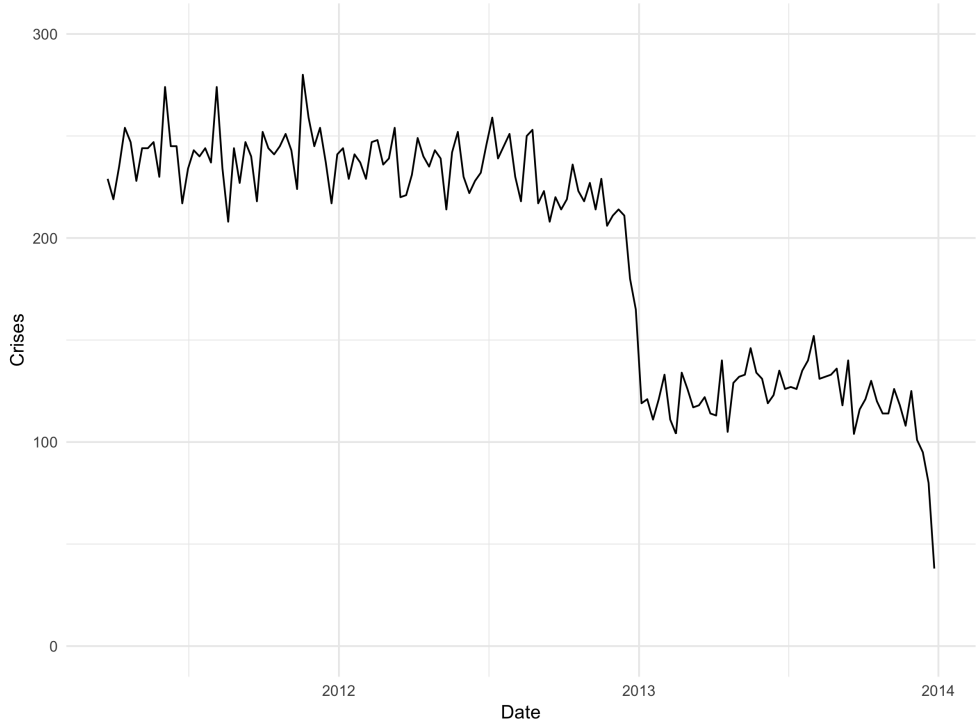


Figure 5.1 – Number of crises per week from March 2011 (week 65) to December 2013 (week 209)

weekly probabilities of getting a crisis for people diagnosed with schizophrenia.

Define $y_{i,t}$ as a binary variable for individual $i = 1, ..., N$ and week $t = 1, ..., M$. If person i has a crisis in week t, then $y_{i,t} = 1$, if there is no sign of crisis $y_{i,t} = 0$. Let $P_{i,t}$ be the predicted probability of crisis in week t for person i. Furthermore, $X_{i,t}$ contains the constructed features based on the 4-, 12- and 64-week history of individual i, as well as the individual characteristics.

5.3.1 Regression

Using regression models for prediction has the advantage of high explainability. In contrast to many of the more advanced machine learning techniques, the parameters of the logistic and hierarchical regression models we discuss can offer insight into the process dynamics.

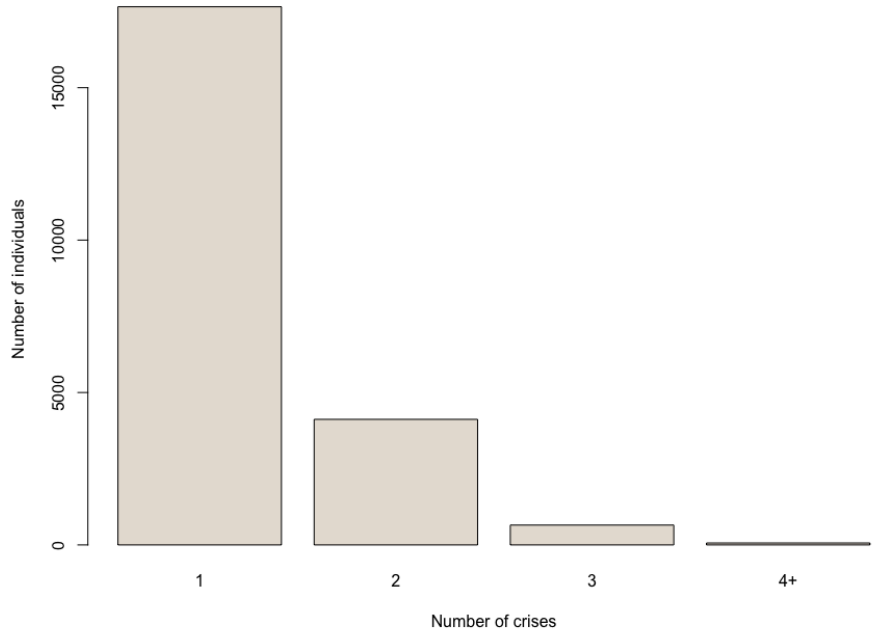


Figure 5.2 – Number of crises per individual from March 2011 (week 65) to December 2013 (week 209)

5.3.1.1 Logistic Regression

Logistic regression is used to model the probabilities of a categorical outcome variable. In this case, the categorical outcome is binary. The model for the probability of crisis $P_{i,t} = P(y_{i,t} = 1|X_{i,t})$ for individual i in week t with vector of predictors $X_{i,t}$ has the form (Hastie et al., 2009)

$$P_{i,t} = \frac{exp(\beta_0 + \beta' X_{i,t})}{1 + exp(\beta_0 + \beta^T X_{i,t})}, \tag{5.1}$$

where β_0 is a vector of constants and β' is the transposed vector of parameters for the predictors $X_{i,t}$. The model is fitted using maximum likelihood. The estimated parameters $\hat{\beta}_0, \hat{\beta}$ can be used for inference and prediction of $P_{i,t+1}$.

5.3.1.2 Hierarchical Regression

Hierarchical modeling is almost always an improvement over single level regression models (Gelman, 2006). In this case, each observation equates to one person-week.

Table 5.1 – Descriptive statistics for the ten most frequent mental healthcare activities from March 2011 (week 65) to December 2013 (week 209) per person-week

Activity	Mean	Std.Dev
Individual contact	4.81	16.92
General no-show/other	1.13	8.04
Individual activating guidance	0.66	7.23
Diagnostics	0.65	8.23
Pharmacotherapy	0.60	4.86
Individual other communication	0.31	3.82
Individual supportive guidance	0.29	4.72
Group contact	0.21	2.91
Patient system	0.17	3.34
Crisis treatment	0.16	4.23

The mental healthcare activities, diagnoses, income changes, and crises leading up to a specific week all relate to a single person. We can model this as a simple two-level hierarchy, with the weeks as the lower level and the individual as the upper level (see Figure 5.3). Suppose we have p_0 predictors on the person-week level and p_1 predictors on the person level. The hierarchical model for the log-odds ratio of the probability of a crisis for individual i in week t is then defined as

$$\log\left(\frac{P_{i,t}}{1-P_{i,t}}\right) \sim N(X_{i,t}\alpha_i, \sigma^2), \text{ for } t = 1, ..., M \text{ (Week level)},$$

where the individual level is modelled as

$$\alpha_i \sim N(\gamma W_i', \Sigma), \text{ for } i = 1, ..., N \text{ (Individual level)},$$

where $X_{i,t}$ is a $1 \times (p_0 + 1)$ row vector of person-week specific variables such as mental healthcare activities and diagnoses (see Table 5.1); α_i is a $(p_0 + 1) \times 1$ vector of parameters for individual i; σ^2 is the variance for the person-week level; γ is a $(p_0 + 1) \times (p_1 + 1)$ parameter matrix determined by the person i that person-week t is a part of; W_i is a $1 \times (p_1 + 1)$ row vector of person specific variables such as age and Σ is the covariance matrix for parameters α_i. We estimate the parameters using restricted maximum likelihood (REML) in the lme4 package in R (Bates et al., 2015).

5.3.2 Machine Learning

A lot of progress has been made in the machine learning domain in recent years. Increased availability of data and computing power extend the range of models that

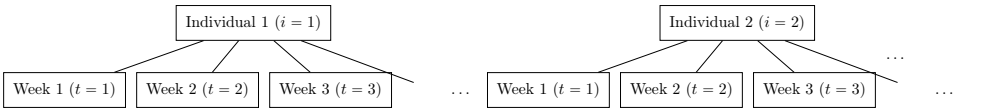


Figure 5.3 – Two-level structure of the case study data with individuals ($i = 1, ..., N$) as the top level. Weeks in the data ($t = 1, ..., M$) belonging to individual i are the bottom level.

can be estimated. In this section, we discuss a gradient boosting framework called extreme gradient boosting and a few alternative techniques.

5.3.2.1 Extreme Gradient Boosting

Gradient boosting is one of the most important recent developments in machine learning (Hastie et al., 2009). XGBoost is an open-source framework to apply gradient boosting in various programming languages (Chen & Guestrin, 2016). The gradient boosting decision tree algorithm within XGBoost creates an ensemble of weak learners that minimize an appropriate loss function. Each weak learner consists of a regression tree grown on the residuals. Each tree has a number of terminal nodes $j = 1, .., J_k$ that each represents a terminal region $R_{k,j}$, containing the predictions for that specific tree. The output of a regression tree k is multiplied by learning rate η and then added to the predictions of tree $k - 1$.

Figure 5.4 gives a fictional example of two trees grown to predict the risk of getting heart disease. The values in the terminal nodes are the log-odds $\log(\frac{P}{1-P})$. The final prediction equals the initial log-odds prediction plus the predictions of the regression trees multiplied by the learning rate η. Given the example in Figure 5.4, the final prediction of the log-odds for a 35-year-old person with a systolic blood pressure of 100 and 90 minutes of physical activity equals $0 - 0.3 \times 2 - 0.3 \times 3 = -1.5$ ($P \approx 0.18$). The log-odds prediction for 25-year-old person with systolic blood pressure 145 and 30 minutes of physical activity equals $0 - 0.3 \times 1 + 0.3 \times 1 = 0$ ($P = 0.5$).

In this study, the outcome is binary thus the logistic loss function is used, given by

$$L(y_{i,t}, P_{i,t}) = -y_{i,t} \log(P_{i,t}) + (y_{i,t} - 1) \log(1 - P_{i,t}), \tag{5.2}$$

for individual i at time t. The algorithm is initialized by setting $P_{i,t,0} = 0.5$. Then a

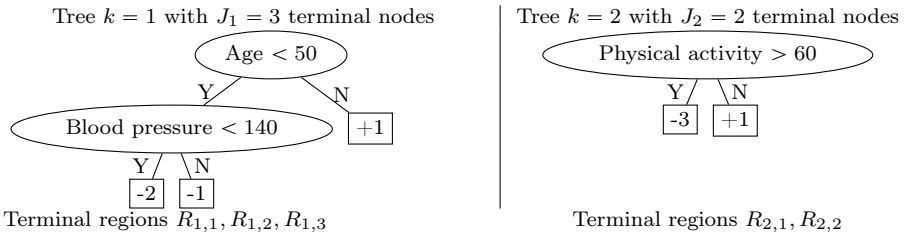


Figure 5.4 – An example of a tree ensemble model for the risk of heart disease using two trees ($K = 2$) and three variables. The learning rate is $\eta = 0.3$. The initial prediction equals $P_0 = 0.5$.

manually specified number of trees $k = 1, 2, ..., K$ is grown on the pseudo-residuals. The pseudo-residuals are $r_{i,t,k} = y_{i,t} - P_{i,t,k-1}$, where $P_{i,t,k-1}$ are the predicted probabilities of the previous iteration $k - 1$.

For each tree k the objective function $obj(t, k)$ for log-odds output values $f_k(X_{i,t})$ consists of the loss function, the pruning term and a regularization term

$$obj(t,k) = \left[\sum_{i=1}^{n} L(y_{i,t}, f_k(X_{i,t}) + f_{k-1}(X_{i,t}))\right] + \chi J_k + \frac{1}{2}\lambda f_k^2(X_{i,t}), \tag{5.3}$$

where χ is a user specified pruning parameter and λ is a regularization parameter.

The output value at time t for the individuals i in terminal node $j = 1, .., J_k$ in tree k that minimizes $obj(t, k)$ is approximated using the second order Taylor expansion. The gradient for $L(y_{i,t}, P_{i,t})$ equals $g_{i,t} = y_{i,t} - P_{i,t}$ and the hessian equals $h_{i,t} = P_{i,t}(1 - P_{i,t})$. The output value for node j in tree k is then given by

$$f(X_{i\in j,t,k}) = \frac{\sum_{i\in j}(y_{i,t} - P_{i,t,k-1})}{\sum_{i\in j}\left[P_{i,t,k-1}(1 - P_{i,t,k-1})\right] + \lambda}. \tag{5.4}$$

The similarity score used to determine the splits when constructing the decision trees is found by plugging the output value back into the second order Taylor approximation and is given by $\frac{(\sum_i (y_{i,t} - P_{i,t}))^2}{\sum_i P_{i,t}(1-P_{i,t}) + \lambda}$.

There are four parameters $\Theta = \{\eta, \lambda, \chi, K\}$ that need to be tuned using cross-validation. Sparsity-aware split finding in the XGBoost framework ensures the algorithm works efficiently for the sparse data in this study. We use the implementation of XGBoost in R (Chen et al., 2019).

5.3.2.2 Other Machine Learning Methods

Several other machine learning techniques can be used in this setting. Examples include (one-class) support vector machines (SVM), decision trees, random forest, and elastic nets. For an overview of machine learning methods in process monitoring see Weese et al. (2016). We applied one-class support vector machines, random forest, and the elastic net methods to the same data in this paper. Compared to the XGBoost predictions, these methods produced inferior results. This is the reason that we have excluded these methods from the analysis. There are some methods, such as recurrent neural networks (see for example Choi et al., 2016), that require more computing power than was available at the Statistics Netherlands terminals.

5.3.3 Estimation

To train the models, weeks 1–175 are used. The first 64 weeks are incorporated into the features, thus the outcomes of weeks 65–175 are used in training. Weeks 176–209 are used to test the predictions. Due to limits in available computing power, the regression models were trained on a random 50% of the people in the training set. To cross-validate the parameters of the XGBoost model we use random splits (75%/25%) of the person-weeks in the training data.

5.3.4 Results

This section describes the results of predicting weekly crises using the logistic regression model, hierarchical model and the XGBoost algorithm described previously. These results are based on the predictions for the test set of weeks 176-209. Because of the highly imbalanced nature of the outcome, we consider the mean assigned probabilities for weeks with crises, $\bar{P}_{t|y=1} = \frac{1}{N}\sum_{i=1}^{N} P_{i,t} I(y_{i,t} = 1)$, and without crises $\bar{P}_{t|y=0} = \frac{1}{N}\sum_{i=1}^{N} P_{i,t} I(y_{i,t} = 0)$. The ratio of these probabilities, $r_t = \frac{\bar{P}_{t|y=1}}{\bar{P}_{t|y=0}}$, is a measure of the predictive power of a model. Values of $r_t \leqslant 1$ indicate that, on average, the procedure assigns the same or a lower probability to person-weeks with positive outcome values. Values $r_t > 1$ show that, on average, the method estimates a higher probability to person-weeks with positive outcome values. We also report the widely used area under the receiver operating characteristic curve (AUC) values

(see Bradley, 1997). AUC values close to 0.5 indicate a total lack of predictive power, values close to 1 represent perfect prediction.

Table 5.2 – Table of the mean estimated probabilities for the three methods, aggregated by the binary crisis outcome variable

Model	$\bar{P}_{t\|y=0}$	$\bar{P}_{t\|y=1}$	r_t	AUC
Logistic regression*	0.0021	0.0026	1.2260	0.6130
Hierarchical regression*	0.0021	0.0030	1.4370	0.5856
XGBoost	0.0017	0.0062	3.5440	0.6533

*Using 50% of the persons in the data due to limited memory size

Table 5.2 shows the measures of performance for the three methods using the test set. The logistic regression model does predict a slightly higher probability for person-weeks with crises with a ratio of 1.226. The limited predictive power is also shown by the AUC value of 0.613. The hierarchical model incorporates some more of the structure in the data. The performance in terms of the ratio of predicted probabilities seems slightly better than the logistic model with $r_t = 1.437$. Conversely, the AUC value is lower than for the logistic regression, which indicates a limited predictive power. Lastly, Table 5.2 shows the results for the XGBoost algorithm predictions on the test set. The mean ratio r_t is around 3.5 and it has the highest AUC score of the three, which shows there is more predictive power than for the logistic and hierarchical regressions. On average, for person-weeks with a crisis, the XGBoost algorithm predicts a probability 3.5 times higher than for person-weeks with no crisis. This suggests the predicted probabilities of the algorithm might be used as a risk score to guide mental health workers towards unstable individuals.

5.4 Monitoring

The three models in the previous section predicted if a crisis is likely to occur for an individual in a given week. This section will discuss the use of these predictions for monitoring.

Using logistic regression, a hierarchical model, or the XGBoost algorithm to model the risk of a crisis will result in weekly estimated probabilities. Monitoring these probabilities requires a probability control limit C. Once a probability $P_{i,t}$ passes this limit, the procedure will signal and practitioners can intervene. The choice of C

will determine the number of false/true signals. A higher value of C will decrease the share of false signals, but also decrease the absolute number of true signals.

As discussed in Section 1.1.1, in process monitoring the performance of a monitoring procedure is often quantified using the False Alarm Rate (FAR) or the Average Run Length (ARL) (see, for example, Jones & Woodall, 1998; Shu et al., 2004; Woodall, 2006). The ARL equals the average time it takes for the procedure to signal. A monitoring procedure is configured to satisfy a required ARL_0 or FAR_0 as determined by the practitioner. This generally involves adjusting parameters based on distributional assumptions or simulation. Distributional assumptions are not realistic with a machine learning method such as XGBoost. Thus, in the following section, we propose a simple tuning procedure that achieves a desired FAR_0 for a predictive monitoring approach. This involves cross-validating the predictions and results in a non-parametric monitoring threshold. Note that the procedure is data-driven, thus no distributional assumptions are needed.

5.4.1 Tuning Procedure

In this section, we propose a tuning procedure to achieve a FAR_0. Assume a practitioner determines a FAR_0 based on the monitoring context. In the case of predictive monitoring, a signal is produced when the predicted probability $P_{i,t}$ exceeds the threshold C. The following steps determine the value of C that produces the desired FAR_0:

1. Set FAR_0, the size of the training set N_0 and the test set N_1, the number of cross-validation splits S, the cross-validation proportion q and initialize an empty vector $\hat{C}$.

2. Split data $\{X, y\}$ of size $N \times k$ into $i = 1, 2, ..., N_0$ training set $\{X_0, y_0\}$ of size $N_0 \times k$ and $i = N_0 + 1, N_0 + 2, ..., N$ test set $\{X_1, y_1\}$ of size $N_1 \times k$ with $N_1 = N - N_0$.

3. Draw a random sample R of integers $1, 2, ..., N_0$ of size qN_0 and define the vector V of size $(1 - q)N_0$ as integers $1, 2, ..., N_0 \notin R$.

4. Split the training set $\{X_0, y_0\}$ into $\{X_R, y_R\}$ using rows R and $\{X_V, y_V\}$ using rows V.

5. Use $\{X_R, y_R\}$ to estimate model $F(X)$ and calculate $P_V = F(X_V)$.

6. Use the grid-search algorithm below to find value $\hat{c}$ for which $\frac{1}{(1-q)N_0} \sum_{i \in V} I(I(P_{i,V} \geqslant \hat{c}) \neq y_{i,V}) \approx FAR_0$, with $I()$ the indicator function.

   (a) Set resolution $r > 2, w = 1/r$, search limit w_{lim} to a small value (i.e. 10^{-5}) and initiate grid $G = \{0, 1/r, 2/r, ..., 1 - 2/r, 1 - 1/r, 1\}$ of length $r + 1$.

   (b) Set $c = \min(g \in G : \frac{1}{(1-q)N_0} \sum_{i \in V} I(I(P_{i,V} \geqslant g) \neq y_{i,V}) < FAR_0)$.

   (c) Calculate $FAR_s = \frac{1}{(1-q)N_0} \sum_{i \in V} I(I(P_{i,V} \geqslant c) \neq y_{i,V})$.

   (d) Update w as $w = 2w/r$ and redefine grid $G = \{c - rw/2, c - rw/2 + w, c - rw/2 + 2w, ..., c - rw/2 + (r-1)w, c - rw/2 + rw\}$ of length $r + 1$.

   (e) If $|FAR_0 - FAR_s| > 0.01 FAR_0$ and $w > w_{lim}$ go back to step (b). If $|FAR_0 - FAR_s| \leqslant 0.01 FAR_0$ set $\hat{c} = c$. If $|FAR_0 - FAR_s| > 0.01 FAR_0$ and $w \leqslant w_{lim}$ set $\hat{c} = NA$.

7. Save value $\hat{c}$ in vector $\hat{C}$, if the length of $\hat{C}$ is smaller than S return to step 3.

8. Set threshold value C_{tuned} as $\max(\hat{C})$ for $\hat{C} \neq NA$. If all values in $\hat{C}$ are NA no threshold was found. Use $\{X_0, y_0\}$ to estimate model $F(X)$ and calculate $P_1 = F(X_1)$. The expected FAR ($EFAR$) then equals $EFAR = \frac{1}{N_1} \sum_{i=N_0+1}^{N} I(I(P_{1,i} \geqslant C_{tuned}) \neq y_{1,i})$.

Note that the maximum estimated value in $\hat{C}$ is set as C_{tuned}. Assuming the model generalizes well to new data, this will result in an expected false alarm rate $EFAR$ that is smaller than FAR_0.

The FAR_0 set by the practitioner in this procedure translates to all observations during monitoring. For example, setting $FAR_0 = 0.01$ will result in 1% of observations being false alarms. The other 99% of observations consist of true/false negatives and true positives. A higher FAR_0 will result in a lower value of C. We will demonstrate the procedure in the following section.

5.4.2 Results

In this section, we monitor the probability of a crisis as predicted by the XGBoost algorithm. Threshold C determines when the procedure signals. As measures of performance we consider the precision and recall values as defined in Section 1.2.

Figure 5.5 shows the average estimated probability per week, grouped by the observed value of $y_{i,t}$. On average, the procedure estimates a visibly higher probability for the individuals that have a crisis for all weeks. We cannot show individual probabilities to ensure the privacy of the persons.

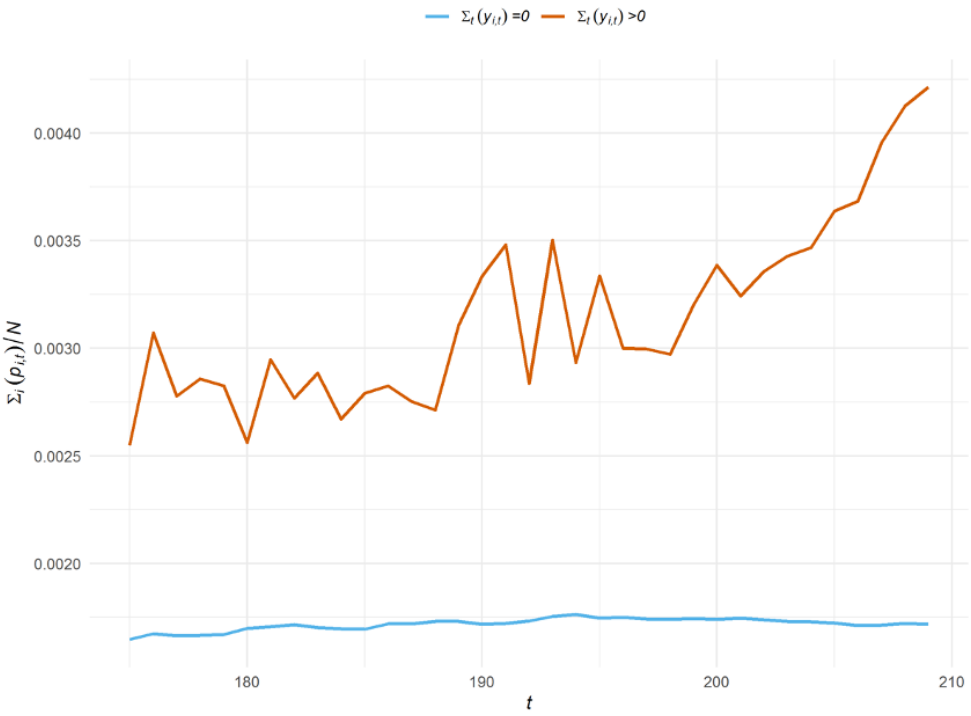


Figure 5.5 – Average probability per week grouped for people that have a crisis in the monitoring time frame (orange) and that do not have a crisis (blue)

Table 5.3 shows the precision and recall values for a wide range of values for C. The table shows perfect recall for $C = 0.0001$, but the precision is very low. Perfect precision is achieved for $C = 0.75$, but the recall is very low. This shows that, although the model does have some predictive power, a high false alarm rate is needed to detect a large portion of the crises.

Table 5.3 – Precision and recall values using the XGBoost estimated probabilities and various values for threshold C

C	Precision	Recall
0.0001	0.0017	1.0000
0.0010	0.0023	0.7955
0.0100	0.0081	0.0517
0.1000	0.1224	0.0087
0.5000	0.5000	0.0018
0.7500	1.0000	0.0003

5.4.2.1 Weeks Before a Crisis

The individual probability plots often show high volatility in the weeks leading up to a crisis. In this subsection, we thus consider the average difference in estimated probability in the weeks before a crisis occurs. Table 5.4 gives the average difference in estimated probabilities over 10 weeks ($\frac{1}{9}\sum_{l=t-9}^{t}(P_{i,l}-P_{i,l-1})$ for week t and individual i), grouped by whether a crisis was observed at the end of those 10 weeks. This shows that, on average, the estimated average difference in probability of a crisis in the 10 weeks leading up to crisis is 135 times higher than if no crisis is observed after these 10 weeks.

Table 5.5 gives the precision and recall values for a range of C values when predicting if a crisis will occur within 10 weeks. This shows a higher precision for low values of $C \leqslant 0.1$ compared to the weekly predictions (cf. Table 5.3).

Table 5.4 – Average estimated difference in probabilities per crisis/no crisis if we consider a 10-week period

Crisis within 10 weeks	Average estimated difference in probability
No	-0.000002
Yes	0.000296
Ratio	-135.846300

5.4.2.2 Tuning the Procedure

In this section, we run the procedure including the tuning algorithm for various required FAR_0s. The desired FAR_0 produces a value for C through the tuning algo-

Table 5.5 – Precision and recall values using the XGBoost estimated probabilities of a crisis within 10 weeks and various values for threshold C

	C	Precision	Recall
1	0.0001	0.0148	1.0000000
2	0.0010	0.0183	0.8244052
3	0.0100	0.0438	0.0265457
4	0.1000	0.1603	0.0010674
5	0.5000	0.5000	0.0001685
6	0.7500	1.0000	0.0000281

rithm outlined in Section 5.4.1. This value C_{tuned} is then used to monitor the test set. We use $S = 10$ splits in the procedure and use cross-validation proportion $q = 0.75$ for all values of FAR_0.

Table 5.6 gives the tuned values C_{tuned} for a set of predetermined false alarm rates $FAR_0 \in \{0.5, 0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001\}$, as well as the precision/recall values in the test set and the actual observed $FAR_{observed}$. The table shows that all the $FAR_{observed}$ values are smaller than their respective FAR_0 values. In step 8 of the procedure in Section 5.4.1 the maximum estimated value in $\hat{C}$ is set as C_{tuned}. This results in $FAR_{observed} \leqslant FAR_0$ for all FAR_0 values of Table 5.6.

Table 5.6 – C_{tuned}-values for various predetermined values of FAR_0, as well as the precision/recall and the observed $FAR_{observed}$

FAR_0	C_{tuned}	$FAR_{observed}$	Precision	Recall
0.50000	0.001961	0.233681	0.003285	0.451351
0.10000	0.006116	0.029413	0.006130	0.106306
0.05000	0.008648	0.014699	0.007784	0.067568
0.01000	0.017988	0.003014	0.014247	0.025526
0.00100	0.051656	0.000327	0.064516	0.013213
0.00010	0.139054	0.000060	0.180556	0.007808
0.00001	0.333604	0.000009	0.400000	0.003604

5.5 Concluding Remarks

In this study, predictive monitoring using extreme gradient boosting (XGBoost) is investigated. We develop a procedure that can produce early warnings of problematic

events and can be tuned to deliver a desired false alarm rate.

Advances in data collection and machine learning techniques can improve process quality control by forecasting and monitoring potential process problems. XGboost is a recently developed powerful machine learning framework that efficiently combines weak learners to minimize an appropriate loss function. The predictive monitoring procedure is demonstrated using a real-life example on mental health in the Netherlands.

Predictive monitoring is an area of tremendous interest in mental health (Hahn et al., 2017). A unique non-public data set on mental health in the Netherlands was provided by Statistics Netherlands. We focused on predictive monitoring of mental health crises in people diagnosed with schizophrenia. These crises are harmful, frequent, and expensive, which motivated the need for early warnings of these events. All 75,000 people diagnosed with schizophrenia in the Netherlands were included in the study. The individual healthcare treatments, diagnoses, admissions, and incomes were aggregated on a weekly interval. Subsequently, we built explanatory variables on three levels of aggregation, short- (4 weeks), medium- (12 weeks), and long-term (64 weeks). The final data set consisted of more than 15 million person-weeks and 264 predictors.

The data was then used to predict the probability of a crisis in a future week. We compared the performance of logistic regression, hierarchical regression, and the extreme gradient boosting algorithm. All three methods showed predictive power, assigning a higher probability of a crisis to individuals that end up in crisis care in the coming week. The XGBoost framework achieved the highest discriminative capacity and was subsequently used in the monitoring procedure.

The predicted probabilities were monitored using a threshold value C. The procedure signals when the predicted probability exceeds this value. Each value of C will result in a number of true/false signals. A higher threshold will result in fewer false positives, but it will also miss more cases of crisis in the monitoring phase. We propose a search-algorithm to find a value for C that results in a desired false alarm rate. This algorithm uses cross-validation on the training data and delivers good results in this case study.

We also considered the monitoring performance looking up to 10 weeks ahead.

More specifically, the average estimated difference in probabilities in the weeks leading up to a crisis was 135 times higher than for weeks that do not lead to a crisis. This can be used by practitioners to produce early warnings of crises in people diagnosed with schizophrenia.

Mental health crises on a weekly basis are rare, occurring in less than 0.3% of the recorded cases. The predictive monitoring procedure shows promising results, although a high degree of uncertainty remains. The administrative changes in the final year of the data made predicting crises more challenging. Further tuning of the parameters could produce more accurate results but are out of the scope of this study. The proposed tuning algorithm provides a tool for practitioners to configure the monitoring procedure based on the available capacity.

Some challenges in the application of the predictive monitoring procedure in mental health remain. The data wrangling operation is extensive. The administration of treatments and diagnoses is complicated causing inconsistencies among healthcare providers. The facilities to process and clean the numerous protected Microdata sources are currently limited in the Netherlands. Improving the consistency and improving the computational facilities will boost the predictive performance.

A logical avenue for further research is the application of recurrent neural networks to a similar data set. This requires more computational capacity than was available in this study. Furthermore, applying the predictive monitoring procedure to a process with a less sparse outcome is of interest.

In summary, predictive monitoring using XGBoost can produce good results in the mental health domain, as well as other areas. It can be tuned to achieve a desired false alarm rate and is capable of handling large amounts of (sparse) data. The tuning procedure and unique data set in this study represent a new direction in process monitoring.

Chapter 6

Multilevel Predictive Process Monitoring in Education

6.1 Motivation

"Early Warning Indicator Reports were invaluable to the success of our school" (high school principal, a quote from the Strategic Data Project Report by Becker et al., 2014). These early warning indicator reports monitor students throughout their school career and warn teachers and staff of students with high dropout risks. According to Romero & Ventura (2019), such early identification of vulnerable students who are prone to fail or drop their courses is crucial for the success of any learning method. Also, monitoring allows for the identification of students who are insufficiently challenged and will benefit from more stimulating classroom material.

Navigating the large body of literature in Statistical Process Monitoring (SPM), predictive monitoring and educational data mining is a daunting task when looking for answers as to what metrics should be monitored and which methods should be implemented.

Multilevel modeling is often a good method in educational settings and can be used for predictive monitoring in quality control. In this chapter, we demonstrate such a procedure and aim to guide researchers and practitioners in monitoring student performance, specifically in a high school setting. To achieve this, we work closely with a Dutch high school to answer the following questions 1) What determines student performance? 2) How can SPM be used in monitoring student progress? 3) What

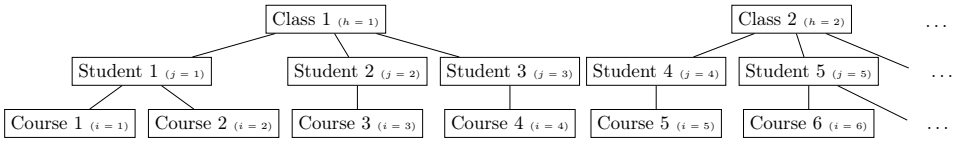


Figure 6.1 – The hierarchical structure of the case study data with classes as the top level. Students within these classes are the middle level and courses followed by these students form the lower level.

method can be used for predictive monitoring of student results? This chapter has been based on Huberts et al. (2020c).

6.1.1 Statistical Process Monitoring

A method that is used for multivariate processes is profile monitoring. Profile monitoring checks the stability of the modeled relationship between a response variable and one or more explanatory variables over time. Often profile monitoring uses regression control charts that were first introduced by Mandel (1969). The current body of regression control charting literature almost exclusively handles the monitoring of linear profiles using classical regression models. Weese et al. (2016) noted that large data sets often contain complex relationships and patterns over time, such as hierarchical structures and autocorrelation.

The case study presented in this chapter contains complex relationships and patterns, notably the hierarchical structure of courses, students, and classes (see Figure 6.1). State-of-the-art multivariate control charting based on linear regression models ignores this structure. However, incorporating hierarchical structures into the models can improve the reliability of a monitoring system. Therefore, we will develop a control chart that can signal at three levels, the class, student, and course level. Also, Woodall & Montgomery (2014) gave an overview of current directions in SPM and highlighted profile monitoring with multiple profiles per group as a topic for further research.

The advantage of using a hierarchical model is an improved estimation of process variability; according to Gelman (2006), hierarchical modeling is almost always an improvement compared to classical regression. The reason is that a hierarchical model includes the effects of both observed and unobserved variables, where unobserved vari-

ables are not explicitly measured but inherent to the group. Another advantage over classical regression is that a multilevel model provides a way to monitor new groups since the model generates some prior beliefs upon which to base the distribution and the prediction for the new groups. Furthermore, in contrast with classical regression, multilevel modeling is capable of prediction for groups with a small number of observations.

Multilevel models have been used in agricultural and educational applications for decades (Henderson et al., 1959; Aitkin & Longford, 1986; Bock, 1989; Aaronson, 1998; Sellström & Bremberg, 2006).

Today, hierarchical models are used in spatial data modeling (Banerjee et al., 2014), extreme value modeling (Sang & Gelfand, 2009), quantum mechanics (Berendsen, 2007) and even in the modeling of intimacy in marriage (Laurenceau et al., 2005). However, to the best of our knowledge, multilevel modeling has not found its way to SPM. Schirru et al. (2010) modeled multistream processes in semiconductor manufacturing using a multilevel model, but it is only applicable to two levels. Qiu et al. (2010) considered nonparametric profile monitoring using mixed-effects modeling, although they did not consider hierarchical modeling.

This chapter will explore process monitoring for a school data set that contains the grades of students in different groups over time. The school is interested in monitoring deviations in student results from what is given by the model, which is a form of profile monitoring. Therefore, we will investigate SPM based on hierarchical Bayesian models. In the next section, we will discuss the use of a hierarchical model to predict outlying results on the student level.

6.1.2 Predictive Monitoring

Becker et al. (2014) emphasized the need for actionable predictive analytics in high schools to keep students on track toward graduation and better prepare them for college and career success. The report discussed three examples of early warning indicator systems that help school teachers and management with early identification of students with a lower probability of passing, based on logistic regressions of student grade and attendance information.

Early prediction of learning performance has gained more traction in the literature,

as showcased by a recent special issue of IEEE Transactions on learning technologies. Together with monitoring big and complex data, predictive monitoring is recently being considered in quality technology literature (for example Kang et al., 2018; Wang et al., 2019). Although our case study focuses on the use of predictive monitoring to improve the quality of education, the presented methods can be used in any setting where clear hierarchical data structures exist. Baghdadi et al. (2019) stated that the ability to estimate when the performance will deteriorate and what type of intervention optimizes recovery can improve the quality and productivity and reduce risk concerning worker fatigue. Our case study offers a very similar approach to improve the quality and productivity of high school education by monitoring student performance.

The hierarchical model will thus be applied in two ways. First, control charting is applied based on the multilevel model. Second, the multilevel model is used for predicting results on the student level. We will compare the results of the multilevel model to one-level regression and the appropriate machine learning method. The final results present a hierarchical early warning indicator system, that can be applied in schools for predictive monitoring of student outcomes.

The outline of this chapter is as follows. The next section describes the relevant educational literature, the practical problem we aim to solve, and the available data. The hierarchical model and its performance are discussed in the section after this, followed by a section that investigates student performance monitoring. The last section summarizes the results.

6.2 Problem Description

In this section, we describe related student performance literature, the goal of the method to be developed, and the data set including the predictor variables.

6.2.1 Student Performance Literature

This section will shortly discuss a selection of determinants of student performance, whose selection has been based on a literature study. The determinants, their expected effects on performance, and their modeling approach are summarized in Table 6.1.

The important variables will be used in the modeling approaches of later sections. The 'unobserved' variables represent variables that were not available in this study, but the hierarchical modeling specification incorporates many of these 'unobserved differences' between students and students within courses.

Table 6.1 – Summary of determinants of student performance according to the literature and modeling approach

Determinant	Effect on performance		
	Student level	Class level	Modeling approach
SES		+	Explanatory variable
Disabilities	−		Explanatory variable
Language	+/−		Explanatory variable
Non-native	+/−	−	Explanatory variable
Student effort	+	+	Student unobserved heterogeneity
Peer associations	+/−	+/−	Student/course unobserved heterogeneity
Parent involvement	+		Student unobserved heterogeneity
School climate	+/−	+/−	Course unobserved heterogeneity
Intelligence	+		Explanatory variable, student unobserved heterogeneity
Grades	+		Time varying explanatory/dependent variable
Absences	−	−	Time varying explanatory variable

Nichols (2003) found a significant relationship between poor performance at the beginning of students' educational careers and later on. Furthermore, students who struggle academically had increased school absences, and students from lower-income families showed a higher probability of poor results. This suggests an important role for family income, absences, and temporal effects in predicting individual high school performance.

Socioeconomic status (SES) has long been argued to significantly affect school performance, although the importance varies greatly among different analyses. Geiser & Santelices (2007) argued omission of socioeconomic background factors can lead to significant overestimation of the predictive power of academic variables, that are strongly correlated with socioeconomic advantage. They based this assumption on a study by Rothstein (2004), which argued the exclusion of student background characteristics from prediction models inflates college admission tests' apparent validity by over 150 percent.

Disabilities can be a determinant of student performance. Dyslexic children fail to achieve school grades at a level that is commensurate with their intelligence (Karande & Kulkarni, 2005). Although they might not be directly linked to learning, disabilities like asthma, epilepsy, and autism can indirectly influence academic performance.

Autistic children can face a lot of problems in school as their core features impair learning. Furthermore, medical problems like visual impairment, hearing impairment, malnutrition, and low birth weight can cause difficulties in school.

The language that children speak at home can influence their academic abilities both positively (Buriel et al., 1998) and negatively (Kennedy & Park, 1994). Collier (1995) found that immigrants and language minority students need 4-12 years of second language development for the most advantaged students to reach deep academic proficiency and compete successfully with native speakers. It has been suggested that the presence of non-native speakers in schools harms the performance of native speakers, but this has been refuted by Geay et al. (2013). In contrast, children who interpret for their immigrant parents; 'language brokers', often perform better academically (Buriel et al., 1998).

Some variables remain unobserved but can be incorporated in models by allowing for unobserved heterogeneity. One is student effort, which is characterized by the level of school attachment, involvement, and commitment displayed by the student (Stewart, 2008). Also, peer influence, i.e. the associations between high school students, matter a great deal to individual academic achievement and development (Nichols & White, 2001). Besides, parent involvement is likely to influence academic achievement. Sui-Chu & Willms (1996) found that the most important dimension of parent involvement towards academic achievement is home discussion. They suggested facilitating home discussion by providing concrete information to the parents about parenting styles, teaching methods, and school curricula. Finally, school climate (a.o. Stewart, 2008) and intelligence (Rohde & Thompson, 2007; Laidra et al., 2007; Parker et al., 2006) are important for academic achievement.

Parent involvement, disciplinary climate, and individual intelligence are usually quite difficult to measure. This study aims to incorporate them nonetheless. Parent involvement is incorporated mostly in student unobserved heterogeneity. Limited observed information on the parents is included in the predictive model (i.e. education level and SES). Disciplinary climate and class disruptions are mostly covered by including absences that equate to dismissals from class and within unobserved course differences. Individual intelligence is approximated using primary school test scores.

Next, some time-varying variables are important. The first variable is the grade.

For each course, specific tests are taken with varying weights. Anytime during the year, these tests determine a current weighted average grade for each student and course. The resulting end-of-year grade is the most important student performance indicator. Also, absences are important as attending class helps students understand the material and motivates their participation (Rothman, 2001). The variables test grades and absences are generated over time. Finally, temporal effects on student performance encompass both inter-year changes and intra-year changes. Students will change the allocation of their effort and time according to their current average grade, their average grade for other courses, seasonal effects, within school changes, and external factors. Ideally, modeling will allow for student and course-specific effects to vary over time. The next section will describe the Dutch high school system.

6.2.2 The Dutch High School System

The Dutch school system in general consists of eight years of primary school, followed by four, five, or six years of high school. There is one level of primary school, but there are multiple levels of high school. Two criteria have been used in recent years to determine the level of high school a child is allowed to go to. Firstly, there is the teacher's advice. The teacher advises the level that fits the child in the final year of primary school. This advice is based on the performance of the child in a specific primary school.

Secondly, the National Institute for Test Development (in Dutch: Centraal Instituut voor Toets Ontwikkeling, abbreviated by CITO) test is a test that is developed by the CITO organization and is scientifically designed to test a child's academic abilities. It was initiated in the Netherlands by the famous psychologist professor A.D. de Groot in 1966 and every primary school is required to conduct the CITO or a similar test in the final year as of 2014.

To pass any specific year of high school, conditions set by the school have to be met. These conditions usually consist of requirements on the end-of-year average grades for all the student's courses. The grades in most Dutch high schools are on a scale from 1 to 10. The end-of-year grades are usually rounded, and a course is failed or 'insufficient' if the rounded grade is below 6. The amount of allowed 'failpoints', i.e. the total points below six, can then be restricted. A school might, for example,

have a student repeat the current year if he or she scores more than two failpoints, which could be a student with a grade of three for a single course or a four and a five or three fives at the end of the year. The restrictions are not limited to the number of failpoints. There can be requirements on the total average grade and certain subtleties emerge once the students start splitting up into high school profiles, where different students do a different set of courses from their fourth year on. These school profiles can have special requirements, usually with more importance assigned to the profile courses.

When implementing a predictive monitoring scheme in a school, the specific rules employed by a school define the passing probability that is estimated. For example, when a student is failing a profile course, this can lead to failing the year directly. If the same student would obtain the same grade for a different course, this would not necessarily mean failing the year. Therefore, different courses have different levels of importance to the probability of success for individual students. The school that has kindly provided the data described in the next section has different passing conditions for each year. Although the implementation at the school incorporates all conditions, the predictive analyses in this chapter reflect a simplified version to demonstrate the detective capabilities of the methods.

6.2.3 Data Set

A large, detailed data set was provided by a Dutch high school. In total there are eight years of data available, comprising of 36 different subjects followed by over 1,700 unique students (about 51% girls) and 711,653 individual tests. The students were born in 38 different countries, spoke 18 different languages, and were taught by 110 different teachers. Out of the unique students, 326 had some kind of disability while at school, 162 had a non-Dutch nationality and 51 students had a serious language barrier. The number of students with parents who have attended university or higher-level academics is 261 and 86% of students were residents of the large city that the school is located in during their time at the Dutch high school.

To incorporate socioeconomic status (SES) in this analysis, nation-wide social status data provided by the Dutch government was used. The relative SES score of a student using a country-wide ranking of his or her postal code was added to the data

set.

Learning disabilities that have been confirmed by the school are included in the data set. The most common learning disabilities reported in the data are Attention-Deficit/Hyperactivity Disorder (ADHD) and dyslexia.

The data used in this chapter contains grades that are on a 1-10 scale. Although easy to interpret, some difficulties arise when using these grades for modeling. First, as Figure 6.2 shows, there are peaks at integer grades and grades on a .5 scale. This is due to teachers grading on an integer or .5 point scale instead of using continuous grades. This becomes less of a problem with average grades, as they are eventually rounded but fairly continuous during the year.

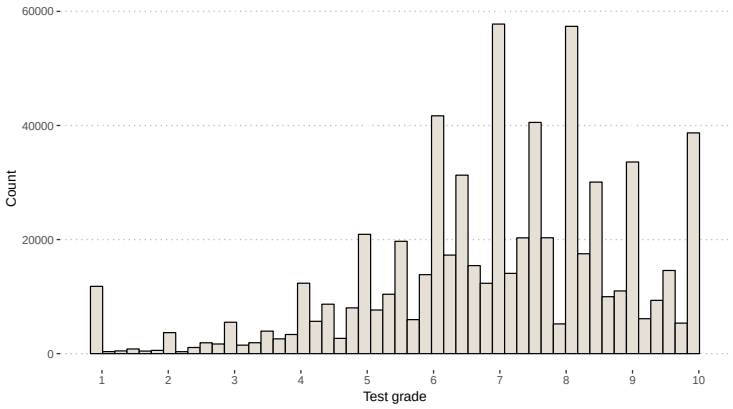


Figure 6.2 – Histogram of the individual test grades in the data

Second, when predicting the precise end-of-year grade, grades below 1 or above 10 should be impossible. However, both grades should have some positive probability, as some students do achieve average grades of 10 for specific courses during a year.

The following section describes the selected predictor variables in the data.

6.2.4 Determinants of Student Performance

We have discussed some of the literature on determinants of high school performance in Section 6.2.1. This section investigates these variables in the data.

The raw values for the most important categorical variables in the data are plotted in Figure 6.3. The first pair of boxplots in Figure 6.3 shows that girls seem to outperform boys in terms of final grades, which is consistent with the literature in

different settings (see Rahafar et al., 2016; Deary et al., 2007; Battin-Pearson et al., 2000, for examples of gender gap findings in academic achievement). The second pair of boxplots in Figure 6.3 indicates that students with a disability achieve lower end-of-year grades, consistent with the findings of Karande & Kulkarni (2005). Children of highly-educated parents seem to perform slightly better at this school in terms of final grades, as depicted in the third pair of boxplots in Figure 6.3.

In line with Buriel et al. (1998), children born outside of the Netherlands do not underperform as shown by the fourth pair of boxplots in Figure 6.3. Students with a different native language do achieve slightly lower grades in the data, supporting conclusions by Collier (1995) and Kennedy & Park (1994). The end-of-year grades are lower towards the end of high school, as indicated in Figure 6.3.

Figure 6.4 shows the two most important numerical independent variables plotted against the final grades. The CITO score has a positive correlation with grades as shown by the positive linear trend in Figure 6.4a. This makes sense, as the CITO test is designed as a predictor of individual intelligence. Furthermore, in line with Rothman (2001), more absences mean lower final grades in the data, as indicated by the negative linear trend in Figure 6.4b.

6.3 Hierarchical Model

The objective is to monitor student progress during the school year, where the school's main interest lies in signaling 'exceptional' students. Exceptional students can be both underperforming and overperforming students. In this section, we introduce a three-level hierarchical model for student grades and compare its performance to simpler models in monitoring student performance.

6.3.1 The Model

Throughout the year, students take tests for every course $i = 1, ..., n_0$. The grades for these tests are defined as $g_{ki} \in [1, 10]$ with $k = 1, .., K_i$, where K_i is the number of tests taken in course i. As these grades are obtained for individual tests, we have a set of cumulative weighted average grades $y_{i,j[i],h[j[i]]}$ for course i, student j and class h. For readability we drop subscripts j and h. The individual test results g_{ki}

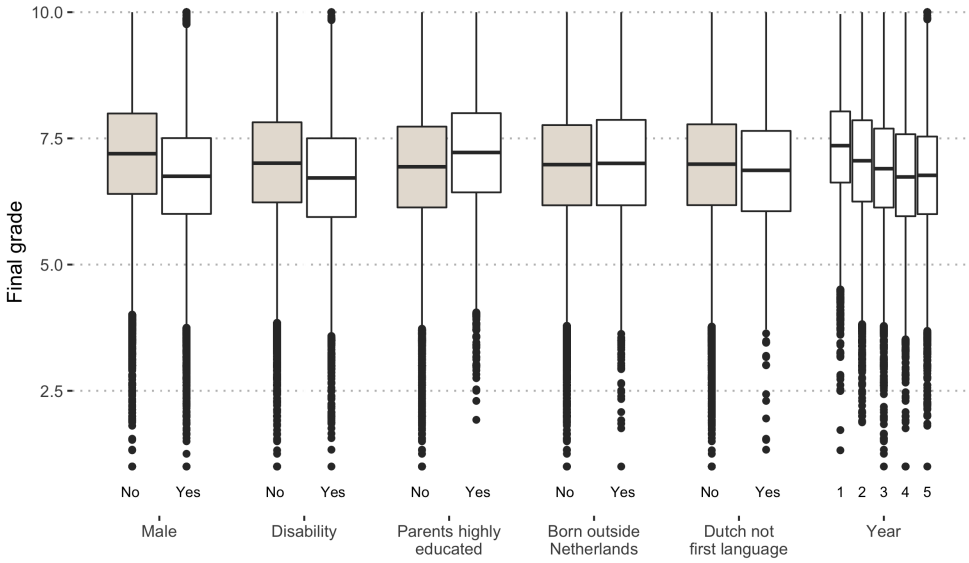


Figure 6.3 – Boxplots of the final grades for the most important categorical predictor variable

and the weights of the tests w_{ki} determine the average grade $y_i = \frac{\sum_{k=1}^{K_i} w_{ki} g_{ki}}{\sum_{k=1}^{K_i} w_{ki}}$, with $y_i \in [1, 10]$.

We consider a hierarchical model with three levels and use the index i ($i = 1, 2, ..., n_0$) to denote the individual course level, j ($j = 1, 2, ..., n_1$) to denote the individual student level and h ($h = 1, 2, ..., n_2$) for the class level (see Figure 6.1). We have p_0 predictors for the course level, p_1 for the student level and p_2 for the class level. We define row vectors $X_i^{(L_0)}, X_j^{(L_1)}$ and $X_h^{(L_2)}$, which consist of the intercept and predictor values for the course, student and class levels respectively.

We model cumulative weighted average grade y_i for course i as

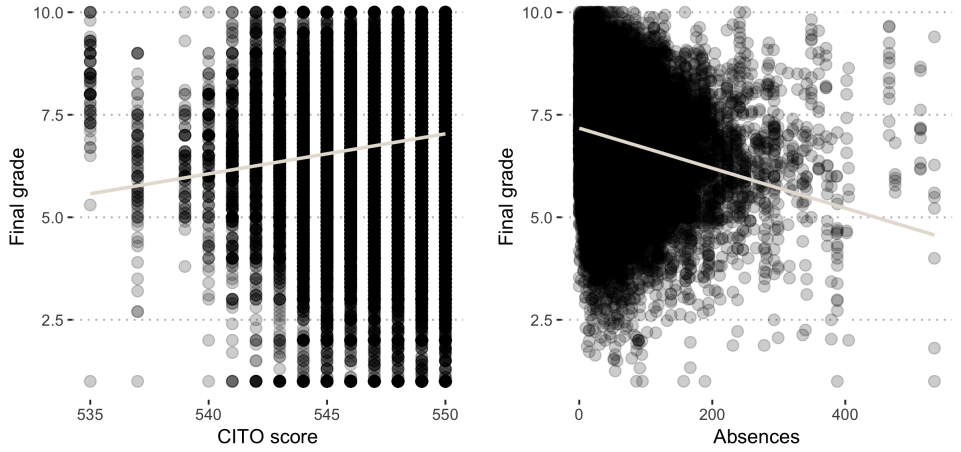
$$y_i \sim N(X_i^{(L_0)} \beta_{j[i]}^{(L_0)}, \sigma^2), \text{ for } i = 1, ..., n_0 \text{ (Course level)},$$

where the student levels are modelled as

$$\beta_j^{(L_0)} \sim N(\beta_{h[j]}^{(L_1)} X_j^{(L_1)\prime}, \Sigma^{(L_1)}), \text{ for } j = 1, ..., n_1 \text{ (Student level)},$$

and the class levels are specified by

$$vec(\beta_h^{(L_1)}) \sim N(\beta^{(L_2)} X_h^{(L_2)\prime}, \Sigma^{(L_2)}), \text{ for } h = 1, ..., n_2 \text{ (Class level)},$$



(a) Final grades versus CITO score

(b) Final grades versus total absences

Figure 6.4 – Scatterplots of the final grades and most important numerical variables with a linear trendline

where $X_i^{(L_0)}$ is a $1 \times (p_0 + 1)$ row vector of subject specific variables such as course content and level; $\beta_{j[i]}^{(L_0)}$ is a $(p_0 + 1) \times 1$ vector of parameters for student j that follows course i; σ^2 is the variance for the course level; $\beta_{h[j]}^{(L_1)}$ is a $(p_0 + 1) \times (p_1 + 1)$ parameter matrix determined by the class h that student j is in; $X_j^{(L_1)}$ is a $1 \times (p_1 + 1)$ row vector of student specific variables such as age, absences and IQ; $\Sigma^{(L_1)}$ is the covariance matrix for parameters $\beta_j^{(L_0)}$; $vec(\beta_h^{(L_1)})$ is the vectorized version of $\beta_h^{(L_1)}$ with dimensions $(p_0 + 1)(p_1 + 1) \times 1$; $\beta^{(L_2)}$ is a $(p_0 + 1)(p_1 + 1) \times (p_2 + 1)$ parameter matrix at the class level; $X_h^{(L_2)}$ is a $1 \times (p_2 + 1)$ row vector of class specific variables such as class size; and $\Sigma^{(L_2)}$ is the covariance matrix for parameters $\beta_h^{(L_1)}$.

6.3.2 Estimation

The parameters of a multilevel model can be estimated using, among other methods, maximum likelihood, generalized least squares, and Bayesian theory (Hox et al., 2017). A discussion of Bayesian and likelihood-based techniques for multilevel models was given by Browne & Draper (2006). These authors show that Bayesian estimation often provides an improvement over likelihood methods in terms of both point and interval estimates as well as the posterior distributions for the parameters. We use

Bayesian estimation to estimate the parameters in this chapter.

The full parameter space $\{\beta^{(L_0)}, \sigma^2, \beta^{(L_1)}, \Sigma^{(L_1)}, \beta^{(L_2)}, \Sigma^{(L_2)}\}$, where $\beta^{(L_0)}$ and $\beta^{(L_1)}$ are constructed by stacking the parameter matrices $\beta_j^{(L_0)}$ and $\beta_h^{(L_1)}$ for all groups j and h respectively, can be estimated based on data that are considered representative, i.e. in control. To estimate the parameters, we use the Bayesian method applying Markov Chain Monte Carlo (MCMC) methods which use the Gibbs sampling procedure. These methods are described in the appendix and are applied using the rJAGS package to link to JAGS (Plummer, 2018).

As the number of parameters increases quickly with added group levels, estimation time increases greatly as well. Thus when defining a multilevel model, there is a tradeoff between added precision and the additional estimation time for a group level. In a two-level model, the number of parameters we need to estimate is 1 for σ^2, $(p_0 + 1)(p_1 + 1)$ for $\beta^{(L_1)}$ and $\frac{1}{2}(p_0 + 1)(p_0 + 2)$ for $\Sigma^{(L_1)}$ ($\beta^{(L_0)}$ is constructed using the estimates for $\beta^{(L_1)}$). For the three-level model this increases, with 1 for σ^2, $\frac{1}{2}(p_0+1)(p_0+2)$ for $\Sigma^{(L_1)}$, $(p_0+1)(p_1+1)(p_2+1)$ for $\beta^{(L_2)}$ and $\frac{1}{2}(p_0+1)(p_1+1)((p_0+1)(p_1+1)+1)$ for $\Sigma^{(L_2)}$ ($\beta^{(L_0)}$ and $\beta^{(L_1)}$ are constructed using the estimates for $\beta^{(L_2)}$). For example, if there are three parameters per level, the number of parameters is 27 for a two-level model and 211 for a three-level model.

After applying the estimation procedure as described in the appendix, we obtain the estimations for the parameters in the three-level model, which we denote by $\{\hat{\beta}^{(L_0)}, \hat{\sigma}^2, \hat{\beta}^{(L_1)}, \hat{\Sigma}^{(L_1)}, \hat{\beta}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$. Later on we can use this three-level model for monitoring the relationships given by the model as well as for predicting results.

6.3.3 Recurrent Neural Network

Recurrent neural networks (RNNs) are a family of neural networks designed to process sequential data sources (Goodfellow et al., 2016). RNNs model a sequence of steps and allows previous outputs to be inputs. The long short-term memory (LSTM) model introduced by Hochreiter & Schmidhuber (1996) is an RNN that has been very successful in recent years (Lipton et al., 2015). It is a gated version of an RNN that adds input-, output- and forget-gates that regulate the flow of information in cells. For details on the LSTM model we refer to Hochreiter & Schmidhuber (1996) and Lipton et al. (2015). We will fit an LSTM model of the 'many-to-one' type as depicted

in Figure 6.5 using Keras in R (Allaire & Chollet, 2019).

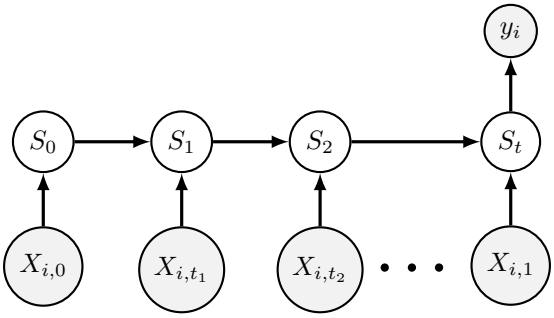


Figure 6.5 – Illustration of a many-to-one long short-term memory model. The inputs are represented by $X_{i,t}$ at various times t during the year, the cells with hidden states by S and y_i represents the single output

6.3.4 Results

In this section, we consider the accuracy of the end-of-year average grade estimates for $N = 3,839$ courses and 268 students during the school year 2014/2015. This subset consists of the first-, second- and third-year students. In the fourth year, students choose a profile, which changes the class compositions. The five school years from 2009 to 2014 are used to estimate the parameters.

As benchmarks, we consider using the weighted average grade (y_i) and a simple one-level linear regression model ($\hat{y}_{sr}$) to predict. The one-level linear regression fits $y_i = X_i\beta + \varepsilon_i$ using the same predictors as the multilevel specification.

As measures of accuracy, we report the Root Mean Squared Errors (RMSE) and the Nearest Neighbors proportions (NN). The RMSE is calculated as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i \in N}(y_i - \hat{y}_i)^2}, \tag{6.1}$$

with i identifying all the predicted grades and N the total number of grades. The RMSE score strongly punishes large errors. The second measure of performance is nearest neighbors percentage (NN)

$$\text{NN} = \frac{1}{N}\sum_{i \in N} I(\hat{y}_i - 1 \leqslant y_i \leqslant \hat{y}_i + 1). \tag{6.2}$$

Note that an alternative criterion is the Mean Absolute Deviation (MAD). However, those results were comparable to the RMSE.

Table 6.2 – RMSE and NN results for the predictions of the 2014/2015 end-of-year grades of 268 students using the average grade (y_i), the simple regression ($\hat{y}_{sr}$), the hierarchical specification ($\hat{y}_H$) and the recurrent neural network ($\hat{y}_{RNN}$)

Time	RMSE				NN			
t	y_i	$\hat{y}_{sr}$	$\hat{y}_H$	$\hat{y}_{RNN}$	y_i	$\hat{y}_{sr}$	$\hat{y}_H$	$\hat{y}_{RNN}$
0	-	1.15	0.86	1.14	-	0.80	0.90	0.80
0.1	1.53	1.07	0.84	1.07	0.70	0.83	0.91	0.83
0.3	1.04	0.83	0.74	0.82	0.86	0.92	0.94	0.92
0.5	0.77	0.67	0.65	0.65	0.93	0.96	0.96	0.97
0.7	0.51	0.48	0.47	0.53	0.98	0.98	0.98	0.98

Table 6.2 reports the RMSE and NN for the recurrent neural network ($\hat{y}_{RNN}$), the hierarchical model ($\hat{y}_H$), the one-level linear regression fit ($\hat{y}_{sr}$) and the weighted average (y_i) at five points in time $t = 0, 0.1, 0.3, 0.5, 0.7$.

The two performance measures in Table 6.2 show the superiority of the hierarchical method $\hat{y}_H$ when predicting end-of-year grades at the beginning of the year ($t = 0$). As the year progresses, the relative advantage of the model decreases over time as more grades accumulate and the final grade is less uncertain. Table 6.2 also shows that the performance of the RNN model of Section 6.3.3 is very similar to the one-level regression model.

A comparison of Tables 6.3 and 6.4 clarifies the advantage of the hierarchical regression model compared to a one-level model. Both tables show the predicted and realized end-of-year grades before the start of the year. The difference in RMSE of 0.292 might not seem worth the trouble at first, but when we compare these two tables, Table 6.4 shows much more granularity in the results. The hierarchical model identifies much more structure in the data, which is especially valuable in predicting far above- and below-average grades.

6.4 Monitoring Student Performance

This section is about monitoring student performance using accumulated test grades. We will consider SPM techniques and predictive monitoring.

Table 6.3 – Confusion matrix of the predictions for the 2014/2015 end-of-year grades of 268 students based on the simple linear regression model at $t = 0$

		Actual grades							
		3	4	5	6	7	8	9	10
Predicted	6	0	1	0	0	0	0	2	0
	7	9	53	208	722	962	747	283	33
	8	0	6	20	134	255	252	140	12

Table 6.4 – Confusion matrix of the predictions for the 2014/2015 end-of-year grades of 268 students based on the three-level model at $t = 0$

		RMSE = 0.860							
		Actual grades							
		3	4	5	6	7	8	9	10
Predicted	3	0	1	1	0	0	0	0	0
	4	0	1	3	2	0	0	0	0
	5	3	10	19	27	11	2	0	0
	6	4	36	114	358	182	55	10	0
	7	2	10	83	425	749	434	79	3
	8	0	2	8	43	267	464	213	14
	9	0	0	0	1	8	44	118	22
	10	0	0	0	0	0	0	5	6

6.4.1 Statistical Process Monitoring

To use a classical control chart technique (i.e. the Shewhart, CUSUM, or EWMA charts of Section 1.1.1) we need a Phase I data set that serves as a training set and a Phase II data set that will be a test set (Vining, 2009). Phase I is used to analyze the model and to estimate the parameters involved. The data used are assumed to be in control, and monitoring begins in Phase II. In this case, and many other practical examples, there is no obvious Phase I at hand. We could use student data from previous years as Phase I. These are not available however, for first-year students, for new courses, and in case of limited data. Furthermore, a second-year course is different from a first-year course and most students don't repeat a year. Identifying a clear Phase I/Phase II setup is thus difficult. These problems are amplified by the

fact that y_i is not i.i.d., violating the assumptions of the basic use of charts.

By modeling y_i, we can correct for a lot of the problems we see for classical control charting techniques. We model y_i at time t using all test grades before time t, with $t \in \{t_I, T\}$ where t_I indicates the start of the school year and T the end of the school year. We then calculate an expected value $\hat{y}_i$. The difference between the expected value and the actual observed value y_i at time t can then be monitored in a Phase II data set using a residuals control chart setup.

6.4.1.1 Three-level Control Chart

In this case, we evaluate whether the relations given by the three-level model still hold. To this end, we monitor the residuals at the three levels. For existing groups, we have estimates of the full parameter space $\{\hat{\beta}^{(L_0)}, \hat{\sigma}^2, \hat{\beta}^{(L_1)}, \hat{\Sigma}^{(L_1)}, \hat{\beta}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$. Then using these estimated parameters, we can calculate the residuals for the three levels for any new observation $\{y_i, X_i^{(L_0)}, X_j^{(L_1)}, X_h^{(L_2)}\}$

$$
\begin{aligned}
r_i^{(L_0)} &= y_i - X_i^{(L_0)} \hat{\beta}_{j[i]}^{(L_0)} \\
r_j^{(L_1)} &= \hat{\beta}_j^{(L_0)} - \hat{\beta}_{h[j]}^{(L_1)} X_j^{(L_1)\prime}, \\
r_h^{(L_2)} &= vec(\hat{\beta}_h^{(L_1)}) - \hat{\beta}^{(L_2)} X_h^{(L_2)\prime},
\end{aligned}
$$

where $r_i^{(L_0)}, r_j^{(L_1)}$ and $r_h^{(L_2)}$ are the residual vectors at the three levels of size 1, (p_0+1) and $(p_0+1)(p_1+1)$ respectively.

In line with traditional SPM techniques, we want to determine if a new observation stems from the in-control Phase I distribution, which was obtained using estimation (i.e. Phase I) data $\{X_I^{(L_0)}, X_I^{(L_1)}, X_I^{(L_2)}, y_I\}$ of size n_0, where $X_I^{(L_0)}$ is the $n_0 \times (p_0+1)$ matrix with the ith row containing the intercept and predictor values for course i. The other matrices are constructed in a similar way. The residuals can be monitored using control charting techniques.

For example, we can use a Shewhart control chart taking the mean and variance estimates from Phase I for $r_i^{(L_0)}$ with upper and lower control limits $\widehat{UCL}_y = 3\hat{\sigma}^2$ and $\widehat{LCL}_y = -3\hat{\sigma}^2$. The chart signals when the residual exceeds one of the control limits, after which the underlying cause can be investigated.

For $r_j^{(L_1)}$ and $r_h^{(L_2)}$, multivariate control charts are needed because these residuals are multidimensional. A multivariate Hotelling T^2 chart offers a solution with test

statistics (cf. 11.23 in Montgomery, 2007)

$$T^2_{(L_1)} = n_0 r_j^{(L_1)\prime} \hat{\Sigma}^{(L_1)} r_j^{(L_1)}, \tag{6.3}$$

$$T^2_{(L_2)} = n_0 r_h^{(L_2)\prime} \hat{\Sigma}^{(L_2)} r_h^{(L_2)}, \tag{6.4}$$

where n_0 is the number of observations used to estimate the covariance matrix. The lower control limit for these T^2 charts is $LCL = 0$, the upper control limit with false alarm percentage α is $UCL_{(L_1)} = \frac{p_1(n_0-1)}{n_0-p_1} F_{\alpha,p_1,n_0-p_1}$ for $T^2_{(L_1)}$ and $UCL_{(L_2)} = \frac{p_2(n_0-1)}{n_0-p_2} F_{\alpha,p_2,n_0-p_2}$ for $T^2_{(L_2)}$.

If the $T^2_{(L_2)}$ chart gives a signal, the root cause analysis can focus on the class level; if the $T^2_{(L_1)}$ chart gives a signal the root cause analysis can focus on the student level; and if the Shewhart chart gives a signal, the root cause analysis can focus on the course level.

Besides monitoring the residuals, there is the option of monitoring the parameter estimates. Similar to Kang & Albin (2000), a T^2 chart can be used to monitor the parameter estimates $\{\hat{\beta}^{L_0}, \hat{\sigma}^2, vec(\hat{\beta}^{(L_1)}), \hat{\Sigma}^{(L_1)}, \hat{\beta}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$.

6.4.1.2 Example

To illustrate this three-level monitoring approach, we monitor the cumulative weighted average y_i at 15 times throughout the school year 2014/2015 using the same subset as in the previous. Phase I consists of the five school years from 2009 to 2014; Phase II is the school year 2014/2015 for the 3,839 courses followed by 268 first-, second- and third-year students. We apply the hierarchical regression model and monitor the residuals using a Shewhart control chart.

The school aims to detect 'exceptional' courses and students. It considers exceptional courses as final grades below 6 or above 8. Each point below 6 is counted as a 'failpoint'. A single course with an end-of-year grade 5 equals 1 failpoint; a single course with an end-of-year grade 3 equals 3 failpoints, and one course grade of 4 and one of 3 equals 5 failpoints, etc. On the other hand, each point above 8 is counted as an 'excelpoint'. Thus the maximum grade of 10 for a course equals 2 excelpoints. An exceptional student is a student with at least four failpoints, and/or at least four excelpoints.

The three-level model estimates have an overall RMSE of 1.172. Figure 6.6 displays

an example of a Shewhart chart monitoring the residuals of the first level $r_i^{(L_0)}$. The chart signals four times near the end of the year. In total, the residuals charts signal 190 times (88 of which (46.32%) are exceptional courses), for 112 different students (36 of which (32.14%) are exceptional students).

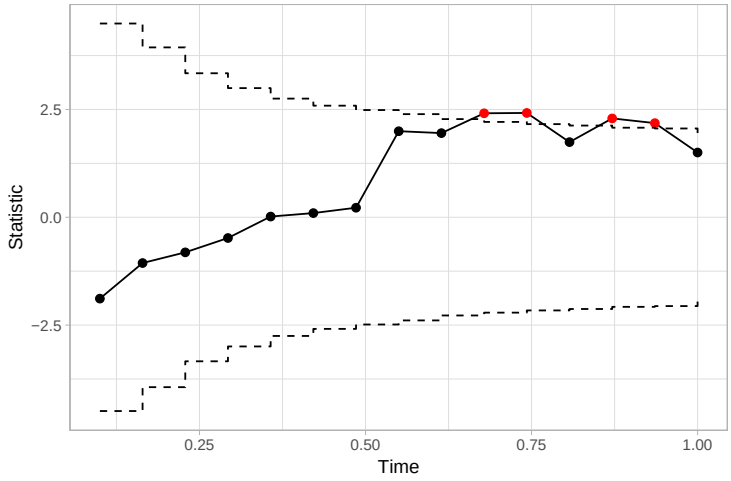


Figure 6.6 – Residual Shewhart control chart monitoring $r_i^{(L_0)}$ based on a three-level regression (signals in red)

As given by Equation (6.3), we can also monitor the student level residuals using a Hotelling T^2 chart. Using the same data as in the previous, the T^2 chart signals at least once for 105 students (38 (36.19%) of which are exceptional students).

The charts signal exceptional cases throughout the year. However, we can not retrospectively determine if at the time of a signal there was some unknown factor that influenced the performance of student j for course i. We are thus unable to distinguish false from true signals. It does, however, out-of-the-box, identify students whom we know have interesting performance during the monitoring phase.

The statistical monitoring approach identifies *incidental* anomalies in the weighted averages. However, the school's main focus is to identify students who need either support or more challenging coursework. This monitoring approach is insufficient for that goal. Therefore, in the next section, we use the hierarchical model to monitor student *expected* end-of-year results to identify under- or overperforming students.

6.4.2 Predictive Monitoring

The high school in this case study aims to predict the end-of-year grades of its students. This enables the school to receive early warnings on exceptional students. In this section, we will thus consider predictive monitoring of student performance.

6.4.2.1 Multilevel Predictive Monitoring

As demonstrated in Section 6.3.4, the predictions of the three-level model are relatively accurate. Furthermore, the three-level model can be used for new students/classes, and when there are a small number of courses per student or students per class. In this section, we will thus use the three-level model for predictive monitoring.

We want to monitor $P(E)_t$, defined as the probability of some event E at time t. $P(E)_t$ summarizes the outcome of the model into a single predictive probability at time t, with $t \in \{t_I, T\}$ where t_I indicates the start of the year and T the end of the year. The chart signals when $P(E)_t$ exceeds threshold C, which is defined as the maximum allowed probability of event E occurring ($0 < C < 1$). Event E concerns the values of y_i, which is context dependent and can take many forms ($y_i = e$, $y_i \geqslant e$, $y_i \leqslant e$, $e_1 \leqslant y_i \leqslant e_2$, $\sum_{i=a}^{b} y_i \geqslant e$ etc., where e, e_1 and e_2 are arbitrary constants and a and b are integers between 1 and n_0). Following the MCMC estimation of the posterior densities of the parameters $\theta = \{\beta^{(L_0)}, \sigma^2, \beta^{(L_1)}, \Sigma^{(L_1)}, \beta^{(L_2)}, \Sigma^{(L_2)}\}$ as described in the supplementary material, we can use the posterior densities to calculate $P(E)_t$.

The steps for predictive monitoring are

1. Define event E and threshold C
2. Specify the multilevel model for y_i
3. Estimate the parameters to obtain $\hat{\theta}_I$ using the Phase I data at time t_I using MCMC, described in the appendix
4. Calculate $P(E)_t$ using the newly available observations at time $t > t_I$
5. Signal if $P(E)_t > C$
6. Re-estimate the parameters to obtain $\hat{\theta}_t$ using all available data at time t and go back to step 4 for a new timepoint $t_{II} > t$.

Assume that we have a large in-control Phase I data set $\{X_I^{(L_0)}, X_I^{(L_1)}, X_I^{(L_2)}, y_I\}$ at time $t = t_I$. At time $t < t_I$ we obtain the estimates for the parameters $\{\hat{\beta}^{(L_0)}, \hat{\sigma}^2, \hat{\beta}^{(L_1)}, \hat{\Sigma}^{(L_1)}, \hat{\beta}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$ based on observations in Phase I. As described in the appendix for the three-level model, using the estimates of the parameters, at any time $t > t_I$ we have a predicted distribution for the outcome variable $\hat{y}_{i,t}$

$$\hat{y}_{i,t} \sim N\Big((X_{i,t}^{(L_0)} \otimes X_{j[i,t]}^{(L_1)\prime})\hat{\beta}^{(L_2)} X_{h[j[i,t]]}^{(L_2)\prime},$$
$$(X_{i,t}^{(L_0)} \otimes X_{j[i,t]}^{(L_1)})\hat{\Sigma}^{(L_2)}(X_{i,t}^{(L_0)} \otimes X_{j[i,t]}^{(L_1)\prime}) + X_{i,t}^{(L_0)}\hat{\Sigma}^{(L_1)}X_{i,t}^{(L_0)\prime} + \hat{\sigma}^2\Big),$$

where $\otimes$ is the Kronecker product. We can use this result to estimate the probability of the outcome $P(E)_t$. The event E can take several forms. Suppose we consider $y_i \leqslant e$, i.e. we study that the grade y_i is less than e. The monitoring scheme we propose uses the posterior distribution of $\hat{y}_{i,t}$ to calculate the probability $P(E)_t$. The chart signals when $P(E)_t > C$, with C the threshold that determines the maximum allowed probability of event E.

Monitoring $P(E)_t$ requires periodic re-estimation of the parameters to incorporate newly available information at time t. Around the time event E occurs, the probability $P(E)_t$ converges to 1 if $t \to T$. The major advantage of monitoring $P(E)_t$ instead of $y_{i,t}$ is that, depending on the predictive capability of the multilevel model, the monitoring scheme provides early warning and the opportunity to intervene before event E occurs. If intervention occurs, it is important to include this in the predictors $\{X^{(L_0)}, X^{(L_1)}, X^{(L_2)}\}$ by including an additional variable, to extract the effect of the intervention on outcome E. Furthermore, there is no need for n_0 control charts. All that is required is a single control chart plotting values of $P(E)_t$ and signaling for observations or groups for which $P(E)_t$ exceeds C.

6.4.2.2 Example

Following the steps outlined before, we define two events: E^f as a student failing the year and E^e as a student excelling that year. E^f occurs if a student has four or more failpoints, as defined in the previous section (the number of points below 6 for all courses a student follows in a year). E^e occurs if a student has four or more excelpoints (the number of points above 8 for all courses a student follows in a year).

The end-of-year rounded grade of student j for course i is defined as y_{ij}. At time

t, the probability of a student failing the year can thus be summarized by $P(E_j^f)_t = P(\sum_{i=1}^{n_j} \max(0, (6 - y_{ij})) \geqslant 4)_t$, where n_j is the number of courses for student j. The probability of a student excelling in the year can then be summarized by $P(E_j^e)_t = P(\sum_{i=1}^{n_j} \max(0, (y_{ij} - 8)) \geqslant 4)_t$ at time t.

Using the same data set as in the previous section, Figure 6.7 shows a control chart of $1 - P(E_j^f)_t$ for $J = 268$ students at 15 points in time. As an example, the threshold $C = 0.05$ is depicted as a dashed line. Note that $1 - P(E_j^f)_t$ equals the probability of passing the year. The $J_p = 238$ students who passed are depicted in blue and the probabilities of the $J_f = 30$ students who failed in red. Although there are some exceptions, overall the model consistently estimates the passing probabilities for the students who fail the year much lower than the students who pass the year. This can also be seen in the probabilities of failure in Table 6.5. This table reports the values of $\frac{1}{J_p} \sum_{j \in J_p} P(E_j^f)_t$ (the average estimated probability of failure for students that pass the year) in the top row and $\frac{1}{J_f} \sum_{j \in J_f} P(E_j^f)_t$ (the average estimated probability of failure for students that fail the year) in the bottom row. The model consistently assigns a higher average probability of failure to students that end up failing the year.

Figure 6.8 plots $P(E_j^e)_t$ for the same $J = 268$ students. The $J_n = 222$ students who did not excel are depicted in red and the probabilities of the $J_e = 46$ students who excelled are depicted in blue. As an example, threshold $C = 0.95$ is depicted as a dashed line. The model has impressive performance, shown also by the differences in average probabilities over time between students who excel, $\frac{1}{J_e} \sum_{j \in J_e} P(E_j^e)_t$, and those that do not, $\frac{1}{J_n} \sum_{j \in J_n} P(E_j^e)_t$, as depicted in Table 6.6.

Table 6.5 – Average estimated probabilities of failing $P(E^f)_t$ for 268 students in 2014/2015, split by observed outcome

	Time						
Failed	0	0.1	0.3	0.5	0.7	0.9	1
No	0.02	0.02	0.04	0.03	0.03	0.01	0.00
Yes	0.27	0.28	0.52	0.61	0.75	0.79	1.00

Depending on the threshold C that determines if the monitoring scheme signals, the model correctly identifies several students who will fail/excel as well as some false positives. Tables 6.7 and 6.8 report the precision and recall values as defined in

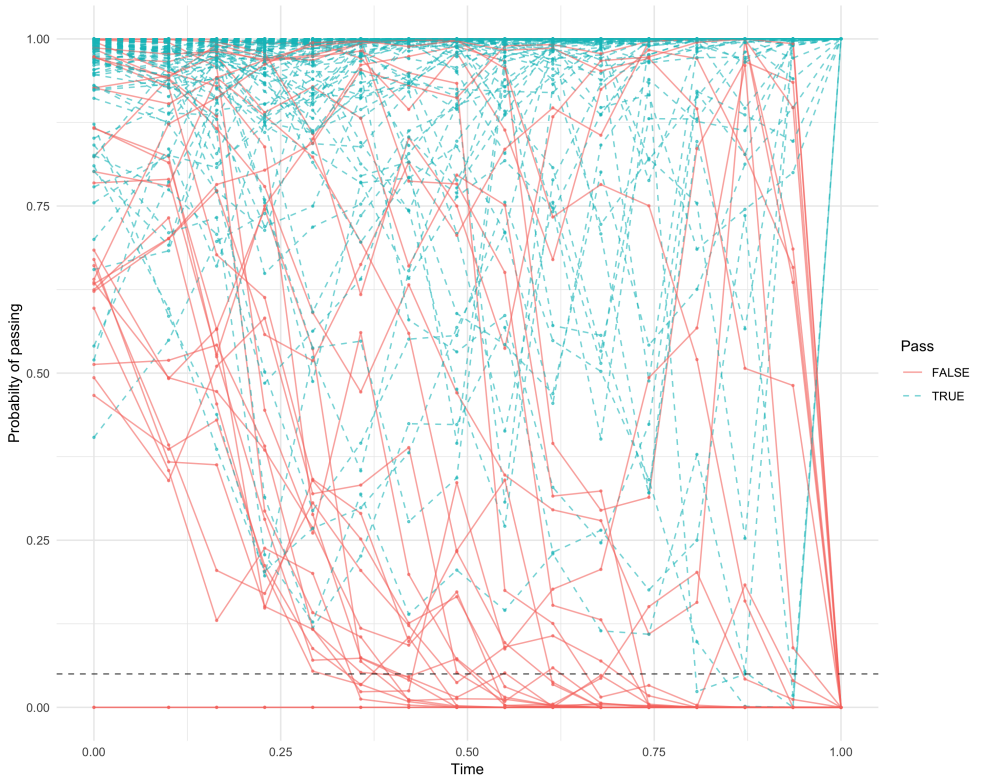


Figure 6.7 – A control chart monitoring the estimated probabilities of passing $1 - P(E^f)_t$ for 268 students in 2014/2015, with dashed threshold $C = 0.05$ in black. The dashed blue lines represent students who passed, the red solid lines students who failed

Section 1.2 when monitoring E^f and E^e respectively.

Table 6.7 shows the procedure correctly identifies students who will fail the year early on. The performance is impressive, where, depending on the chosen level of C, multiple early warnings are generated aiding in the student support system. For example, setting C at 0.75, the procedure identifies almost half (14 out of 30) of the students who will fail before the start of the year with only 26% (5) false positives.

Table 6.8 shows the precision and recall values when predicting excelling students. Depending on the school's preferences, high precision or recall can be achieved early on in the year. For example, setting C at 0.50, the procedure identifies half (23 out of 46) of the students who will excel before the start of the year with only 15% (4) false positives.

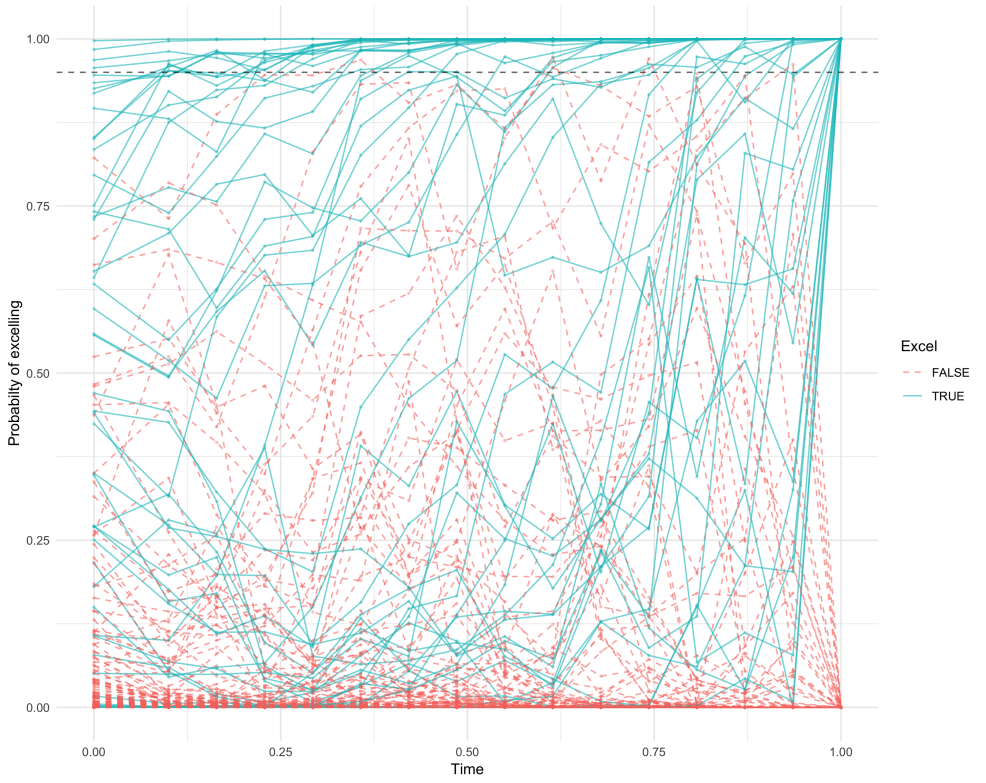


Figure 6.8 – A control chart monitoring the estimated probabilities of excelling $P(E^e)_t$ for 268 students in 2014/2015, with dashed threshold $C = 0.95$ in black. The solid blue lines represent students who excelled, the red dashed lines students who did not excel

The multilevel monitoring procedure has shown its value in a high school setting, as it adequately provides expected end-of-year grades for all students and subjects. This can aid in classifying at-risk students who need support, as well as the areas in which they need help. On the other side of the spectrum, the model successfully identifies excelling students who can benefit from more challenging schoolwork. The model further provides easily interpretable results, as well as good explainability for the parameters.

Table 6.6 – Average estimated probabilities of excelling $P(E^e)_t$ for 268 students in 2014/2015, split by observed outcome

	Time						
Excelled	0	0.1	0.3	0.5	0.7	0.9	1
No	0.05	0.04	0.04	0.06	0.04	0.03	0.00
Yes	0.50	0.49	0.50	0.61	0.67	0.81	1.00

Table 6.7 – Precision$_t(C)$ (Recall$_t(C)$) results when monitoring $P(E^f)_t$ with various values of C and t using the three-level model predictions of end-of-year grades for 268 students in 2014/2015

		C					
		0.05	0.1	0.25	0.5	0.75	0.999
Time	0	1 (0.07)	1 (0.07)	1 (0.07)	0.67 (0.13)	0.74 (0.47)	0.25 (0.93)
	0.1	1 (0.07)	1 (0.07)	1 (0.07)	1 (0.27)	0.71 (0.40)	0.25 (0.93)
	0.3	1 (0.10)	1 (0.20)	0.85 (0.37)	0.76 (0.53)	0.67 (0.67)	0.27 (1)
	0.5	1 (0.33)	1 (0.43)	0.94 (0.53)	0.79 (0.63)	0.67 (0.67)	0.34 (0.97)
	0.7	1 (0.57)	1 (0.63)	0.88 (0.73)	0.77 (0.70)	0.70 (0.77)	0.40 (0.97)
	0.9	0.90 (0.63)	0.86 (0.63)	0.88 (0.70)	0.81 (0.70)	0.81 (0.73)	0.59 (0.90)
	1	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)

6.5 Concluding Remarks

This study has considered three research questions concerning high school students' performance. We worked together with a Dutch high school in attempting to answer the following questions (1) What determines student performance? (2) How can SPM be used in monitoring student progress? (3) What method can be used for predictive monitoring of student results? This resulted in the use of a three-level model in a predictive monitoring scheme, that can be applied when monitoring hierarchical data. We discuss our results in the following section.

6.5.1 What Determines Student Performance?

The detailed data set made available by a Dutch high school has shown interesting determinants of student performance. These are generally in line with the educational literature and are useful when monitoring student progress.

Female students were found to obtain higher final grades. In line with the litera-

Table 6.8 – Precision$_t(C)$ (Recall$_t(C)$) results when monitoring $P(E^e)_t$ with various values of C and t using the three-level model predictions of end-of-year grades for 268 students in 2014/2015

		C					
		0.99	0.95	0.75	0.5	0.25	0.01
Time	0	1 (0.02)	1 (0.09)	0.93 (0.3)	0.85 (0.5)	0.69 (0.72)	0.38 (0.89)
	0.1	1 (0.04)	1 (0.2)	0.94 (0.35)	0.72 (0.46)	0.71 (0.65)	0.45 (0.87)
	0.3	1 (0.09)	1 (0.28)	0.89 (0.37)	0.83 (0.54)	0.68 (0.54)	0.45 (0.85)
	0.5	1 (0.37)	1 (0.41)	0.92 (0.52)	0.77 (0.59)	0.65 (0.67)	0.47 (0.96)
	0.7	1 (0.43)	1 (0.48)	0.89 (0.54)	0.88 (0.61)	0.72 (0.78)	0.57 (0.93)
	0.9	1 (0.57)	1 (0.63)	0.94 (0.74)	0.88 (0.83)	0.8 (0.87)	0.64 (1)
	1	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)

ture, students with disabilities perform slightly worse. Children with highly-educated parents were found to outperform their peers with less-educated parents in this case study.

The nationality and language barrier variables represent an interesting case study of the discussed theory on immigrant and language barriers in academia. Consistent with work by Geay et al. (2013) and the "language broker" effect of Buriel et al. (1998), students born abroad achieve similar performance to their locally born peers. A serious language barrier seems to produce slightly lower grades. This, in turn, is consistent with findings by Kennedy & Park (1994) and Collier (1995).

Students show a decrease in performance through their high school career, with around half a point difference in grades between the first and fourth years of high school. Absences seem to have a strong negative correlation with grades. On a policy level, the relationship between the primary school test scores (CITO) and student grades should be considered towards current discussion around the determinants of the high school level.

The main goal of the school was to monitor student performance as the process output throughout the year. Therefore, statistical and predictive monitoring techniques were considered.

6.5.2 Statistical Process Monitoring

Classical SPM techniques are often insufficient when applied to complex processes, for which increasingly large data sets are available. When a hierarchical structure is present in the data set, multilevel modeling improves the reliability of process monitoring. Using multilevel models improves estimation accuracy and explainability over regular linear regression models. Furthermore, the method is essential for predictive modeling of new students/classes or students/classes with small sample sizes.

Univariate SPM techniques proved insufficient in this case study and one-level linear regression models did not provide satisfactory results. We have discussed a three-level model together with the monitoring options. Residual control charting at the three levels was proposed as the multilevel statistical monitoring method for online monitoring of process output. The proposed multilevel monitoring framework did provide promising results.

6.5.3 Predictive Monitoring

A predictive monitoring method has been developed to enable an early warning monitoring system. This method monitors the probability of an event, rather than a process output. The three-level model was used to continuously predict end-of-year individual grades. Using a Bayesian hierarchical model, probability distributions for the student outcomes are obtained. These can be used to monitor unwanted results in the form of under- and overperforming students using a single predictive control chart setup. This predictive monitoring approach was shown to be very useful in practice, as the school obtains valuable early warnings on both under- and overperforming students.

The proposed multilevel process monitoring framework can be useful across many applications, including industrial processes (batch production, multiple factories), market monitoring, HR analytics, sports, and more. Implementation of multilevel models can be challenging, however, especially in a Bayesian setting. Sampling procedures can be used to simplify the analysis. We have provided a full analysis of the three-level model and its estimation in the supplementary material, where we used Gibbs sampling to estimate the parameters. Using these parameters, predictions

were made for the monitoring period, after which the parameters can be updated to improve the predictive power of the model. Predictive monitoring results in early warning systems, that can greatly aid in early detection and prevention of special cause variation.

We argue the importance of predictive monitoring in general. As more and more data are available, the use of more complex models can extract more information towards valuable predictions. Summarizing complex processes into simple and interpretable results is essential. Multilevel modeling is one method that achieves this, which is applicable in cases where a clear hierarchy is present. There are of course many more statistical and machine learning methods that can be applied. We encourage research that investigates the use of these methods in a predictive monitoring setting.

Concluding this chapter, early warning indicator systems have the potential to improve the educational system at a low cost. These systems can add a layer of sophistication to school and teacher performance evaluation and work towards fulfilling individual student needs.

6.6 Appendix

6.6.A Predictive Distribution

If we represent the three-level model as

$$\begin{aligned} y_i &= X_i^{(L_0)}\beta_{j[i]}^{(L_0)} + \varepsilon_i^{(L_0)} \ , \varepsilon^{(L_0)} \sim N(0, \sigma_y^2) \\ \beta_j^{(L_0)} &= \beta_{h[j]}^{(L_1)} X_j^{(L_1)\prime} + \varepsilon_j^{(L_1)} \ , \varepsilon^{(L_1)} \sim N(0, \Sigma^{(L_1)}) \\ vec(\beta_h^{(L_1)}) &= \beta^{(L_2)} X_h^{(L_2)\prime} + \varepsilon_h^{(L_2)} \ , \varepsilon^{(L_2)} \sim N(0, \Sigma^{(L_2)}), \end{aligned} \tag{6.5}$$

we can summarize the model as

$$y_i = X_i^{(L_0)} vec^{-1}(\beta^{(L_2)} X_{h[j[i]]}^{(L_2)\prime}) X_{j[i]}^{(L_1)\prime} + X_i^{(L_0)} vec^{-1}(\varepsilon_h^{(L_2)}) X_{j[i]}^{(L_1)\prime} + X_i^{(L_0)} \varepsilon_{j[i]}^{(L_1)} + \varepsilon_i^{(L_0)}.$$

We obtain parameter estimates $\{\hat{\beta}^{(L_0)}, \hat{\sigma}^2, \hat{\beta}^{(L_1)}, \hat{\Sigma}^{(L_1)}, \hat{\beta}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$ using the observations during Phase I time period $t < t_I$. At any time $t > t_I$ we have a predicted distribution for the outcome variable $\hat{y}_{i,t}$. Considering the distributions of the error

terms $\hat{y}_{i,t}$ has a normal distribution

$$\hat{y}_{i,t} \sim N\Big((X^{(L_1)}_{j[i,t]} \otimes X^{(L_0)}_{i,t})\hat{\beta}^{(L_2)} X^{(L_2)\prime}_{h[j[i,t]]},$$
$$(X^{(L_1)}_{j[i,t]} \otimes X^{(L_0)}_{i,t})\hat{\Sigma}^{(L_2)}(X^{(L_1)}_{j[i,t]} \otimes X^{(L_0)}_{i,t})' + X^{(L_0)}_{i,t}\hat{\Sigma}^{(L_1)} X^{(L_0)\prime}_{i,t} + \hat{\sigma}^2\Big),$$

where $\otimes$ is the Kronecker product and we use the relationship $vec(ABC) = (C' \otimes A)vec(B)$.

6.6.B Prior Distributions

The full parameter space $\theta = \{\beta^{(L_0)}, \sigma^2, \beta^{(L_1)}, \Sigma^{(L_1)}, \beta^{(L_2)}, \Sigma^{(L_2)}\}$, where $\beta^{(L_0)}$ and $\beta^{(L_1)}$ are constructed by stacking the parameter matrices $\beta^{(L_0)}_j$ and $\beta^{(L_1)}_h$ for all groups j and h respectively, are estimated using the Gibbs sampler (Casella & George, 1992). The Gibbs sampler approximates the posterior distribution by sampling from the full conditional distributions of the parameters. We use the rJAGS package in R to link to JAGS (Plummer, 2018).

The estimation requires prior distributions for the unknown parameter space. Parameters $\beta^{(L_0)}$ and $\beta^{(L_1)}$ have priors given explicitely by the model. Proper diffuse priors are chosen for parameters $\{\sigma^2, \Sigma^{(L_1)}, \beta^{(L_2)}, \Sigma^{(L_2)}\}$.

The vector $vec(\beta^{(L_2)})$ has a multivariate normal prior $N(a, B)$, with diagonal covariance matrix B and larger values of B reflecting greater uncertainty. Thus proper but diffuse priors were determined, with $a = 0$ and $B = 1000I$, where I is the identity matrix.

The covariance matrix $\Sigma^{(L_1)}$ associated with level 1 student unobserved differences and the covariance matrix $\Sigma^{(L_2)}$ for unobserved group level 2 differences are both defined as positive definite matrices with Inverse Wishart priors $W^{-1}(C, (p_0 + 1) + 1)$ for $\Sigma^{(L_1)}$ and prior $W^{-1}(D, (p_0 + 1)(p_1 + 1) + 1)$ for $\Sigma^{(L_2)}$. C and D are diagonal matrices, where smaller values correspond to more diffuse priors. Values for these inverse Wishart distributions are set at $C = D = diag(0.001)$.

For the variance parameter σ^2 of the error term in the model the inverse Gamma distribution, $IG(a, b)$, was chosen. We use an uniformative prior, with parameters $a = 0.001; b = 1$; $\sigma^2 \sim IG(0.001, 1)$.

6.6.C Full Conditional Distributions

The Gibbs sampling procedure uses the full conditional distributions of the unknown parameter space. Although they are not necessary when using rJAGS (Plummer, 2018), we report them below to be used in a Gibbs sampler or similar Markov Chain Monte Carlo (MCMC) sampling methods.

The likelihood function of the n_0 observed grades is the joint density of the data conditional on the parameters.

$$L(\theta) = \prod_{i=1}^{n_0} f(y_i|\theta) = (2\pi)^{-n_0/2}\sigma^{-n_0}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n_0}(y_i - X_i^{(L_0)}\beta_{j[i]}^{(L_0)})^2\right).$$

Let vector Y of size n_0 contain the observed values y_i. The full conditional distributions of the individual parameters are each proportional to $p(Y, \theta)$:

$$f(Y,\theta) \propto f(\beta^{(L_2)})f(\sigma^2)f(\Sigma^{(L_1)})f(\Sigma^{(L_2)})\prod_{h=1}^{n_2} f(\beta_h^{(L_1)}|\Sigma^{(L_2)}, \beta^{(L_2)}) \prod_{j=1}^{n_1} f(\beta_{jh}|\Sigma^{(L_1)}, \beta_h^{(L_1)})\prod_{i=1}^{n_0} f(y_{ijh}|\beta_{jh}, \beta_h^{(L_1)}, \sigma^2). \tag{6.6}$$

We then calculate the full conditional distributions by multiplying the prior by the likelihood and simplifying.

Calculation of the Full Conditional Distribution of $\beta_j^{(L_0)}$

We calculate the full conditional distribution of $\beta_j^{(L_0)}$ using 6.6 and simplifying, i.e.

$$f(\beta_j^{(L_0)}, |Y, \sigma^2, \beta_h^{(L_1)}, \beta^{(L_2)}, \Sigma^{(L_1)}, \Sigma^{(L_2)}) \propto f(\beta_j^{(L_0)}|\Sigma_j, \beta_h^{(L_1)}) \times \prod_{i=1}^{N_j} f(y_{ij}|\beta_j^{(L_0)}, \beta_h^{(L_1)}, \sigma^2) \propto$$

$$N(V^{-1}M, V),$$

with $V = (\Sigma^{(L_1)-1} + \sigma^{-2}X_{i\in j}^{(L_0)\prime}X_{i\in j}^{(L_0)})^{-1}$ and $M = (\Sigma^{(L_1)-1}\beta_h^{(L_1)}X_j^{(L_1)\prime} + \sigma^{-2}X_{i\in j}^{(L_0)\prime}Y_{i\in j})$.

Calculation of the Full Conditional Distribution of $\beta_h^{(L_1)}$

We calculate the full conditional distribution of $vec(\beta_h^{(L_1)})$ using 6.6 and simplifying, giving

$$f(vec(\beta_h^{(L_1)})|Y,\sigma^2,\beta_j^{(L_0)},\beta^{(L_2)},\Sigma^{(L_1)},\Sigma^{(L_2)})\propto$$
$$f(vec(\beta_h^{(L_1)})|\Sigma^{(L_2)},\beta^{(L_2)})\prod_{j=1}^{N_{j\in h}} f(\beta_j^{(L_0)}|\Sigma^{(L_1)},\beta_h^{(L_1)})\prod_{i=1}^{N_{i\in h}} f(y_{ijh}|\beta_j^{(L_0)},\beta_h^{(L_1)},\sigma^2)\propto$$
$$exp\Bigg[-(vec(\beta_h^{(L_1)})-\beta^{(L_2)}X_h^{(L_2)\prime})'\Sigma^{(L_2)-1}(vec(\beta_h^{(L_1)})-\beta^{(L_2)}X_h^{(L_2)\prime})-$$
$$\sum_{j\in h}\left((\beta_j^{(L_0)}-\beta_h^{(L_1)}X_j^{(L_1)\prime})'\Sigma^{(L_1)-1}(\beta_j^{(L_0)}-\beta_h^{(L_1)}X_j^{(L_1)\prime})\right)\Bigg].$$

For further calculations, we define $A = vec(\beta_h^{(L_1)})$, $B = \beta^{(L_2)}X_h^{(L_2)\prime}$, $C = \Sigma^{(L_2)}$, $D_j = \beta_j^{(L_0)}$, E=$\Sigma^{(L_1)}$, $X_j = X_j^{(L_1)}$, this gives

$$f(vec(\beta_h^{(L_1)})|Y,\sigma^2,\beta_j^{(L_0)},\beta^{(L_2)},\Sigma^{(L_1)},\Sigma^{(L_2)})\propto P(A|B,C,D_j,E,X_j)\propto$$
$$exp\Bigg[-(A-B)'C^{-1}(A-B)-\sum_{j\in h}\Big[D_j'E^{-1}D_j-2(X_j\otimes(D_j'E^{-1}))A+A'(X_j\otimes(X_j'\otimes E))A\Big]\Bigg]\propto$$
$$exp\Bigg[-A'\left(C^{-1}+\sum_{j\in h}\left(X_j\otimes(X_j'\otimes E^{-1})\right)\right)A+A'(C^{-1}B)+\left(B'C^{-1}+2\sum_{j\in h}\left(X_j\otimes(D_j'E^{-1})\right)\right)A\Bigg],$$

which shows that the full conditional distribution of $vec(\beta_h^{(L_1)})$ is a multivariate normal distribution with covariance matrix

$$\left(\Sigma^{(L_2)-1}+\sum_{j\in h}\left(X_j^{(L_1)}\otimes(X_j^{(L_1)\prime}\otimes\Sigma^{(L_1)-1})\right)\right)^{-1}$$

and mean

$$\left(\Sigma^{(L_2)-1}+\sum_{j\in h}\left(X_j^{(L_1)}\otimes(X_j^{(L_1)\prime}\otimes\Sigma^{(L_1)-1})\right)\right)(\Sigma^{(L_2)-1}\beta^{(L_2)}X_h^{(L_2)\prime}).$$

Calculation of the Full Conditional Distribution of $\Sigma^{(L_1)}$

We calculate the full conditional distribution of $\Sigma^{(L_1)}$ using 6.6 and simplifying, i.e.

$$p(\Sigma^{(L_1)}|Y,\sigma^2,\beta_j^{(L_0)},\beta_h^{(L_1)},\beta^{(L_2)},\Sigma^{(L_2)})\propto f(\Sigma^{(L_1)})\prod_{j=1}^{n_1} f(\beta_j^{(L_0)}|\Sigma^{(L_1)},\beta_h^{(L_1)})\propto$$
$$|\Sigma^{(L_1)}|^{-(n_1+\nu+p_0+2)/2}exp(-tr((S^{L_1}+C)\Sigma^{(L_1)-1}/2)$$

with $S^{L_1}=\sum_{j=1}^{n_1}(\beta_j^{L_0}-\beta_{h[j]}^{(L_1)}X_j^{(L_1)\prime})'(\beta_j^{L_0}-\beta_{h[j]}^{L_1}X_j^{(L_1)\prime})$, which shows that the full conditional distribution of $\Sigma^{(L_1)}$ is $W^{-1}(S^{L_1}+C,n_1+\nu)$.

Calculation of the Full Conditional Distribution of $\Sigma^{(L_2)}$

We calculate the full conditional distribution of $\Sigma^{(L_2)}$ using 6.6 and simplifying, giving

$$p(\Sigma^{(L_2)}|Y,\sigma^2,\beta_j^{(L_0)},\beta_h^{(L_1)},\beta^{(L_2)},\Sigma^{(L_1)})\propto f(\Sigma^{(L_2)})\prod_{h=1}^{n_2} f(vec(\beta_h^{(L_1)})|\Sigma^{(L_2)},\beta^{(L_2)})\propto$$
$$|\Sigma^{(L_2)}|^{-(n_2+\nu+2p_0p_1+2)/2}exp(-tr((S^{L_2}+D)\Sigma^{(L_2)-1}/2)$$

with $S^{L_2}=\sum_{h=1}^{n_2}(vec(\beta_h^{L_1})-\beta^{L_2}X_h^{(L_2)\prime})'(vec(\beta_h^{L_1})-\beta^{L_2}X_h^{(L_2)\prime})$, which shows that the full conditional distribution of $\Sigma^{(L_2)}$ is $W^{-1}(S^{L_2}+D,n_2+\nu)$.

Calculation of the Full Conditional Distribution of $\beta^{(L_2)}$

We calculate the full conditional distribution of $\beta^{(L_2)}$ using 6.6 and simplifying, i.e.

$$p(vec(\beta^{(L_2)})|Y,\sigma^2,\beta_j^{(L_0)},\beta_h^{(L_1)},\Sigma^{(L_1)},\Sigma^{(L_2)})\propto f(vec(\beta^{(L_2)}))\prod_{h=1}^{n_2} f(vec(\beta_h^{(L_1)})|\Sigma^{(L_2)},\beta^{(L_2)}),$$

which, similarly to $\beta_h^{(L_1)}$, has a multivariate normal distribution with covariance matrix

$$\left(B^{-1}+\sum_h\left(X_h^{(L_2)\prime}\otimes(X_h^{(L_2)}\otimes\Sigma^{(L_2)-1})\right)\right)^{-1}$$

and mean

$$\left(B^{-1}+\sum_h\left(X_h^{(L_2)\prime}\otimes(X_h^{(L_2)}\otimes\Sigma^{(L_2)-1})\right)\right)(B^{-1}a).$$

Calculation of the Full Conditional Distribution of σ^2

We calculate the full conditional distribution of σ^2 using 6.6 and simplifying, giving

$$p(\sigma^2|Y,\beta_j^{(L_0)},\beta^{(L_2)},\beta_h^{(L_1)},\Sigma^{(L_1)},\Sigma^{(L_2)})\propto f(\sigma^2)\prod_{i=1}^{n_0} f(y_{ijh}|\beta_{jh},\beta_h^{(L_1)},\sigma^2)\propto$$
$$\sigma^{-(a+n_0/2+1)}exp(-\sigma^{-2}(b+\sum_{i=1}^{n_0}(y_i-X_i^{(L_0)}\beta_{j[i]}^{(L_0)})^2/2),$$

which shows that the full conditional distribution of σ^2 is proportional to an $IG(a+\frac{n_0}{2},b+\frac{1}{2}\sum_{i=1}^{n_0}(y_i-X_i^{(L_0)}\beta_{j[i]}^{(L_0)})^2)$ distribution.

Chapter 7

Summary

In this dissertation, Statistical Process Monitoring (SPM) and Predictive Process Monitoring (PPM) for big data sets have been discussed. We analyzed the use of classical control charting techniques as well as predictive solutions.

In Chapter 2 of this thesis, we investigated the use of the Central Limit Theorem (CLT) in monitoring large data streams. Because averages are normally distributed under certain conditions, according to the CLT, this should largely resolve the issue of non-normally distributed data. However, we showed that the tail behavior for the means of non-normally distributed subsamples deviates strongly from normality. The degree to which the distribution of the mean deviates depends on various factors: the sample size, the number of samples, the specified desired performance of the control chart, and the degree of the deviation from normality. For example, when the deviation from normality is substantial due to heavy tails (t_4) or substantial skewness (lognormal), the tail behavior can not be accurately approximated by the normal distribution even when the sample size is 1000. The implications are especially relevant for process monitoring.

Chapter 3 of this thesis is concerned with the continuous updating of parameters during process monitoring. We studied the effects of updating in various scenarios for three types of control charts. The results support updating control limits as long as the reason for out-of-control signals is known and the origin can be retraced. If this is not the case, the best strategy depends on the size of the expected mean deviation. We suggest further research on the behavior of updating the limits for various subgroup

sample sizes, as well as on performance for varying distributional assumptions.

In Chapter 4 a procedure to introduce a delay in updating control chart parameters is discussed. As discussed in Chapter 3, updating using contaminated samples should be avoided. The methods described in this chapter prevent these contaminated updates while maintaining the improvements in parameter estimation. In a case study using COVID-19 related data, we demonstrated the added value of updating control chart parameters for mortality rates in the Netherlands.

The second part of this thesis considers PPM. In Chapter 5 we considered the use of various machine learning techniques in PPM. A wide range of predictive techniques is available that are largely data-driven. We introduce a procedure to tune these predictions towards a desired false alarm rate in monitoring. Using a unique non-public data set on mental health, we investigate the performance of machine learning techniques. The Extreme Gradient Boosting (XGBoost) algorithm is subsequently used to monitor the risk of relapse in people diagnosed with schizophrenia. The procedure can aid healthcare workers in identifying people that are likely to need preventive care. Future research using more consistent data and a longer timeframe is encouraged. Neural networks can potentially improve predictions, as well as the addition of high-frequency data sources.

In Chapter 6 of this thesis, we introduced multilevel process monitoring. Modeling the hierarchical structure of a process can improve parameter estimates and the predicted probabilities. Furthermore, using a multilevel model allows monitoring at the different measurement levels. An educational case study was presented to illustrate this approach. Bayesian hierarchical modeling was used in a predictive monitoring procedure. This method produced more accurate predictions than the appropriate machine learning method. The procedure allows early warnings for students that have 'exceptional' performance. This aids schools in personalizing education and quality control. We suggest further research of the procedure using industrial process data of a hierarchical nature and varying the Bayesian priors in analyses.

In conclusion, the increase in available data and improvements in technology enable a new phase in SPM and PPM. Updating process parameter estimates will improve the use of control charts. Introducing a delay in these updates can prevent the use of contaminated data. Furthermore, early intervention based on PPM in services

and industry can support the efficient use of resources and prevent processes and people from spiraling out of control.

Bibliography

Aaronson, D. (1998). Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *The Journal of Human Resources*, *33*(4), 915–946.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (General)*, *149*(1), 1–26.

Ali, A., Khelil, A., Shaikh, F. K., & Suri, N. (2012). Efficient predictive monitoring of wireless sensor networks. *International Journal of Autonomous and Adaptive Communications Systems*, *5*(3), 233–254.

Allaire, J., & Chollet, F. (2019). *keras: R Interface to 'Keras'*. R package version 2.2.5.0. `https://CRAN.R-project.org/package=keras`.

Alwan, L. C., & Roberts, H. V. (1995). The problem of misplaced control limits. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *44*(3), 269–278.

Amarasingham, R., Patzer, R. E., Huesch, M., Nguyen, N. Q., & Xie, B. (2014). Implementing electronic health care predictive analytics: considerations and challenges. *Health Affairs*, *33*(7), 1148–1154.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders*. Arlington, VA: American Psychiatric Association, 5 ed.

Baghdadi, A., Cavuoto, L. A., Jones-Farmer, A., Rigdon, S. E., Esfahani, E. T., & Megahed, F. M. (2019). Monitoring worker fatigue using wearable devices: A case study to detect changes in gait parameters. *Journal of Quality Technology*. `https://doi.org/10.1080/00224065.2019.1640097`.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Battin-Pearson, S., Newcomb, M. D., Abbott, R. D., Hill, K. G., Catalano, R. F., & Hawkins, J. D. (2000). Predictors of early high school dropout: A test of five theories. *Journal of Educational Psychology*, *92*(3), 568–582.

Becker, J., Hall, L. S., Levinger, B., Sims, A., & Whittington, A. (2014). *Student success and college readiness: Translating predictive analytics into action*. Strategic Data Project, SDP Fellowship Capstone Report. `http://sdp.cepr.harvard.edu/files/cepr-sdp/files/sdp-fellowship-capstone-student-success-college-readiness.pdf`.

Berendsen, H. J. (2007). *Simulating the physical world: Hierarchical modeling from quantum mechanics to fluid dynamics*. Cambridge: Cambridge University Press, 1 ed.

Billingsley, P. (1995). *Probability and measure*. Wiley: New York, 32 ed.

Blyth, C. R. (1986). Convolutions of cauchy distributions. *The American Mathematical Monthly*, *93*(8), 645–647.

Bock, R. D. (1989). *Multilevel analysis of educational data*. San Diego: Academic Press, 1 ed.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514.

Buriel, R., Perez, W., de Ment, T. L., Chavez, D. V., & Moran, V. R. (1998). The relationship of language brokering to academic performance, biculturalism, and self-efficacy among Latino adolescents. *Hispanic Journal of Behavioral Sciences*, *20*(3), 283–297.

Capizzi, G., & Masarotto, G. (2020). Guaranteed in-control control chart performance with cautious parameter learning. *Journal of Quality Technology*. `https://doi.org/10.1080/00224065.2019.1640096`.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Chakraborti, S., Human, S. W., & Graham, M. A. (2009). Phase I statistical process control charts: An overview and some results. *Quality Engineering*, *21*(1), 52–62.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, *1*, 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2019). *xgboost: Extreme Gradient Boosting*. R package version 0.90.0.2. `https://CRAN.R-project.org/package=xgboost`.

Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, *24*(2), 361–370.

Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J., & Tarassenko, L. (2013). Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, *18*(3), 722–730.

Cloutier, M., Aigbogun, M. S., Guerin, A., Nitulescu, R., Ramanakumar, A. V., Kamat, S. A., DeLucia, M., Duffy, R., Legacy, S. N., Henderson, C., et al. (2016). The economic burden of schizophrenia in the United States in 2013. *The Journal of Clinical Psychiatry*, *77*(6), 764–771.

Collier, V. P. (1995). Acquiring a second language for school. *Directions in Language and Education*, *1*(4), 3–13.

Crosier, R. B. (1986). A new two-sided cumulative sum quality control scheme. *Technometrics*, *28*(3), 187–194.

Cryer, J. D., & Ryan, T. P. (1990). The estimation of sigma for an X chart: MR/d_2 or S/c_4? *Journal of Quality Technology*, *22*(3), 187–192.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21.

Diko, M. D., Chakraborti, S., & Does, R. J. M. M. (2019). An alternative design of the two-sided CUSUM chart for monitoring the mean when parameters are estimated. *Computers & Industrial Engineering*, *137*(106042), 1–14.

Emsley, R., Chiliza, B., Asmal, L., & Harvey, B. H. (2013). The nature of relapse in schizophrenia. *BMC Psychiatry*, *13*(50), 1–8.

Fenton, L. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems*, *8*(1), 57–67.

Fraker, S. E., Woodall, W. H., & Mousavi, S. (2008). Performance metrics for surveillance schemes. *Quality Engineering*, *20*(4), 451–464.

Frisén, M. (2009). Optimal sequential surveillance for finance, public health, and other areas (with discussion). *Sequential Analysis: Design Methods and Applications*, *28*(3), 310–337.

Gandy, A., & Kvaløy, J. T. (2013). Guaranteed conditional performance of control charts via bootstrap methods. *Scandinavian Journal of Statistics*, *40*(4), 647–668.

Geay, C., McNally, S., & Telhaj, S. (2013). Non-native speakers of English in the classroom: What are the effects on pupil performance? *The Economic Journal*, *123*(570), 281–307.

Geiser, S., & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. *UC Berkeley: Center for Studies in Higher Education*, *6*(7), 1–35.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432–435.

George, E., & Mudholkar, G. (1983). On the convolution of logistic random variables. *Metrika*, *30*(1), 1–13.

Goedhart, R., Da Silva, M. M., Schoonhoven, M., Epprecht, E. K., Chakraborti, S., Does, R. J. M. M., & Veiga, A. (2017a). Shewhart control charts for dispersion adjusted for parameter estimation. *IISE Transactions*, *49*(8), 838–848.

Goedhart, R., Schoonhoven, M., & Does, R. J. M. M. (2016). Correction factors for Shewhart X and $\bar{X}$ control charts to achieve desired unconditional ARL. *International Journal of Production Research*, *54*(24), 7464–7479.

Goedhart, R., Schoonhoven, M., & Does, R. J. M. M. (2017b). Guaranteed in-control performance for the Shewhart X and $\bar{X}$ control charts. *Journal of Quality Technology*, *49*(2), 155–171.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Hahn, T., Nierenberg, A. A., & Whitfield-Gabrieli, S. (2017). Predictive analytics in mental health: Applications, guidelines, challenges and perspectives. *Molecular Psychiatry*, *22*(1), 37–43.

Hall, P. (1927). The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*, *19*(3/4), 240–245.

Hany, M., & Mahmoud, M. A. (2016). An evaluation of the Crosier's CUSUM control chart with estimated parameters. *Quality and Reliability Engineering International*, *32*(5), 1825–1835.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer, 2 ed.

Hawkins, D. M. (1987). Self-starting CUSUM charts for location and scale. *The Statistician*, *36*(4), 299–316.

Henderson, C. R., Kempthorne, O., Searle, S. R., & von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, *15*(2), 192–218.

Hochreiter, S., & Schmidhuber, J. (1996). Lstm can solve hard long time lag problems. *Advances in Neural Information Processing Systems*, *9*, 473–479.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. New York: Routledge, 3 ed.

Huberts, L. C. E., Does, R. J. M. M., Ravesteijn, B., & Lokkerbol, J. (2020a). Predictive monitoring using machine learning algorithms and a real-life example on schizophrenia [submitted for publication]. *Quality and Reliability Engineering International*.

Huberts, L. C. E., Goedhart, R., & Does, R. J. M. M. (2020b). Improved control chart performance using cautious parameter learning [conditionally accepted]. *Computers & Industrial Engineering*.

Huberts, L. C. E., Schoonhoven, M., & Does, R. J. M. M. (2019). The effect of continuously updating control chart limits on control chart performance. *Quality and Reliability Engineering International*, *35*(4), 1117–1128.

Huberts, L. C. E., Schoonhoven, M., & Does, R. J. M. M. (2020c). Multilevel process monitoring: A case study to predict student success or failure. *Journal of Quality Technology*. `https://doi.org/10.1080/00224065.2020.1828008`.

Huberts, L. C. E., Schoonhoven, M., Goedhart, R., Diko, M. D., & Does, R. J. M. M. (2018). The performance of $\bar{X}$ control charts for large non-normally distributed datasets. *Quality and Reliability Engineering International*, *34*(6), 979–996.

Jensen, W. A., Jones-Farmer, L. A., Champ, C. W., & Woodall, W. H. (2006). Effects of parameter estimation on control chart properties: a literature review. *Journal of Quality Technology*, *38*(4), 349–364.

Jones, L. A., & Woodall, W. H. (1998). The performance of bootstrap control charts. *Journal of Quality Technology*, *30*(4), 362–375.

Jones, M. A., & Steiner, S. H. (2012). Assessing the effect of estimation error on risk-adjusted CUSUM chart performance. *International Journal for Quality in Health Care*, *24*(2), 176–181.

Kang, L., & Albin, S. L. (2000). On-line monitoring when the process yields a linear profile. *Journal of Quality Technology*, *32*(4), 418–426.

Kang, L., Kang, X., Deng, X., & Jin, R. (2018). A Bayesian hierarchical model for quantitative and qualitative responses. *Journal of Quality Technology*, *50*(3), 290–308.

Karande, S., & Kulkarni, M. (2005). Poor school performance. *The Indian Journal of Pediatrics*, *72*(11), 961–967.

Karve, S. J., Panish, J. M., Dirani, R. G., & Candrilli, S. D. (2012). Health care utilization and costs among medicaid-enrolled patients with schizophrenia experiencing multiple psychiatric relapses. *Health Outcomes Research in Medicine*, *3*(4), 183–194.

Keefe, M. J., Woodall, W. H., & Jones-Farmer, L. A. (2015). The conditional in-control performance of self-starting control charts. *Quality Engineering*, *27*(4), 488–499.

Kennedy, E., & Park, H.-S. (1994). Home language as a predictor of academic achievement: A comparative study of Mexican- and Asian-American youth. *Journal of Research & Development in Education*, *27*(3), 188–194.

Laidra, K., Pullmann, H., & Allik, J. (2007). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and Individual Differences*, *42*(3), 441–451.

Laurenceau, J.-P., Barrett, L. F., & Rovine, M. J. (2005). The interpersonal process model of intimacy in marriage: A daily-diary and multilevel modeling approach. *Journal of Family Psychology*, *19*(2), 314–323.

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. `arxiv.org/abs/1506.00019`.

Lucas, J. M., & Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, *32*(1), 1–12.

Luo, J. (2020). Predictive monitoring of COVID-19. SUTD Data-Driven Innovation Lab. `https://ddi.sutd.edu.sg/`.

Mandel, B. (1969). The regression control chart. *Journal of Quality Technology*, *1*(1), 1–9.

Metzger, A., Leitner, P., Ivanović, D., Schmieders, E., Franklin, R., Carro, M., Dustdar, S., & Pohl, K. (2015). Comparing and combining predictive business process monitoring techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *45*(2), 276–290.

Montgomery, D. C. (2007). *Introduction to statistical quality control*. Hoboken, NJ: John Wiley & Sons, 6 ed.

Moreno-Küstner, B., Martin, C., & Pastor, L. (2018). Prevalence of psychotic disorders and its association with methodological issues. a systematic review and meta-analyses. *PloS One*, *13*(4), 1–25.

National Institute for Public Health and the Environment (2017). Costs of diseases 2017 [data file]. `http://statline.rivm.nl/RIVM/nl/dataset/50050NED/table?ts=1597217173008`.

Nichols, J. D. (2003). Prediction indicators for students failing the state of Indiana high school graduation exam. *Preventing School Failure: Alternative Education for Children and Youth*, *47*(3), 112–120.

Nichols, J. D., & White, J. (2001). Impact of peer networks on achievement of high school algebra students. *The Journal of Educational Research*, *94*(5), 267–273.

Nie, H., & Chen, S. (2007). Lognormal sum approximation with type IV Pearson distribution. *IEEE Communications Letters*, *11*(10), 790–792.

OECD & European Observatory on Health Systems and Policies (2019). Netherlands: Country health profile 2019. `https://www.oecd-ilibrary.org/content/publication/9ac45ee0-en`.

Olfson, M., Gerhard, T., Huang, C., Crystal, S., & Stroup, T. S. (2015). Premature mortality among adults with schizophrenia in the United States. *JAMA Psychiatry*, *72*(12), 1172–1181.

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

Palmer, B. A., Pankratz, V. S., & Bostwick, J. M. (2005). The lifetime risk of suicide in schizophrenia: A reexamination. *Archives of General Psychiatry*, *62*(3), 247–253.

Parker, J. D., Hogan, M. J., Eastabrook, J. M., Oke, A., & Wood, L. M. (2006). Emotional intelligence and student retention: Predicting the successful transition from high school to university. *Personality and Individual Differences*, *41*(7), 1329–1336.

Paxton, C., Niculescu-Mizil, A., & Saria, S. (2013). Developing predictive models using electronic medical records: Challenges and pitfalls. In *AMIA Annual Symposium Proceedings*, vol. 2013, (pp. 1109–1115). American Medical Informatics Association.

Plummer, M. (2018). *rjags: Bayesian graphical models using MCMC*. R package version 4-8. `ttps://CRAN.R-project.org/package=rjags`.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Psarakis, S., Vyniou, A. K., & Castagliola, P. (2014). Some recent developments on the effects of parameter estimation on control charts. *Quality and Reliability Engineering International*, *30*(8), 1113–1129.

Qiu, P., Zou, C., & Wang, Z. (2010). Nonparametric profile monitoring by mixed effects modeling. *Technometrics*, *52*(3), 265–277.

Quesenberry, C. P. (1991). SPC Q charts for start-up processes and short or long runs. *Journal of Quality Technology*, *23*(3), 213–224.

Rahafar, A., Maghsudloo, M., Farhangnia, S., Vollmer, C., & Randler, C. (2016). The role of chronotype, gender, test anxiety, and conscientiousness in academic achievement of high school students. *Chronobiology International*, *33*(1), 1–9.

Reifman, J., Rajaraman, S., Gribok, A., & Ward, W. K. (2007). Predictive monitoring for improved management of glucose levels. *Journal of Diabetes Science and Technology*, *1*(4), 478–486.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, *1*(3), 239–250.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, *35*(1), 83–92.

Romero, C., & Ventura, S. (2019). Guest editorial: Special issue on early prediction and supporting of learning performance. *IEEE Transactions on Learning Technologies*, *12*(2), 145–147.

Rothman, S. (2001). School absence and student background factors: A multilevel analysis. *International Education Journal*, *2*(1), 59–68.

Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics*, *121*(1-2), 297–317.

Ruetsch, C., Un, H., & Waters, H. C. (2018). Claims-based proxies of patient instability among commercially insured adults with schizophrenia. *ClinicoEconomics and Outcomes Research: CEOR*, *2018*(10), 259–267.

Saleh, N. A., Mahmoud, M. A., & Abdel-Salem, G. A. S. (2013). The performance of the adaptive exponentially weighted moving Average control chart with estimated parameters. *Quality and Reliability Engineering International*, *29*(4), 595–606.

Saleh, N. A., Mahmoud, M. A., Keefe, M. J., & Woodall, W. H. (2015). The difficulty in designing Shewhart $\bar{X}$ and X control charts with estimated parameters. *Journal of Quality Technology*, *47*(2), 127–138.

Saleh, N. A., Zwetsloot, I. M., Mahmoud, M. A., & Woodall, W. H. (2016). CUSUM charts with controlled conditional performance under estimated parameters. *Quality Engineering*, *28*(4), 402–415.

Sang, H., & Gelfand, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environmental and ecological statistics*, *16*(3), 407–426.

Schirru, A., Pampuri, S., & De Nicolao, G. (2010). Multilevel statistical process control of asynchronous multi-stream processes in semiconductor manufacturing. In *2010 IEEE International Conference on Automation Science and Engineering*, (pp. 57–62). IEEE.

Sellström, E., & Bremberg, S. (2006). Is there a "school effect" on pupil outcomes? a review of multilevel studies. *Journal of Epidemiology & Community Health*, *60*(2), 149–155.

Shewhart, W. A. (1926). Quality control charts. *Bell System Technical Journal*, *5*(4), 593–603.

Shu, L., Tsung, F., & Tsui, K.-L. (2004). Run-length performance of regression control charts with estimated parameters. *Journal of Quality Technology*, *36*(3), 280–292.

Spiewak, S., Duggirala, R., & Barnett, K. (2000). Predictive monitoring and control of the cold extrusion process. *CIRP Annals*, *49*(1), 383–386.

Statistics Netherlands (2017). Gezondheid, leefstijl, zorggebruik en -aanbod, doodsoorzaken; kerncijfers [data file]. `https://opendata.cbs.nl/CBS/nl/dataset/81628NED/table?ts=1597217730991`.

Stewart, E. B. (2008). School structural characteristics, student effort, peer associations, and parental involvement: The influence of school-and individual-level factors on academic achievement. *Education and Urban Society*, *40*(2), 179–204.

Substance Abuse and Mental Health Services Administration (2018). Key substance use and mental health indicators in the united states: Results from the 2017 national survey on drug use and health. `https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHFFR2017/NSDUHFFR2017.pdf`.

Sui-Chu, E. H., & Willms, J. D. (1996). Effects of parental involvement on eighth-grade achievement. *Sociology of Education*, *69*(2), 126–141.

Sullivan, S., Northstone, K., Gadd, C., Walker, J., Margelyte, R., Richards, A., & Whiting, P. (2017). Models to predict relapse in psychosis: A systematic review. *PloS one*, *12*(9), 1–12.

Tax, N., Verenich, I., La Rosa, M., & Dumas, M. (2017). Predictive business process monitoring with lstm neural networks. In *International Conference on Advanced Information Systems Engineering*, (pp. 477–492). Springer.

Teinemaa, I., Dumas, M., Rosa, M. L., & Maggi, F. M. (2019). Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *13*(2), 1–57.

Vera do Carmo, C., Lopes, L. F. D., & Souza, A. M. (2004). Comparative study of the performance of the CuSum and EWMA control charts. *Computers & Industrial Engineering*, *46*(4), 707–724.

Vigod, S. N., Kurdyak, P. A., Seitz, D., Herrmann, N., Fung, K., Lin, E., Perlman, C., Taylor, V. H., Rochon, P. A., & Gruneir, A. (2015). Readmit: a clinical risk index to predict 30-day readmission after discharge from acute psychiatric units. *Journal of Psychiatric Research*, *61*, 205–213.

Vining, G. (2009). Technical advice: Phase I and phase II control charts. *Quality Engineering*, *21*(4), 478–479.

Vos, T., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulkader, R. S., Abdulle, A. M., Abebo, T. A., Abera, S. F., et al. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, *390*(10100), 1211–1259.

Wang, Y. F., Tseng, S. T., Lindqvist, B. H., & Tsui, K. L. (2019). End of performance prediction of lithium-ion batteries. *Journal of Quality Technology*, *51*(2), 198–213.

Weese, M., Martinez, W., Megahed, F. M., & Jones-Farmer, L. A. (2016). Statistical learning methods applied to process monitoring: An overview and perspective. *Journal of Quality Technology*, *48*(1), 4–24.

Woodall, W. H. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, *38*(2), 89–104.

Woodall, W. H., & Montgomery, D. C. (2014). Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology*, *46*(1), 78–94.

World Health Organization (2019). Schizophrenia fact sheet. `https://www.who.int/news-room/fact-sheets/detail/schizophrenia`.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and xgboost. *IEEE Access*, *6*(1), 21020–21031.

Zhou, J., Li, X., Andernroomer, A. J., Zeng, H., Goh, K. M., Wong, Y., & Hong, G. S. (2005). Intelligent prediction monitoring system for predictive maintenance in manufacturing. In *31st Annual Conference of IEEE Industrial Electronics Society, 2005. IECON 2005.*, (pp. 2314–2319). IEEE.

Zwetsloot, I. M., & Ajadi, J. O. (2019). A comparison of EWMA control charts for dispersion based on estimated parameters. *Computers & Industrial Engineering*, *127*(1), 436–450.

Acknowledgements

This thesis is a culmination of more than three years of research in collaboration with and supported by a large number of people. First, I want to thank my supervisor Prof. Dr. Ronald Does and co-supervisor Dr. Marit Schoonhoven, who recruited me to the Amsterdam Business School in 2016 and have been the core support team in my research and teaching. Marit, I would like to thank you for your enthusiasm, your thorough and valuable feedback, and your guidance in teaching. I have a lot of respect for the way you combine the demands of motherhood with your great work. Ronald, thank you for your clear guidance and decisiveness, your detailed input, and roaring laugh. Based on a short speech I gave on a completely unrelated topic in 2016 you somehow decided I would be a good addition to the group. Hopefully, I did not disappoint. Over the years we have made the leap to predictive monitoring, traveled the world, had a lot of 'plopmoments' and many laughs.

Secondly, many thanks to the committee for taking the time to read this thesis and provide constructive feedback. I hope we can (continue to) work together in the future.

Thanks to my co-writers for your input in this thesis. Besides my supervisors, Dr. Rob Goedhart, Dr. Mandla D. Diko, Dr. Joran Lokkerbol, and Dr. Bastian Ravesteijn all had a vital role in the articles contained in this manuscript. In particular, Rob and Joran, thank you for your energy and initiative.

Thank you to my brilliant colleagues/friends that got me through the rough patches. Rob for all the fun we had together all around the world; Yannik for being my buddy and your terrible taste in music; Alex for all the walks, talks, and support; Stevan for going crazy together from Amsterdam to Hamburg; Robert for all your inspo and insane humor; Ujjwal for being the sweetest and smartest of the bro's; Bart

for always being open for a chat and showing what being efficient really means; Inez for the support in work and fun; Thomas for putting things in perspective; Reza for all of your help and smiles; Chintan for accepting me as a roommate; Atie for making sure things get done and all the ABS PhD's for the fun weekends and activities.

I would also like to thank my BSc and MSc thesis supervisor Dr. Maurice J.G. Bun. Together we published my BSc thesis "The impact of higher fixed pay and lower bonuses on productivity" in the Journal of Labor Research, and my MSc thesis was the basis for the final chapter of this thesis. I'm glad we're still in touch – thank you for your positivity and guidance.

Thanks to all the great people in and around the UvA: Sandy, Sam, Sue-Ellen, and all the others for the positive vibes.

Special thanks to my parents Leo and Carla and brother Krystan for their unlimited love and support, and for their votes on Christmas' eve 2016. Pap and mam, you have always convinced me I could do more than I actually can. You have been my biggest fans for everything from my DJ-career, through my five bachelor's studies to Delph and back to UvA. Thank you for all the Nam Kees, Verandas, weekends in Rome and Venice, etc., etc. I am grateful to Geesje and my two little niffos Alexander and Thomas, for always putting a smile on my face.

I want to thank Ruben Polak and Ruben Spruit and Lyor Kooistra, for living with and supporting me throughout the writing of this thesis. Thanks to Youri Vink for always having my back. Thank you, Margriet Bosman, for being a friend and mentor for over 18 years now. Thank you to my amazing 'Bolle is Koning' friends, you've been there through it all. Thanks to my football team for all the fun times. Thanks, Mark Verhagen for asking the right questions. Thanks to Gijs Overgoor for being my main ally across the pond.

To my beautiful girlfriend Renate, thank you for being there during the ups and downs and your direct contributions to this thesis. I'm so happy you're crazy enough to be with me 24/7, through lockdowns and 'avondklokken'. You inspire me.

I feel very lucky to have so many amazing people in my life.

About the Author



Leo C.E. Huberts (1991) holds MSc and BSc degrees in Econometrics and a BSc in Natural and Social Sciences from the University of Amsterdam (UvA). During his studies, he worked at the financial department of the Stichting CPNB, was a strategy consultant for the Kleine Consultant Amsterdam (doing pro bono strategy projects for small and medium-sized companies), served on various committees, and was board member of the Analytics Academy (running pro bono data science projects for cultural and social institutions).

During his master, he was a manager for the Big Data Alliance and founded Delph, a data science consultancy and development firm with clients ranging from political parties to municipalities and construction companies. After two years with Delph, he decided to pursue a Ph.D. degree at the University of Amsterdam, of which this thesis is the result.

Leo currently works as a researcher and teacher at the Amsterdam Business School, continuing his work on applied statistics and machine learning. He aims to do meaningful work in bridging the gap between theory and practice. For more information on current activities and publications visit `linkedin.com/in/leohuberts/` or `researchgate.net/profile/Leo_Huberts2`.

In this thesis, we investigate the possibilities of the increase in the size and frequency of data for both statistical and predictive process monitoring. This includes adjusting statistical process monitoring techniques based on large samples using the Central Limit Theorem and updating parameter estimates to increase the flexibility for high-frequency data. Furthermore, combining the increase in data with advances in modeling techniques paves the way for predictive monitoring. Signaling as early as possible can be imperative in taking preventive measures in sectors such as healthcare, education, manufacturing, maintenance, and more. It can be vital to ensure the quality of products and services.

LEO C.E. HUBERTS



UNIVERSITY OF AMSTERDAM

MONITORING COMPLEX PROCESSES IN THE AGE OF BIG DATA