PHASE I ANALYSIS OF VARIABLES TYPE CORRELATED DATA

NİTELİK TİPİ KORELASYONLU VERİ İÇİN FAZ I ANALİZİ

ERDİ DAŞDEMİR

PROF. DR. MURAT CANER TESTİK
Supervisor

Submitted To
The Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfilment of the Requirements
for the Award of the Degree of Master of Science
in Industrial Engineering

2015

This work named **"Phase I Analysis of Variables Type Correlated Data"** by **ERDİ DAŞDEMİR** has been approved as a thesis for the Degree of **MASTER of SCIENCE in INDUSTRIAL ENGINEERING** by the below mentioned Examining Committee Members.

Prof. Dr. İmdat Kara
Head

Prof. Dr. Murat Caner Testik
Supervisor

Yrd. Doç. Dr. Güldal Güleryüz
Member

Yrd. Doç. Dr. Ceren Tuncer Şakar
Member

Yrd. Doç. Dr. Barbaros Yet
Member

This thesis has been approved as a thesis for the Degree of **MASTER of SCIENCE in INDUSTRIAL ENGINEERING** by Board of Directors of the Graduate School of Science and Engineering.

Prof. Dr. Fatma SEVİN DÜZ
Director of the Graduate School of
Science and Engineering

This thesis is dedicated to
my wonderful parents, İsmet and Ayşe,
my sweet sister, Sevgi,
and my loving fiancée, Sevinç
for all their support.

ETHICS

In this thesis study, prepared in accordance with the spelling rules of the Graduate School of Science and Engineering of Hacettepe University,

I declare that

- all the information and documents have been obtained in the base of the academic rules
- all audio-visual and written information and results have been presented according to the rules of scientific ethics
- in case of using others Works, related studies have been cited in accordance with the scientific standards
- all cited studies have been fully referenced
- I did not do any distortion in the data set
- and any part of this thesis has not been presented as another thesis study at this or any other university.

16/06/2015

ERDİ DAŞDEMİR

ABSTRACT

PHASE I ANALYSIS OF VARIABLES TYPE CORRELATED DATA

Erdi Daşdemir

Master of Science, Department of Industrial Engineering

Supervisor: Prof. Dr. Murat Caner Testik

June 2015, 67 Pages

Phase I analysis of a control chart implementation comprises parameter estimation, chart design, and outlier filtering, which are performed iteratively until reliable control limits are obtained. These control limits are then used in Phase II for online monitoring and prospective analyses of the process to detect out-of-control states. Although a Phase I study is required only when the true values of the parameters of a process are unknown, this is the case in many practical applications. In the literature, however, process parameters had often been assumed as known and parameter estimation step of Phase I analysis was often skipped to simplify the development of control charts. Recently, effects of parameter estimation on the performance of process monitoring in Phase II were recognized as an important research field. However, these studies consider availability of complete and clean data sets, without outliers and missing observations, and neglect the iterative nature of Phase I analysis, in which outliers are identified and control charts are designed through revisions. In the traditional use of control charts, it is also assumed that process observations are independent and identically distributed. The real industrial processes, however, often have correlated observations. It is well known that autocorrelation effects parameter estimation and so control chart design. Many charting methods have been proposed for autocorrelated data.

In this thesis, AutoRegressive models of order 1, AR(1), are considered and the effects of two extreme cases for Phase I analysis are studied; the case where all outliers are filtered from the data set (parameter estimation from incomplete but clean data) and the case where all outliers remain in the data set during estimation. To investigate these effects, autocorrelated observations from AR(1) process are generated for different observation lengths and autocorrelation parameters. For the case of all outliers remaining in the dataset, some of the generated observations are randomly selected (for different rates) and contaminated to simulate outliers. Then, process parameters are estimated without filtering the outliers. For the missing observations case, randomly selected observations from the dataset are assumed as outliers and filtered from the data set before estimations. For estimating the parameters under both cases in Phase I, performance of the maximum likelihood and conditional sum of squares estimators are evaluated. As an approach to evaluate control charts, firstly control chart is designed with the estimated parameters and then average run length performance for Phase II is computed by using the obtained design. Results indicate that presence of outliers may have severe effects on the estimates of control chart parameters. Moreover, the effect of not detecting outliers in Phase I can be also severe on the Phase II application of a control chart. A real world example is provided to illustrate the proposed method and importance of an appropriate Phase I analysis.

Key Words: Statistical process control, Phase I analysis, Phase II analysis, control charts, autocorrelation, outliers, maximum likelihood estimator, conditional sum of squares estimator.

ÖZET

NİTELİK TİPİ KORELASYONLU VERİ İÇİN FAZ I ANALİZİ

Erdi Daşdemir

Yüksek Lisans, Endüstri Mühendisliği Bölümü

Tez Danışmanı: Prof. Dr. Murat Caner Testik

Haziran 2015, 67 Sayfa

İstatistiksel süreç kontrolünde, kontrol grafikleri Faz I ve Faz II olmak üzere iki aşamalı olarak uygulanmaktadır. Faz I aşamasında, süreçle ilgili geriye dönük analizler yapılarak sistemin kontrol altında olduğu duruma ait parametreler tahmin edilir. Kontrol grafiklerinde Faz I uygulaması, parametre tahmini, kontrol grafiğinin tasarımı ve kontrol dışı noktaların (aykırı noktaların) filtrelenmesi işlemlerini içerir. Bu işlemler, güvenilir kontrol limitleri elde edilene kadar yinelenir. Elde edilen bu güvenilir limitler, Faz II aşamasında süreci gerçek zamanlı izlemek için kullanılır. Bu aşamada sürecin kontrol dışına çıktığı durumlar, bir başka deyişle sinyaller tespit edilir.

İstatistiksel süreç kontrolü literatüründe genellikle sürece ait parametrelerinin bilindiği varsayılmakta, bu nedenle de Faz I aşamasının parametre tahmin işlemi göz ardı edilerek doğrudan Faz II uygulamaları üzerinde durulmaktadır. Oysaki birçok gerçek endüstriyel uygulamada süreç parametreleri bilinmemekte ve bu nedenle de Faz I analizinin yapılması gerekmektedir. Bununla birlikte kontrol grafiklerinin, Faz II aşamasındaki çevrimiçi gözlem performansları, Faz I analizi sonuçlarından etkilenmektedir. Bu da Faz I uygulamasının önemini artırmaktadır. Son yıllarda parametre tahminlerinin Faz II aşamasındaki gözlem performansı üzerindeki etkileri önemli bir araştırma alanı olarak ortaya çıkmıştır. Ancak yapılan çalışmaların birçoğu temiz ve eksiksiz veri setine sahip olunduğu varsayılarak

yapılmakta, aykırı veya eksik gözlemler dikkate alınmamakta ve de aykırı gözlemlerin tespit edilip kontrol grafiklerinin tasarımlarının revize edildiği Faz I'in yinelemeli yapısı hesaba katılmamaktadır.

Literatürdeki ikinci bir geleneksel yaklaşım da, kontrol altındaki süreç gözlemlerinin normal dağılıma sahip, bağımsız ve aynı dağılımlı olduğu varsayımıdır. Süreç gözlemleri bağımsız olduğunda, Faz I uygulaması esnasında kontrol dışına çıkan gözlemler süreçten sorunsuz bir şekilde çıkarılabilmektedir. Ancak, gerçek endüstriyel süreçlerin birçoğunda gözlemler bir zaman serisi olarak önceki gözlemlere bağımlıdır. Birbirini takip eden gözlemler arasında otokorelasyon olduğu zaman, Faz I uygulaması esnasında kontrol dışına çıkan gözlemlerin süreç dışında bırakılması, süreç parametrelerinin tahminlerini ve buna bağlı olarak kontrol grafiğinin performansını etkilemektedir. Bu konuyla ilgili çalışmalar yapılmış, otokorelasyonlu veriler için birçok yeni kontrol grafiği yöntemi önerilmiştir. İstatistiksel süreç kontrolü literatüründe, otokorelasyonlu gözlemlere sahip süreçler için Faz I uygulaması araştırmaya oldukça açık bir konudur.

Bu tezde, bağımlı gözlemlerin Faz I analizi temel alınmaktadır. Otokorelasyonlu gözlemler için Faz I analizinde nasıl bir yaklaşım izlenmesi gerektiği ve bu yaklaşımın Faz II aşamasındaki kontrol grafiklerinin performansına olan etkileri incelenmiştir. Bu doğrultuda yapılan simülasyon çalışmasında, özbağlanımlı AR(1) süreci ele alınmış ve Faz I analizinde iki ekstrem durumun etkileri incelenmiştir. Bu durumlar, parametre tahmini öncesinde tüm aykırı gözlemlerin süreç dışında bırakıldığı durum ve tüm aykırı gözlemlerin süreç içerisinde bırakıldığı durumdur. Etkilerin araştırılması için, farklı otokorelasyon parametresine sahip AR(1) süreçlerinden farklı uzunluklarda veri setleri üretilmiştir. Aykırı gözlemler veri setinde tutularak parametrelerin tahmin edildiği durumda, üretilen veri setlerinden farklı oranlarda rasgele gözlemler seçilmiş, bu gözlemler aykırı gözlemlere dönüştürülmüştür. Bu şekilde parametre tahminleri yapılmıştır. Aykırı gözlemlerin filtrelenerek veri seti dışında bırakıldığı durumda ise, üretilen veri setinden belirli oranlarda rasgele gözlemler seçilmiş, bu gözlemler aykırı gözlem olarak varsayılmış ve parametre tahmini öncesinde veri seti dışında bırakılmıştır.

Bu amaçla gerçekleştirilen simülasyon çalışmasında; 0.3, 0.5 ve 0.7 olmak üzere üç farklı otokorelasyon parametresine sahip veri setleri üretilmiştir. Bu üç farklı değerin her biri için 50, 100 ve 200 olmak üzere üç farklı uzunlukta veri setleri üretilmiştir. Yukarıda bahsedilen iki durum (aykırı gözlemlerin süreç içerisinde bırakıldığı ve filtrelendiği durumlar) için veri setlerinin uzunluklarının 0, 0.02, 0.05, 0.1 ve 0.25 oranlarında aykırı gözlemler yaratılmıştır. Faz I aşamasındaki parametre tahminleri aşamasında, her iki durum için de En Büyük Olabilirlik ve Koşullu Kareler Toplamı tahmin edicilerinin performansları incelenmiştir. Yaklaşım olarak, öncelikle tahmin edilen parametrelerle kontrol grafikleri tasarlanmıştır. Faz I sonunda elde edilen bu tasarımlar kullanılarak Faz II aşamasında kontrol grafiklerinin ortalama tespit uzunluğu (ARL) performansları hesaplanmıştır. Çalışmanın sonunda, gürbüz tahmin konusu da kısaca değerlendirilmiştir.

Bu araştırmanın sonunda elde edilen sonuçlar, otokorelasyonlu bir veride %2 oranındaki aykırı gözlemin bulunmasının bile Faz I aşamasındaki parametre tahminlerini oldukça çok etkilediğini ve güçlü bir negatif etki oluştuğunu ortaya koymaktadır. Bu olumsuz etki, En Büyük Olabilirlik ve Koşullu Kareler Toplamı tahmin edicilerinin her ikisinde de ortaya çıkmaktadır. Ayrıca. Faz II aşamasındaki ortalama tespit uzunluğu performansları incelendiğinde de, Faz I aşamasındaki aykırı gözlemlerin parametre tahmini öncesinde filtrelenmesinin gerekliliği ve önemi anlaşılmıştır. Aykırı gözlem oranının yükselmesiyle birlikte, ARL_0 değerlerinin arttığı açık bir şekilde görülmüştür. Aykırı gözlemlerin dışarıda bırakılması sonucunda elde edilen temiz ancak eksik veri setinin tahminlerde kullanılmasının, anlamlı bir şekilde tahmin performanslarını etkilemediği görülmüştür. Ancak, aykırı gözlem oranı yükseldiğinde, bir başka deyişle çok fazla sayıda gözlemin filtrelenmesi gerektiğinde, En Büyük Olabilirlik tahmin edicisinin kullanılması kullanıcıya önerilmektedir.

Anahtar Kelimeler: İstatistiksel süreç kontrolü, Faz I analizi, Faz II analizi, kontrol grafikleri, otokorelasyon, aykırı gözlemler, en büyük olabilirlik tahmin edicisi, koşullu kareler toplamı tahmin edicisi

ACKNOWLEDGMENTS

Master thesis is a somewhat beginning step for one who plan to build a successful academic career. I am deeply grateful to my thesis supervisor Prof. Dr. Murat Caner Testik for his efforts, guidance and contributions throughout my first professional academic study. He has not only guided me perfectly throughout the study, but also helped me to begin a bright career as an ideal figure of an academician.

During my thesis, I worked with two more great professors, Prof. Dr. Christian Weiss (HSU Hamburg, Germany) and Prof. Dr. Sven Knoth (HSU Hamburg, Germany). They provided valuable contributions and criticisms, which helped me to improve my research findings. Their guidance not only helped me in all steps of my thesis, but also introduced me to the world of solid academic research. I am also grateful to Prof. Dr. Roland Fried (TU Dortmund, Germany) for his very helpful advice concerning the robust estimation of the AR(1) parameters.

I would like to express my gratitude to my family members: my father İsmet Daşdemir, my mother Ayşe Daşdemir, and my sister Sevgi Daşdemir. They supported me in every step I have taken since the very beginning of my academic life. Without their love, such a work would not exist.

I would like to express my deepest appreciation to my fiancée Sevinç Çayıroğlu for her continuous moral and support I have received throughout my studies. Her encouragement, which shed light on my way during tough challenges, was the most valuable of all.

Finally, I would like to thank my dear friends Macit Mete Oğuz, Murat Atalay, Volkan Bilgin, Muhammed Alan and Rıfat Can Aktaş for their invaluable friendships, support, and their encouragements to me for beginning an academic career. I always miss our PES nights.

CONTENTS

TABLES

FIGURES

SYMBOLS AND ABBREVIATIONS

Symbols

$\bar{x}$	Sample average
n	Sample size
μ	Mean
σ	Standard deviation
α	Type I error
σ^2	Variance
β	Type II error
ϕ	The autoregressive parameter
ε_t	The white noise process

Abbreviations

AR (1)	AutoRegressive process model of order 1
ARL	Average Run Length
ARL_0	In-control ARL
ARL_1	Out-of control ARL
CL	Center Line
CSS	Conditional Sum of Squares Estimator
CUSUM	Cumulative Sum
EWMA	Exponentially Weighted Moving Average
FAP	False alarm probability
FAR	False alarm rate
i.i.d.	Independent and identically distributed
L	Distance of the limits from the center line, in terms of standard deviation units
LCL	Lower Control Limit
ML	Maximum Likelihood Estimator
OCAP	Out-of-control-action plan
SPC	Statistical Process Control
SQC	Statistical Quality Control
T	Data lengths
UCL	Upper Control Limit

1. INTRODUCTION

The fundamentals of statistical quality control (SQC) were proposed in the 1920s. Although there had been many important advances until the mid to late 1970s, it was the quality revolution in 1980s that made the field really attractive for the researchers. The quality revolution was the result of increased competition in global market. Since high quality products were required to survive in global competition, research activities in the field of quality were greatly increased. Quality control has gained a relative importance and numerous methods and tools that supports continuous improvement have been proposed by the researchers [1].

Statistical Process Control (SPC) as a field of SQC, is a collection of tools to improve processes by reducing variability and ensuring stability. A process may be operating in an in-control state, where only common causes of variation exist. Alternatively, a process may be operating in an out-of-control state, in which special causes of variation exist. Although common causes of variation may be considered to be acceptable due to technological or economic reasons, out-of-control process states should be rapidly detected for taking corrective actions to improve processes. As a powerful SPC tool, control charts have been widely used in the industry to monitor transitions between these two process states and detect unusual variations occurring due to the assignable causes. Through the SPC history, numerous statistical methods have been incorporated into SPC procedures. Today, SPC is a field that has a very rich and comprehensive research literature.

1.1. Problem Definition

Monitoring a process using control charts first requires defining the in-control state of the process, and then this information is used to detect out-of-control process states. Hence, there are two different phases of a control chart application. Phase I is the retrospective analysis of process observations. Parameter estimation, chart design, and outlier filtering are performed iteratively during Phase I, until reliable control limits for actual process monitoring are obtained [2]. Control chart design obtained in Phase I is then used in Phase II for online monitoring and prospective analyses of the process to detect out-of-control states [3]. In Phase II, a point exceeding the control limits is an alarm signal, which is an indication of a potential out-of-control process state. Hence, a widely used metric for evaluating control chart performance in Phase II is the average run length (ARL), which is the average number of points plotted on a control chart until a signal is triggered. A classification of the ARL metric is used to differentiate the performance; for the false alarms (ARL_0) when the process is actually in-control, and for the detection performance when the process shifts out-of-control (ARL_1).

Studies in the literature mainly focus on theoretical/empirical sampling distributions of parameters without considering the iterative nature of a Phase I analysis, in which outliers are identified and control charts are designed through revisions. Process parameters have been commonly assumed as known, which is not the case in practice. Often, unknown process parameters need to be estimated in practice and it is known that estimation during Phase I effects control chart design. Furthermore, process observations are usually assumed as independent and identically distributed, which may not be the case in real processes.

In this thesis, effects of Phase I analysis on the parameter estimates as well as performance of control charts in Phase II implementations are investigated for autocorrelated time series data. Note that, this research does not simply focus on investigating the effects of autocorrelation in parameter estimation, which has been already studied in the literature. This research also investigates the effects of outliers in autocorrelated data on the parameter estimates in Phase I and on the control chart performance in Phase II implementations, i.e. use of the control chart for online process monitoring. The findings of this research may be a solution to the problem of deciding for whether filtering or keeping outliers in autocorrelated data during Phase I implementation.

1.2. Motivation

In the literature, the parameter estimation step of a Phase I analysis is often neglected and in-control process parameters are assumed to be known for simplifying development and evaluation of control charts [2] [4]. In practice, however, process parameters are often unknown and control chart design first requires parameter estimation, which is then used to determine control limits of the chart [5]. It is well known that parameter estimation in Phase I analysis has significant effects on the actual performance of process monitoring in Phase II. More specifically, the use of estimated parameters might cause a dramatic increase in the false alarm rates [6]. A comprehensive literature review by Jensen et al. [7] concluded that the effects of parameter estimation have to be considered while investigating control chart performance. Moreover, Montgomery and Woodall [1], and Woodall and Montgomery [8] pointed out the effect of parameter estimation on control chart performance as an active research area.

Although SPC researchers often consider independent observations, monitoring of autocorrelated observations is also common in industry. For autocorrelated observations, applying standard control charts to the residuals of an appropriate time series model or modifying the control limits of known charts for the correlation are the two mainly proposed approaches in SPC [9] [10]. Yet, effects of Phase I analysis are often neglected in evaluating performance of control charts for autocorrelated observations.

1.3. Contribution of the Thesis

As a main difference from the literature, here outliers for autocorrelated data in Phase I are considered in a more realistic scenario for practical applications. Effects of outliers are investigated under two extreme cases: all outliers are filtered from data and all outliers remain in data. Note that filtering of outliers would result in incomplete time series data. Time series observations from an **A**uto**R**egressive process model of order 1, i.e. an AR(1) process, are simulated by using different data lengths and autocorrelation parameters. For the first case, some of the autocorrelated data are randomly assumed as outliers and filtered from the dataset before parameter estimations. For the second case, outliers are randomly generated in the autocorrelated dataset. To generate outliers, first some of data are selected with different rates. Then, these selected data are contaminated according to a certain rule and parameters are estimated without filtering these contaminated data. The behavior of the maximum likelihood (ML) and conditional sum of squares (CSS) estimators under the two cases are described and compared. The obtained designs at the end of Phase I are then used to investigate effects of outliers in autocorrelated observations on average run length performance of the control chart in Phase II. Furthermore, the topic of robust estimation is briefly considered.

The results of this research show that even 2% outliers in the autocorrelated data set heavily affects the parameter estimates and a strong negative bias can be observed in Phase I both with ML and CSS estimators. It is also shown that filtering outliers in Phase I is important to obtain a better Phase II process monitoring performance. Additionally, it can be suggested to use ML estimator if many data need to be filtered due to the outliers.

1.4. Thesis's Organization

The rest of this thesis is organized as follows. In Chapter 2, the related literature is discussed, Phase I and Phase II implementations of control charts are explained. Chapter 3 presents the methodology employed in this research. Details regarding the investigation of the performances of estimators through simulation experiments are described. The method to investigate the control charts' performance, using the designs obtained from a Phase I analysis, is also provided in Chapter 3. The results of the thesis's research and their interpretations are provided in Chapter 4. The proposed solutions for parameter estimation and the chart design are illustrated by a real-world example in Section 5. Finally, conclusions and remarks concerning future research, including a discussion of robust parameter estimation, are given in the Section 6.

2. LITERATURE REVIEW

SQC is a field where industrial statistics and quality improvement philosophies are integrated. More specifically, SQC comprises the areas of acceptance sampling, statistical process control, design of experiments, and capability analysis [3]. Since this thesis research is mainly related with the field of SPC, the following literature review will be focusing on the details of SPC and control charting.

2.1. Statistical Process Control (SPC)

In the beginning, quality researchers were focusing mostly on keeping the mean of a quality control characteristic in between certain specification limits. Later on, researchers like Deming, an important figure in quality history, emphasized that meeting specification limits is not enough by itself to claim a good quality; variability also need to be considered. Another quality guru Taguchi supported the aim of reducing the variability, as long as remaining in the feasible region in terms of cost [11]. Montgomery [3] also stated that in order to meet customer's quality expectations, a product should be produced by a stable process in terms of deviation from the target quality characteristic. Hence, today quality improvement is considered as reduction in variability for critical quality characteristics.

SPC is often employed to understand, analyze, and continuously improve the processes over time [11]. Since SPC comprises useful tools for achieving process stability, it has a key role in understanding and reducing the variations existing in the processes. The tools employed in SPC are also named as "***magnificent seven***" in the literature, as shown in Figure 1.

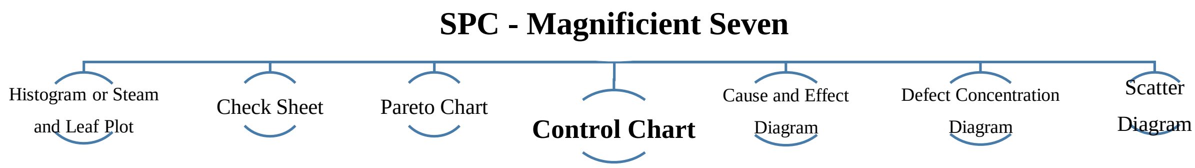


Figure 1. The Seven Tools (Magnificent Seven) used in SPC

Among these tools, control chart is the most complex and technical tool that is capable to reduce variability and improve process quality. Since control charting is the major SPC tool used in this thesis, the following sections will focus on it. Any reader also interested with other six tools may see Montgomery [3] for their details.

2.2. Control Charting

Walter A. Shewhart, who is considered as the father of SPC, was the first one who proposed the fundamentals of control charting in 1920s and 1930s [12]. Since first control chart in the history was introduced by Shewhart, a basic control chart is also referred to as Shewhart Control Chart in the literature.

Control charting is primarily used to reduce variability through monitoring of the processes. In order to understand principles of control charts, one must first understand the role of variability in the process control. In terms of variability, a process may be operating in two states, whether statistically ***in-control or out-of-control***. There are mainly two distinct type of causes for the variation occurring in the processes [1]:

- **Common Causes of Variation (Chance Causes of Variation):** This type of causes are considered as they inherently exist in the processes. In SPC terminology, if a process operates only with chance causes, it is assumed as statistically in-control.

- **Assignable Cause of Variation (Special Causes of Variation):** This type of causes are considered as the unusual variation indicators. The process is assumed as statistically out-of-control, if any assignable cause exists.

Although there are various definitions for in-control and out-of-control processes in the literature, it is easy to realize that they all describe the same concept by only using different words. For example, Woodall [11] stated that a process is in statistical control if the probability distribution of the quality characteristic is stable over time, and out of control if the distribution changes over time.

A control chart detects assignable causes of variation and then signals it to the users. When a chart signals, corrective actions must be taken to reduce variability. These actions primarily include analyzing the process, identifying the assignable causes of variation, and removing them from the process.

2.2.1. Principles of Shewhart Type Control Charts

In a basic Shewhart type Control Chart, there are three lines drawn on the two dimensional coordinate system (X-Y coordinates). Center line (CL) represents the mean of the monitored control statistic. The Upper Control Limit (UCL) and Lower Control Limit (LCL) lines represent the boundaries of the allowable in-control interval for the control statistic. A typical control chart is illustrated in Figure 2. These limits are first determined through statistical analysis of the relevant process, and then used for monitoring the actual process. In process monitoring, a control statistic obtained from the observations of a monitored quality characteristic are plotted on this coordinate system, over time. A control chart gives an alarm, if the value of a point goes outside these limits. Please remember that even if all the observations fall between the limits, if they show any systematic or nonrandom behavior this may also indicate that the process is out-of-control.

As an example, assume that there is a metal bar production process and the quality characteristic of interest is the lengths of the produced bars. The following Figure 2 may represent a simple control chart for an in-control bar production process. Here, the center line represents the average value of the quality characteristic corresponding to the in-control state. Since all the observations fall between the limits, this process is considered to be in statistical control.

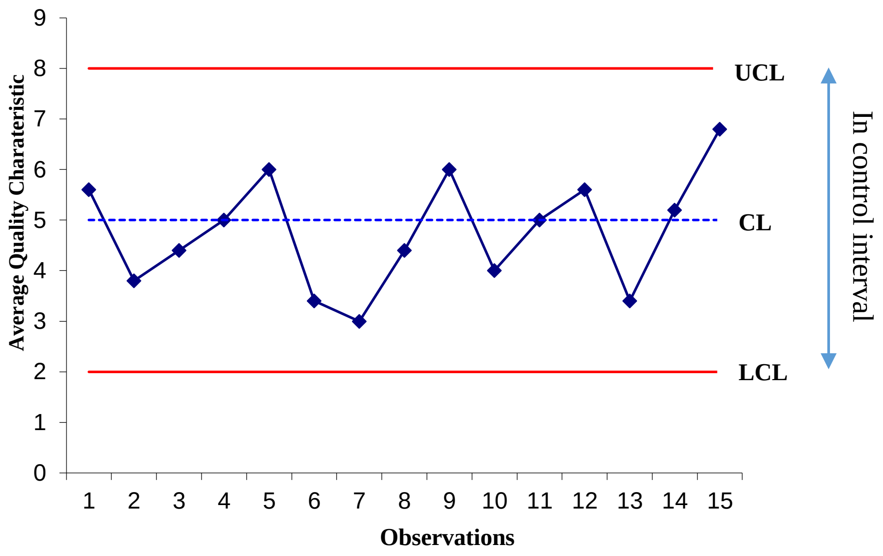


Figure 2. A typical control chart representing an in-control process

When a certain quality characteristic obtained from observations falls outside the in-control interval**,** the process is considered as out-of-control and the control chart alarms. An alarming control chart is an indication that there is an unusual variability, e.g. an assignable cause, in the process. Going back to the previous example, Figure 3 represents an out-of-control bar production process, where an alarm is triggered at point 15.

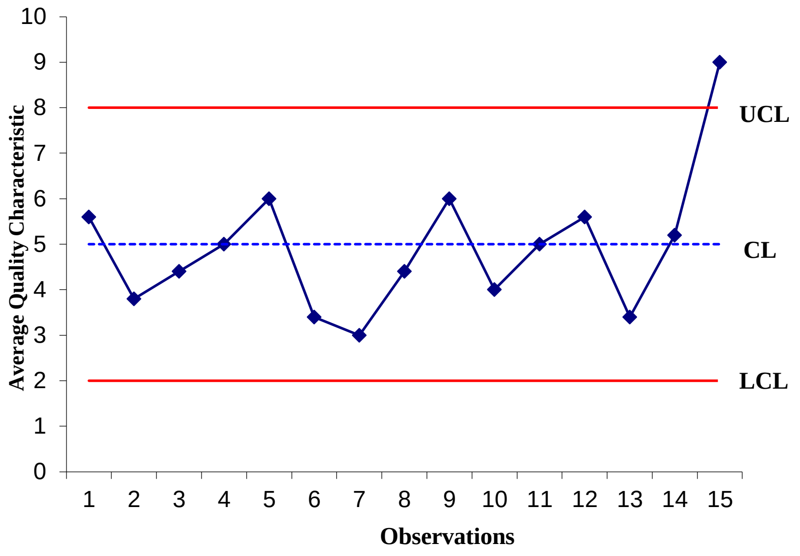


Figure 3. A simple control chart representing an out-of-control process

2.2.2. Statistical Calculations for the Chart Construction

Construction of a Shewhart Control Chart requires simple statistical calculations from a retrospective dataset. CL, UCL and LCL are calculated according to the certain statistics that are obtained through the observed values of a defined quality characteristic. Montgomery [3] provides a general model for a Shewhart-type control chart:

$$UCL = \mu_w + L\sigma_w$$
$$CL = \mu_w$$
$$LCL = \mu_w - L\sigma_w$$

Here, *w* is a control statistic that measures some quality characteristic of interest, mean of *w* is denoted by μ_w, and standard deviation of w is denoted by σ_w. As an example, consider sample average ($\bar{x}$) as the control statistic (*w*). The following equations are given for this example to describe how the control limits are determined by using the sample averages of the monitored quality characteristic:

$$UCL = \mu_{\bar{x}} + L\sigma_{\bar{x}}$$
$$CL = \mu_{\bar{x}}$$
$$LCL = \mu_{\bar{x}} - L\sigma_{\bar{x}}$$

For a "***sample i***" with size *n*, the average of the sample, denoted by $\bar{x}_i$, is calculated as

$$\bar{x}_i = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

The CL is determined through the mean value (μ) of the sample averages, represented with $\boldsymbol{\mu_{\bar{x}}}$ in the formula. For example, if m samples are considered for determining the process mean, $\boldsymbol{\mu_{\bar{x}}}$ is calculated through

$$\mu_{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \ldots + \bar{x}_m}{m}$$

In calculations, $\boldsymbol{L}$ represents the distance of the limits from the center line. This distance is usually expressed in terms of standard deviation units. The general consideration for the L is 3σ that is also referred to as ***"three sigma control limits"***. Upper and lower control limits (UCL and LCL) are determined through the multiplication of L and the sample standard deviation $\sigma_{\bar{x}}$,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of individual observations. Note that, in the design of the chart, here it is assumed that an observation come from an in-control process and no out-of-control data are presented. Since this is often unknown, a Phase I implementation of control charts is often performed. Details regarding how the control limits are obtained in Phase I are described in the next sections.

2.2.3. Type I - II errors and Their Relationship with Selection of "L"

Montgomery [3] stated that there is a close relationship between control charts and hypothesis testing. In control charting, the process mean is tested with respect to the hypothesis of the process is in a state of statistical control. During process monitoring, if a sample average $\boldsymbol{\bar{x}_i}$ falls between the defined control limits, process mean is considered as ***in-control*** which may be expressed as $\mu_{\bar{x}} = \mu_0$ in hypothesis testing**.** On the other hand, when a sample average $\boldsymbol{\bar{x}_i}$ falls outside the in-control interval, process mean is considered as ***out-of-control*** and may be expressed as $\mu_{\bar{x}} \neq \mu_0$ in hypothesis testing**.** In spite of this close relationship, there are also differences between these two concepts. Interested readers may find detailed discussions in the literature regarding the similarities and differences of control charting and hypothesis testing, see for example Woodall [11].

The most important contribution of hypothesis testing to the control charting is the performance analysis. As in the hypothesis testing, two types of errors emerges in a control chart, as a result of the selected limits.

- ***Type I (α) error*** is the probability of concluding the process is out of control when it is really in control.
- ***Type II (β) error*** is the probability of concluding the process is in control when it is really out of control.

Although three sigma control limits are commonly preferred in control chart design, deciding to the value of *L*, is always an issue for the practitioners; because, it is directly related with the Type I and Type II errors. In terms of Type I (α) error, *L* can be defined as,

$$L = Z_{\alpha/2}$$

Therefore, in a typical control chart, selection of *L* requires careful considerations. For example, Type I (α) error is 0.0027 in the common selection of three sigma control limits (*L* = *3*) when the distribution of observations is normal. This means that a Type I (α) error, which is a false out-of-control alarm, will occur 27 times out of 10,000 observations on average. This values is obtained as

$$Z_{0.0027/2} = Z_{0.000135} \rightarrow Z = 3 \text{ (See the standard normal distribution table).}$$

When one desire to reduce Type I (α) error to 0.001 in one direction, the *"L"* should be selected as 3.09 sigma:

$$Z_{0.002/2} = Z_{0.001} \rightarrow Z = 3.09 \text{ (See the standard normal distribution table).}$$

Selecting a greater value for *L* means that control limits will move away from the center line and the in-control interval will expand. Since the probability of falling outside the limits will decrease, the probability of Type I (α) error will reduce. However, since the points will have a higher probability of falling to the in control interval when it is wider, Type II (β) errors also increase. As a conclusion the following two behavior will occur according to the selection of *L*:

- Larger L → wider in-control interval → Type I (α) will decrease, Type II (β) will increase.
- Smaller L → narrower in-control interval → Type I (α) will increase, Type II (β) will decrease.

There are several discussions still have been going on regarding the selection of *L*. Montgomery [3] concluded that selecting the "***L=3 sigma"*** as a multiple of standard deviation is common since it gives good results in practice.

Selection of *L* also has a close relationship with performance evaluation of control charts. Average run length (ARL) metric is calculated trough Type I (α) and Type II (β) errors. Details regarding the relationship of *L* selection and chart performance will be given in Section 2.4.

2.2.4. Further Details on Shewhart Control Charts

Although the in-control interval is usually determined according to the **±** ***3 sigma limits***, there are other proposed methods to increase sensitivity to the small shifts in the mean. Additionally, there are some generally accepted run rules, which have been used to detect unusual patterns on the control charts. These rules are mostly based on "non-random" patterns on the chart. Interested readers may see Western Electric Handbook [13] for the patterns indicating out of control processes. These run rule are usually good in terms of improving sensitivity; however, they are also poor in terms of increasing false alarm rates [11].

There are several control charts that are originated from the Shewhart Type Control Charts. *"R"* and *"s"* charts are used to monitor a process's variability, *"p"* and *"np"* charts are used to monitor non-conforming products, and *"c"* and *"u"* charts are used to monitor nonconformities [3]. Moreover, for monitoring the parameters of all the standard probability distributions, there are various proposed control charts in the literature [1].

Although Shewhart control charts are very useful in practice, they have a major disadvantage. Montgomery [3] stated that Shewhart control charts are relatively insensitive to the small shifts; because, they only consider the last sample mean and ignore the information from previous samples. One may find various research attempts for Shewhart control charts, aiming to use the information from previous samples and thereby increase the sensitivity of the charts. They, however, have not became so popular since they decrease the simplicity and performance of Shewhart control charts [14].

After researches had focused more on to the control charting, some complementary charts such as ***Cumulative Sum (CUSUM) and Exponentially Weighted Moving Average (EWMA)*** type control charts were developed. The main difference between these new charts and Shewhart type control charts is that CUSUM and EWMA allows information to be accumulated over time. Therefore, EWMA and CUSUM are also referred as time-weighted control charts in the literature. Since these charts accumulate information from previous samples, they are effective charts when small shifts are of concern. Interested readers are referred to Hawkins and Olwell [15], Lucas and Saccucci [16], and Montgomery [3] for details regarding to the CUSUM and EWMA control charts.

2.3. Phase I and Phase II Implementations in SPC

SPC is implemented in two phases when process parameters are unknown: ***Phase I,*** in which retrospective analysis is implemented, and ***Phase II,*** in which process monitoring is implemented [5]. Being a retrospective study of the observations from a process, Phase I analysis aims to understand the process, characterize its in-control process performance and establish an online monitoring scheme, by investigating historical data. On the other hand, Phase II analysis aims to monitor the process in real time (online) and alarm if any abnormality occurs.

Control charts have a major role in Phase I implementation. Phase I analysis is an iterative process and first control limits are assumed as trial limits. Assignable causes of variation indicate that there is an abnormality in the process; thus, corrective actions must be immediately taken when any assignable cause is detected. Samples with assignable causes are investigated, and when the sources of the causes can be determined and removed, the samples are removed from the data. For each removal of the assignable cause, control limits are revised. This process continues until obtaining a stable process performance [5]. The output of a Phase I analysis is a model for the in-control state, which is then used to design a control chart for online process monitoring in Phase II (not necessarily the same type of chart as it was used for Phase I analysis). In phase II, the process data is monitored online to quickly detect process shifts. For each phase, different statistical methods should be used [17].

Despite the importance of Phase I analysis for SPC, most researchers ignored the distinction between Phase I and II in the current teaching and practice of SPC. It is common to assume that control chart parameters are known for Phase II implementation. In practice, however, this is not the case; because, Phase II control limits are obtained through a Phase I implementation. Vining [18] stated that the distinction between Phase I and II implementations of control charts is an issue that have attracted the attention of researchers recently.

2.3.1. Phase I Implementation

The major purpose of Phase I analysis is ensuring that the process is operating in an in-control state. An in control process should be operating around an acceptable target without assignable causes. Several statistical tools including control charts, graphical and numerical analysis tools, and data handling methods are employed in this phase [5]. Among these methods, control charts have a substantial position for Phase I analysis. More specifically, Shewhart Control Charts are appropriate for Phase I implementation; because, they are used for general purposes and effective in detecting large shifts. The following steps are implemented sequentially as a part of Phase I Analysis [4] [19].

i. As an initial step, identification of an appropriate probabilistic model for observations is essential.

ii. Based on the considered model of the observations, an appropriate control chart is selected.

iii. The parameters required for the control chart design are estimated.

Although the parameter estimation step of a Phase I analysis may be skipped when the parameters are assumed to be known, in practice this is not the case since there is often not enough historical knowledge and expert opinion. In the literature, the situation in which parameters are known is referred as the "standards known case or Case K"; the situation in which the parameters are unknown is referred as the "standards unknown case or Case U".

iv. Estimated parameters are first used to calculate trial control limits of the selected control chart, which is then used to identify outliers.

v. Outliers, as the out-of-control points exceeding the control limits, are investigated for possible process upsets and corrective actions are taken accordingly.

vi. Outliers that indicate a deviation from the in-control state of a process are then filtered and the parameter estimates are revised.

vii. The control limits are recalculated from the revised estimates and these iterations are continued until all the remaining observations fall between the control limits. Mind that in-control a false alarm rate is expected, therefore, some observations may fall close but outside the control limits.

Sometimes, it may not be possible to classify an outlier as a process upset. In such cases, possible actions are;

a. Filter the observation, which is common when it is believed that the point does not represent an in-control process state.
b. Retain the observation, which is common when it is believed that the point does represent an in-control process state.

Phase I analysis will lead to parameter estimation from incomplete data if outliers are filtered. On the other hand, if such outliers are erroneously kept in the sample, they are expected to influence the parameter estimates used in chart design, and consequently the process monitoring performance in Phase II. As conclusion, the performance of process monitoring in Phase II depends on to the success of the former Phase I implementation. See Figure 4 for a flow chart of the activities used in Phase I implementation.

2.3.2. Phase II Implementation

After obtaining a stable and reliable process in Phase I, the final control chart design is assumed as representative of an in-control process performance. This design is then used in Phase II, in which future process monitoring is executed. Therefore, Phase II implementation of control charts is for process monitoring, not achieving an in-control chart design. An alarming control chart during process monitoring is an indicator of an abnormality in the process. When such a situation exists, the ***out-of-control-action plan (OCAP)*** related with the control chart must be activated. The OCAP aims to resolve out of control condition and eliminate the assignable cause [3].

As mentioned in previous sections; in the literature, process parameters are often assumed as known before the Phase II implementation. However, a chart designer usually must estimate process parameters in real world applications. Hence, parameter estimation in Phase I analysis will directly effects the chart design and process monitoring performance in Phase II [7]. Statistical performance of Phase II is measured by the probability of a signal or some parameter of the run length distribution. Average run length (ARL) is the most common performance indicator used in control charts. Since it is more important to detect reasons of out-of-control points rather than false alarms in Phase I, ARL is not a good measure for that phase. On the other hand, monitoring the process is the main task of Phase II implementation in which false alarm rates are important. Thus, ARL is a good performance evaluator for Phase II control charts [3]. See the following section for the performance metrics used in control charts. See the following Figure 4 for a flow chart of the activities used in Phase II implementation.

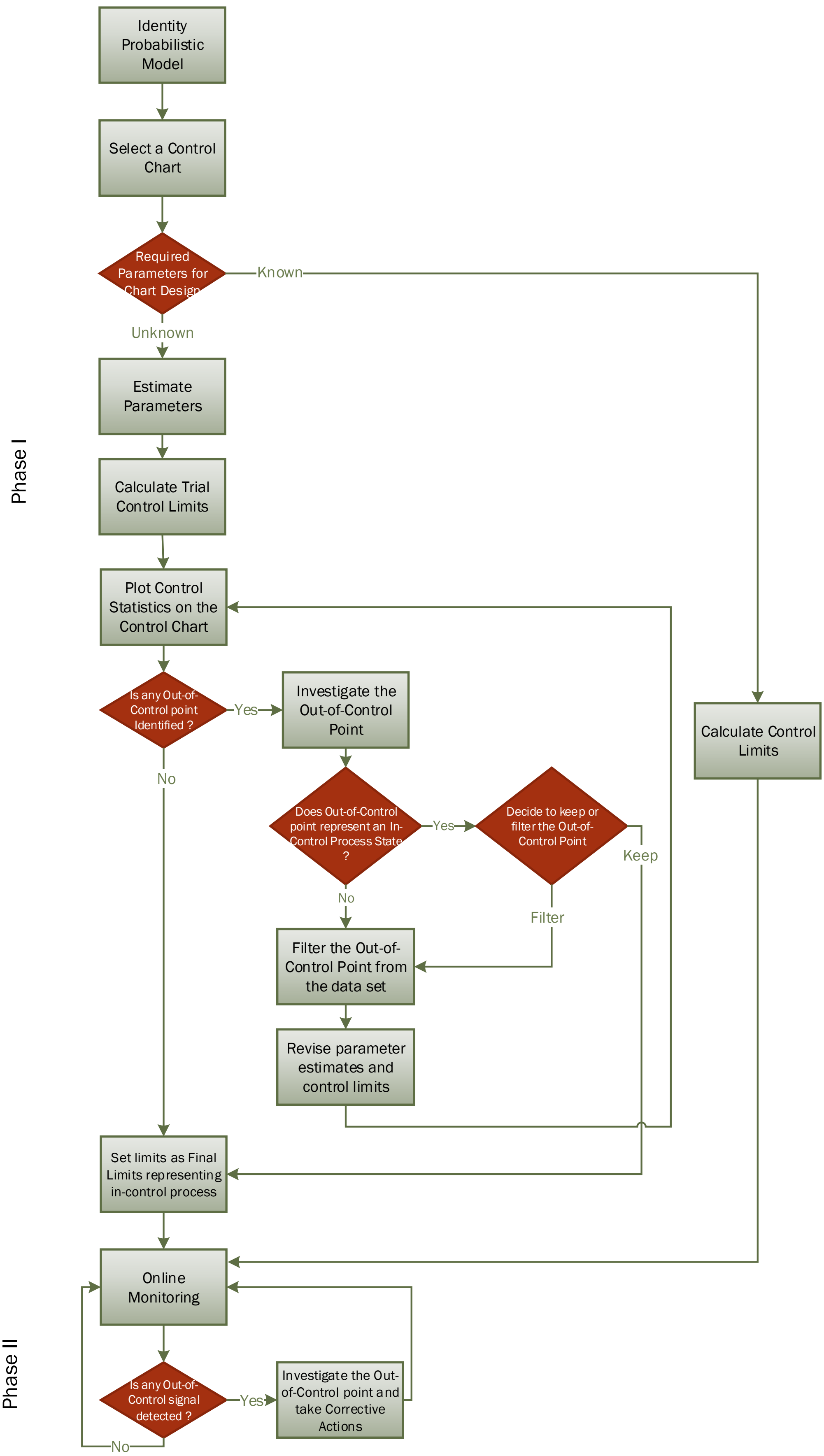


Figure 4. Flowchart of Phase I-II implementations

2.4. Metrics for Control Chart Performance

The number of samples required until the chart alarms is referred to as run length. Although the run length distribution is usually skewed to the right, the average run length (ARL) is the most popular performance metric used in control charts [1]. The primary purpose of ARL is to describe the average number of points plotted before a point falls outside of the control limits. For any Shewhart Type Control Chart (when observations are uncorrelated), the ARL can be calculated as [3]:

$$\text{ARL} = \frac{1}{p}$$

Here, ***p*** is the probability that any point falls outside of the in-control interval. To evaluate the monitoring performance of a control chart, the average number of plotted statistics before a false signal is investigated as the ***in-control ARL*** **(*ARL_0*)**, while its performance to quickly detect out-of control points is quantified through ***out-of control ARL*** **(*ARL_1*).**

Remember the Type I (α) and Type II (β) errors described previously. The former is the probability of concluding the process is out of control when it is really in control; the latter is the probability of concluding the process is in control when it is really out of control. Since Type I (α) error represent the probability of a point to fall outside of the limits for an in-control process, the probability "α" represents the probability "p" in ARL_0 calculations. For example, remember that Type I (α) error is 0.0027 for the "three sigma control limits (L=3)" and when the observations are normal distributed. Therefore, ARL_0 is calculated as:

$$ARL_0 = \frac{1}{\alpha} = \frac{1}{0.0027} = 370$$

Note that, $ARL_0 = 370$ means that, during process monitoring, a Shewhart Control Chart will signal a false alarm in on average 370 samples. In the other hand, out of control ARL (ARL_1) is used to measure the performance of a control chart to detect an out of control point (e.g. shifts in the mean). Note that, in a control chart design, sample size and the frequency of sampling are the two parameters that determines the performance of a control chart to detect small shifts in the process. When sample size increases, probability of detecting a shift in the mean increases [20].

Since Type II (β) error represents the probability of an out of control point to fall in between the control limits, the probability "β" represents the probability 1-p for out of control ARL (ARL_1) calculations. Thus, ARL_1 is calculated as

$$ARL_1 = \frac{1}{1 - \beta}$$

The probability of a point to fall between control limits (β) can be determined by using Operating-Characteristic Curves. Operating-characteristic curves for some control charts can be found in [3].

The performance of the control charts in terms of ARL results are also closely related with selection of *"L"*. Details regarding how selection of *"L"* effects control chart performance were already described in Section 2.2.3. Table 1 is given to show how different *"L"* selections for a control chart design effects ARL_0 and ARL_1 performances. As it is seen, increased *"L"* will result in enhanced ARL_0 performance, which means that the designed control chart will be giving less false alarms. On the other hand, ARL_1 performance, however, will be worse with increased *"L"*. In other words, it will take longer to detect process shifts for a chart designed with increased *"L"*. Additionally, one may also observe how Shewhart Type Control charts perform worse for small shifts, e.g. "0.5σ ".

Table 1. The relationship between the selection of L value, ARL_0 and ARL_1 performances

	ARL_0	**ARL_1 (mean shifts, in terms of "σ")**			
		0.5	1	2	4
L = 3.00	370.4	155.2	43.9	6.3	1.2
L = 3.09	499.6	201.4	54.6	7.3	1.2
L = 3.30	1034.3	380.6	93.2	10.3	1.3

2.5. Performance Evaluation in Phase I

For performance evaluation in Phase II, the major concern is detecting process shifts as quickly as possible. ARL is common metric used in performance comparison of control charts in Phase II. For Phase I, the major concern is controlling the false alarm rate. There are two common methods used to measure the statistical performance of a control chart in Phase I implementation: False alarm rate (FAR), and False alarm probability (FAP) [17].

FAR is the probability of a false alarm at every sampling stage. Control chart limits may be designed according to a desired FAR such as 0.0027. Unfortunately, some problems causing increased false alarms occur when a fixed FAR is used in Phase I. The major problems related with FAR are dependency and simultaneous comparison. Since same control limits are used to compare all subgroups, signals given by chart are somewhat statistically dependent. Additionally, comparing subgroups at the same time results in a simultaneous comparison problem [5].

FAP is the probability of a control chart to alarm at least one time when the process is in-control [11]. A specific FAP that helps to control the probability of false alarms, may be used as a metric to decide control limits on Phase I charts. Calculation of control limits for a certain FAP is the recommended method in the literature for Phase I chart design. FAP considers dependency issues between signals and handle the simultaneity problems [21].

Chakraborti et al. [5] explain the effects of FAR and FAP usages in control chart performance with the following very simple, but clear example. Assume that one desire to design a Shewhart type control chart with "3 sigma" control limits. For a control chart with "3 sigma" limits, FAR is 0.0027. This probability is the false alarm probability of each sample. Thus, the probability of not alarming is (1-0.0027). For example, if 25 samples simultaneously compared in Phase I, the probability of at least one false alarm, FAP, is equal to $1-(1-0.0027)^{25}= 0.0654$ that is unacceptably high for an in-control process. As it seen, using FAR in chart design results with a higher false alarm probability that causes a poor control chart performance. Alternatively, one may desire to have a FAP that is equal to 0.0027, instead of 0.0654. Then, $1-(1-FAR)^{25}$ should be equal to the 0.0027. Hence, FAR can be calculated as $1-(1-0.0027)^{1/25} = 0.0001$. This means that one who desires to have a FAP equals to 0.0027 need to choose "3.69 sigma control limits" in Phase I.

As a conclusion, FAP is a more reasonable performance metric used in Phase I to design control limits; because, it considers effects of estimation, dependency, and the simultaneous comparison issues [5].

3. METHODOLOGY

Phase I analysis of control charting includes parameter estimation, trial chart design, and outlier filtering, which are performed iteratively until reliable control limits are obtained. These final control limits are then used in Phase II for online monitoring to detect out-of-control states. During Phase I and Phase II implementations, there have been two common assumptions in SPC control charting literature.

In order to simplify development of control charts, process parameters have generally been assumed as known. In practice, however, process parameters are often unknown and one needs to estimate them to decide on an appropriate chart design (Chakraborti et al., 2008). Many studies have presented that the parameter estimation in Phase I implementation have significant effects on the actual chart performance of process monitoring in phase II [7]. Since effects of parameter estimation on control chart performance are significant, Phase I implementations have gained attention from the SPC researchers.

As a second assumption in the studies, process observations are generally token as independent and identically distributed over time. Although this assumption facilitate the works of SPC researchers, real industrial processes in practice often have correlated observations. When correlation exists, Phase I implementation becomes a more challenging issue. Autocorrelation between observations effect parameter estimations, and so control chart design. Adams and Tseng [22] showed that the error in estimating autocorrelation parameters causes a poor ARL performance. Furthermore, Boyles [23] considered that autocorrelation is an important issue for control chart design and proposed an approach for Phase I analysis of autocorrelated data. As a conclusion, researchers have conducted various studies to implement control charting for autocorrelated observations. In SPC, applying standard charts to the residuals of an appropriate time series model or modifying the known charts for correlated data are the mainly proposed two approaches for dependent observations [9]. Both of these two methods are explained in the Section 3.1.

The methodology employed in this thesis research deals of with both these two assumptions. Both effects of parameter estimation on chart design during Phase I implementation, and effects of Phase I analysis on performance of control charts in Phase II implementation are investigated for autocorrelated time series data. The main difference of this research from the others investigating effects of autocorrelation on the control chart performance is that, this research also investigates the effects of outliers on the parameter estimations for correlated data. When the outliers are erroneously retained in the set of observations for Phase I analysis, this will have effect on the estimates of the control chart parameters used in the design, where the effect will propagate to the monitoring performance of the chart in Phase II [19]. For instance, Ledolter [24] showed that outliers have significant effects on the forecasts of time series models.

As a methodology, autocorrelated observations from AR (1) processes are simulated by using different observation lengths and autocorrelation parameters. To generate outliers, some data with certain rates are randomly selected and converted to the outliers according to a certain rule. For the missing case, randomly selected data are assumed as outliers and filtered from the generated AR(1) observations. For Phase I implementation, these simulated data were analyzed for two extreme cases: all outliers are retained in the data, and all outliers are discarded from the data. Note that filtering of outliers here results in incomplete time series data for parameter estimation. For both cases, parameters are estimated and control charts are designed according to the estimations. Maximum Likelihood (ML) and Conditional Sum of Squares (CSS) estimators are used for parameter estimation process. The approach employed in this study has two stages:

1. The control chart design with estimated parameters,
2. The computation of the ARL performance using the obtained design.

Therefore, after the final control chart design is obtained at the end of Phase I, Phase II performance of this control chart design is also investigated through the ARL metric. As a conclusion, this research develops a general strategy for Phase I analysis of autocorrelated data with outliers [19].

3.1. Control Charts for Autocorrelated Variables Data

Standard control charts explained in Chapter 2 were proposed for the cases where consecutive observations are independent and identically distributed (*i.i.d.*). Autocorrelated observations, however, are also common in many industrial applications, and such autocorrelation may significantly affect the ARL performance of standard control charts. In fact, Maragah and Woodall [25] showed that ARL performance may be misleading if the chart design is done by ignoring the actual level of autocorrelation. Knoth and Schmid [10] showed that applying standard control charts to the dependent data affects the false alarm rates and leads to an unexpected control chart performance. Interested readers are also referred to [19], Alwan and Roberts [26], Montgomery, et al. [27], Alwan [28], Harris and Ross [29], and Runger and Willemain [30] for researcher that investigated effects of autocorrelation on control chart performance.

There are two common approaches to monitor autocorrelated observations [9]:

i. **Residual Control Charts:** Explain process dynamics by an appropriate time series model [31] and monitor the forecast residuals. If the process can be modeled adequately, residuals will be independent and standard control charts can be used on the residuals rather than the actual observations. This is a widely accepted approach for dealing with autocorrelated processes [10]. Hence, residual control charts have been intensively discussed in the literature (see for example; Alwan and Roberts [26], Harris and Ross [29], Montgomery, et al. [27], and Schmid [32]. Note that, time series parameters are often estimated from a sample of observations that is believed to represent the normal operating conditions of the process. Although the differences between the estimated and exact parameter values are expected to be small for larger sample sizes, it is not possible to achieve a purely independent residual sequence [9].

ii. **Modified Control Charts:** Monitor the autocorrelated observations by modifying the standard control limits to account for the autocorrelation. For the modified CUSUM, EWMA, and Shewhart control charts, a review can be found in Knoth and Schmid [10]. Schmid [32] compared the performance of the modified Shewhart chart with that of the residual Shewhart chart in the case of an underlying AR(1) process. He concluded that the modified chart provides the better out-of-control performance if the autocorrelation parameter is positive (as it is usually observed in practice). Analogous conclusions in the case of estimated parameters or concerning CUSUM and EWMA charts are drawn by Kramer and Schmid [33] and Knoth and Schmid [10], respectively.

There are several other researches in the literature investigating the residuals and modified charts. Interested readers are referred to, for example, Runger, Willemain and Prabhu [34], Wiel [35], Apley and Shi [36], Runger [37], Lu and Reynolds [38], Kramer and Schmid [39], Jiang [40].

In this research, in view of the findings by Schmid [41] and Kramer and Schmid [33], a two-sided modified Shewhart control chart for individual observations, which is well-suited for positively correlated data ($\phi > 0$) is used. Since the normally distributed data is considered, the control chart has symmetric control limits with the standardized control limit denoted by L. Hence, an alarm is triggered if $|X_t - \mu| \geq$ L·σ.

3.2. Investigating Effects of Estimation on the Control Chart Design Parameters in Phase I

To illustrate the effect of estimation on the control chart parameters, a stationary AR(1) model for autocorrelated observations, X, is considered as an important special case of the AutoRegressive Integrated Moving Average (ARIMA) models for time series data [31].

Consider the AR(1) model

$$X_t = \phi\, X_{t-1} + \varepsilon_t, \qquad t=1,\, 2\,, \ldots,$$

where ϕ is the autoregressive parameter, ε_t is the white noise process of *i.i.d.* normal random variables having mean 0 and variance σ_ε^2, and t is the time index. Note that an AR(1) model of a time series process regresses the observation X_t at time t on the observation X_{t-1} at time $t-1$. For a stationary process, the autocorrelation parameter assumes a value $|\phi| < 1$, the mean and variance of the process X_t are $\mu = 0$ and $\sigma^2 = \sigma_\varepsilon^2/(1-\phi^2)$, respectively. Without loss of generality but for simplicity, we let $\sigma_\varepsilon^2 = 1 - \phi^2$ so that σ^2 =1, i.e., we have the same marginal distribution for any choice of ϕ.

In this research, like as Weiß and Testik [4] and [19], two extreme cases regarding the availability of data are considered for estimating process parameters:

Case (i): All outliers are filtered from the dataset before estimation.

Case (ii): All outliers remain in the dataset while estimation.

In these settings, case (i) can be considered to be the ***"the best-case scenario"*** in terms of identifying out-of-control process data in Phase I. Note that filtering of outliers here results in incomplete time series data for parameter estimation. Because of the autocorrelation, the exact position of the filtered observations is required for the estimation methods, also see Weiß and Testik [4]. Similarly, case (ii) can be considered to be ***"the worst-case scenario"*** for identifying out-of-control process data in Phase I, where contaminated but complete time series data are available for parameter estimation.

For estimating the process parameters, here the maximum likelihood (ML) and conditional sum of squares (CSS) estimators are considered. These and further approaches have been studied intensively in the literature, both for complete and incomplete time series data (e.g., R. H. Jones [42]). In the present research, we use the readily available implementations in R's arima command [43]. The following command is used in the simulations to fit AR (1) model to the simulated time series data:

- arima(x, order = c(1,0,0), transform.pars=FALSE, method = ("ML", or "CSS")).

Here, x represents a univariate time series data. Order represents the specifications of the non-seasonal part of the ARIMA model: the components (p, d, q) are the AR oder, the degree of differencing, and the MA order, respectively. Since the time series data are simulated from an AR (1) process in this study, see the next section, order is (1, 0, 0). Transform.pars is a logical argument. If it is set as true, the AR parameters are transformed to ensure that they remain in the region of stationarity. This argument is not used for CSS method. Finally, the method represents the fitting method: ML or CSS. In this research, both methods are fitted to the generated time series data.

3.3. Experimental Settings for Simulations

Simulations were realized by using R's arima.sim command [43]. The following describes how R's arima.sim command runs to simulate time series data from ARIMA models [43]:

- arima.sim(model=list(ar=c(ϕ)), n= T, rand.gen = rnorm, mean=0, sd =$\sqrt{1-\phi^2}$)

Here, "model" represents a list with component "ar" giving the AR(1) coefficient. As stated, 0.3, 0.5, and 0.7 are used for the model argument. The length of output series is represented with "n", which is selected 50, 100, and 200 respectively. "Rand.gen" is the function to generate innovations. Here, the data are generated from a normal distribution (rnorm); thus the "mean" and "standard deviation (sd)" are the parameters of normal distribution.

By using generated data from the AR(1) model, the ML and CSS estimators were evaluated as follows:

- The mean and variance of the simulated process X_t were $\mu = 0$ and σ^2 =1, respectively.
- The values of the autocorrelation parameter ϕ used in the simulations of the process were 0.3, 0.5, and 0.7, which satisfy the stationary condition.

- In the simulations, data lengths of size *T*= 50,100, and 200 were considered to obtain the estimates $\hat{\phi}, \hat{\mu},$ and $\hat{\sigma}^2$ for $\phi, \mu,$ and σ^2, respectively. To differentiate ML and CSS estimates, subscripts are used in the following.
- 10,000 replications were conducted for each scenario and, hence, empirical distributions of the ML and CSS estimators were obtained.

Table 2 summarizes all considerations for the simulations. Here, the outliers' rate column represents the rates for both for outliers and missing observations cases. Time series data were generated for all the simulation scenarios.

Table 2. Experimental Settings Scenarios

ϕ	T	Outliers' Rate (Missing Rate)	Estimators	Estimates
0.3 0.5 0.7	50 100 200	0 0.02 0.05 0.1 0.25	ML CSS	$\hat{\phi}$ $\hat{\mu}$ $\hat{\sigma}^2$

For the case (i) above (perfect outlier identification), in each replication, observations from the generated time series with length T were randomly selected with rates $r = 0, 0.02, 0.05, 0.1$ and 0.25. Then these selected observations were treated as randomly scattered outliers within the time series data and therefore filtered (value "NA" at the respective positions), which leaves an incomplete time series dataset for estimation.

In the case (ii), where it is assumed that all the outliers are in the time series during estimation (worst case), a similar procedure is applied to randomly select the observations with rates $r = 0.02, 0.05, 0.1$ and 0.25. This time, without filtering the randomly selected observations, random variates from a normal distribution with mean 4 and variance 1 are generated and added to the selected observations, to have contaminated time series data with additive outliers [44] at rates r. This particular choice of 4σ shift in the mean for generating outliers was motivated by the popular 3σ rule, where the extension of the outliers will usually be sufficiently large to be detected by a chart with 3σ limits.

For the two cases above, the rate $r = 0$ implies no outliers and no missing observations in the time series process. In fact, the special case $r = 0$ corresponds to the studies on the effects of estimation on control charts' performance (see, e.g., Jensen, et al. [7]), where outliers and missing observations are not considered. The following simulation scenarios tree in Figure 5 is given to explain how simulation cases are branched. In total, 486 scenarios were simulated.

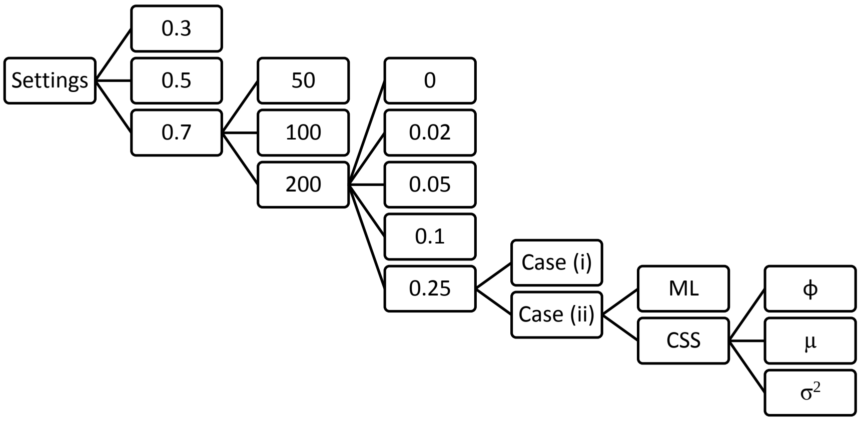


Figure 5. Simulation Scenarios Tree

3.4. Investigating Effects of Phase I Analysis on Control Chart Performance in Phase II

As it is stated in the beginning of the Chapter 3, the approach employed during this research has two stages: First, estimating parameters and obtaining a final control chart design in Phase I; second, computing ARL performance for this final design to investigate effects of Phase I Analysis on Phase II performance.

In the second stage, the effects of these parameter estimates are studied through the chart design on the performance of control charts, again considering the cases (i) and (ii). During Phase I, the use of Shewhart charts is recommended [3]. Remember that details regarding Shewhart type control charts are discussed in Section 2.2. The control charts for the autocorrelated data are also discussed in Section 3.1. Please remember that, there are two common approaches to monitor autocorrelated data: Residual Control Charts and Modified Control Charts. After some simulation studies, Schmid [41] and Kramer and Schmid [33] concluded that modified Shewhart chart with estimated parameters should be preferred for AR(1) processes with positive correlation ($\phi > 0$). As stated before, in view of the findings of these researches, a two-sided modified Shewhart control chart is used in this thesis. Normally distributed data is considered; thus, the control chart has symmetric control limits with the standardized control limit at a distance “L”, in terms of standard deviation unit. Hence, whenever $|X_t - \mu| \geq L \cdot \sigma$, an alarm is triggered.

The simulation principles and settings for parameter estimation were already discussed in previous sections. In the second stage, the estimates for ϕ from each replication were used to determine control limits through ARL_0 evaluations, where the AR(1) model with the estimated parameter $\hat{\phi}$ is used to yield an expected ARL_0 performance of 370.4. For this purpose, ARL_0 computations were conducted by means of R’s spc package [45], see Appendix A for the details. This design is then used in the second stage of our approach, where the true process parameters are used to compute the true ARL_0 performance of the chart having the control limits obtained by use of process parameter estimates. Note that, here ARL is a function of the estimators and, hence, itself random. Therefore, as the performance measures, mean ARL_0 and median ARL_0 were computed from the results of 10,000 simulation replications.

4. RESULTS AND DISCUSSION

During simulations, AR (1) time series processes are generated for different observation lengths and autocorrelation parameters, and certain parameters are estimated. As explained previously, the outliers were generated for different rates and randomly placed to the time series data before estimations. Remember that, while estimating parameters in Phase I, two cases are considered: all outliers are retained in the data (case i), and all outliers are discarded from the data (case ii). Each simulation scenario is replicated for 10,000 times; thus, parameter estimation results are obtained from these 10,000 repetitions. As stated before, to estimate the parameters, Maximum Likelihood (ML) and Conditional Sum of Squares (CSS) estimators are used. The estimation results presented in this chapter are both for ML and CSS estimators. After obtaining parameter estimates and the final control chart design at the end of Phase I, Phase II control chart performance based on this chart design is also investigated through the ARL metric.

In this chapter, the results of parameter estimations will be presented first. Then, the ARL performance of the control chart designed through the estimated parameters will be discussed.

4.1. Simulation Results for Estimates of Control Chart Design Parameters

Means and standard deviations of the parameter estimates are provided and discussed for both case (i) and case (ii) in the following sections. The results are provided for the data lengths T = 200,100, and 50 respectively.

As expected, main differences of the results when T = 200 from the corresponding results of T = 100 and 500 are; means of the estimates are closer to the true values of the parameters, and standard deviations of the estimates are smaller with the larger data length T = 200. As an example, means and standard deviations of 10,000 ϕ estimates for case (i) are compared in Figure 6. One may see how estimation results behave for different data lengths T = 200, 100, and 50. The mean results (left graph) are more close to the actual value and standard deviations (right graph) are smaller with larger data length T = 200, compared to the T = 100, and T = 50. The comparison are only provided for ML estimator here; however, CSS estimator behaves similar. Additionally, this conclusion is also same for case (ii) and for $\mu \ and \ \sigma^2$ estimations.

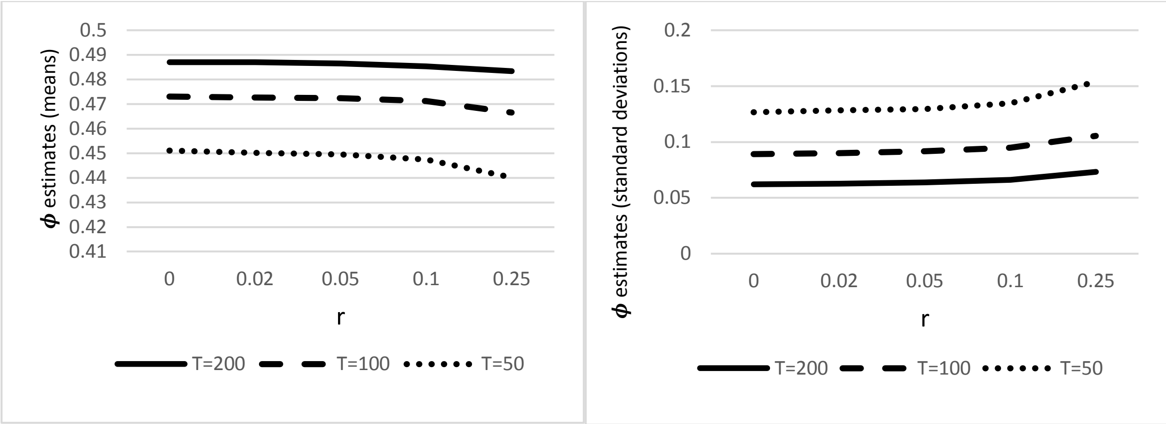


Figure 6 – Mean and standard deviation of $\phi = 0.5$ estimates obtained by ML estimator

Remember that this research focuses on the effects of outliers on parameter estimations when the data is correlated. The parameter estimation results of T = 200, estimated from 10,000 repetitions, for Case (i) and for Case (ii), are given in the Table 3 and Table 4, respectively. Note that, there are two major parts in the tables: Mean of Parameter Estimates (left part), and Standard Deviation of Parameter Estimates (right part). Here, "mean" represents the average value of 10,000 estimates, and "standard deviation" represents the standard deviation of 10,000 estimates.

Table 3. Means and standard deviations of parameter estimates obtained from 10,000 replications with $T = 200$: The case of outliers being filtered from the dataset before estimation, Case (i).

		Mean of Parameter Estimates						**Standard Deviation of Parameter Estimates**					
$\boldsymbol{\phi}$	$\boldsymbol{r}$	$\hat{\boldsymbol{\phi}}_{\text{ML}}$	$\hat{\boldsymbol{\phi}}_{\text{CSS}}$	$\hat{\boldsymbol{\mu}}_{\text{ML}}$	$\hat{\boldsymbol{\mu}}_{\text{CSS}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{ML}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{CSS}}$	$\hat{\boldsymbol{\phi}}_{\text{ML}}$	$\hat{\boldsymbol{\phi}}_{\text{CSS}}$	$\hat{\boldsymbol{\mu}}_{\text{ML}}$	$\hat{\boldsymbol{\mu}}_{\text{CSS}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{ML}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{CSS}}$
	0	0.2908	0.2908	-0.0024	-0.0024	0.9931	0.9930	0.0674	0.0674	0.0954	0.0959	0.1090	0.1091
	0.02	0.2906	0.2906	-0.0025	-0.0025	0.9927	0.9925	0.0684	0.0686	0.0958	0.0979	0.1101	0.1118
0.3	0.05	0.2904	0.2905	-0.0024	-0.0026	0.9927	0.9924	0.0703	0.0712	0.0964	0.1011	0.1110	0.1144
	0.1	0.2901	0.2900	-0.0025	-0.0022	0.9927	0.9914	0.0733	0.0749	0.0978	0.1072	0.1137	0.1211
	0.25	0.2879	0.2890	-0.0022	-0.0024	0.9905	0.9900	0.0854	0.0904	0.1020	0.1279	0.1235	0.1479
	0	0.4871	0.4871	0.0011	0.0013	0.9835	0.9835	0.0621	0.0621	0.1210	0.1219	0.1268	0.1273
	0.02	0.4870	0.4871	0.0013	0.0015	0.9834	0.9836	0.0626	0.0633	0.1213	0.1246	0.1275	0.1300
0.5	0.05	0.4865	0.4865	0.0011	0.0007	0.9836	0.9835	0.0637	0.0652	0.1217	0.1282	0.1284	0.1343
	0.1	0.4854	0.4854	0.0009	0.0013	0.9822	0.9819	0.0660	0.0696	0.1231	0.1361	0.1305	0.1420
	0.25	0.4834	0.4843	0.0008	0.0003	0.9814	0.9820	0.0733	0.0835	0.1252	0.1625	0.1376	0.1730
	0	0.6851	0.6850	0.0007	0.0005	0.9741	0.9739	0.0518	0.0520	0.1662	0.1689	0.1663	0.1675
	0.02	0.6850	0.6850	0.0008	0.0007	0.9739	0.9745	0.0521	0.0530	0.1662	0.1714	0.1666	0.1714
0.7	0.05	0.6849	0.6845	0.0008	0.0007	0.9739	0.9735	0.0526	0.0550	0.1664	0.1779	0.1672	0.1767
	0.1	0.6846	0.6843	0.0004	0.0003	0.9737	0.9755	0.0533	0.0577	0.1666	0.1880	0.1684	0.1876
	0.25	0.6832	0.6828	0.0011	-0.0020	0.9727	0.9816	0.0571	0.0707	0.1684	0.2279	0.1727	0.2372

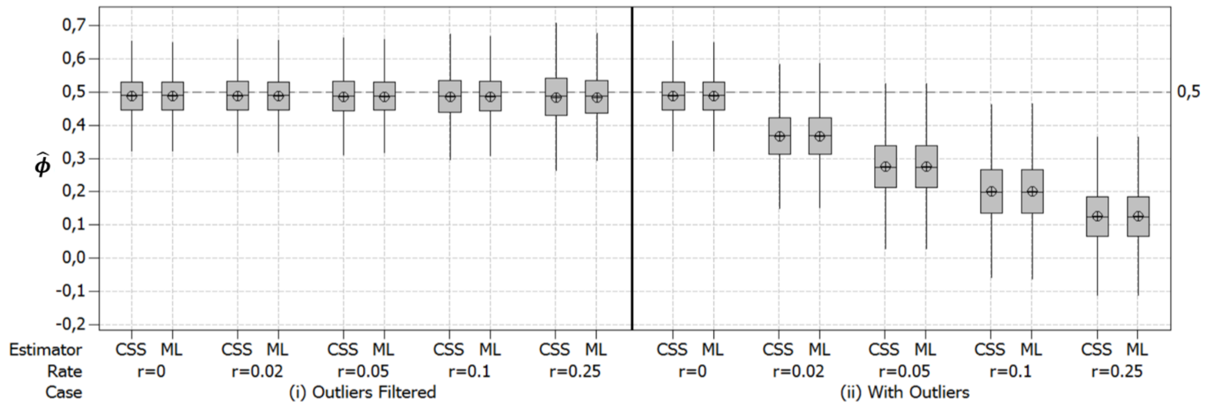
Table 4. Means and standard deviations of parameter estimates obtained from 10,000 replications with $T = 200$: The case of outliers being in the dataset while estimation, Case (ii).

		Mean of Parameter Estimates						**Standard Deviation of Parameter Estimates**					
ϕ	r	$\hat{\phi}_{ML}$	$\hat{\phi}_{CSS}$	$\hat{\mu}_{ML}$	$\hat{\mu}_{CSS}$	$\hat{\sigma}^2_{ML}$	$\hat{\sigma}^2_{CSS}$	$\hat{\phi}_{ML}$	$\hat{\phi}_{CSS}$	$\hat{\mu}_{ML}$	$\hat{\mu}_{CSS}$	$\hat{\sigma}^2_{ML}$	$\hat{\sigma}^2_{CSS}$
	0	0.2908	0.2908	-0.0024	-0.0024	0.9931	0.9930	0.0674	0.0674	0.0954	0.0959	0.1090	0.1091
	0.02	0.2186	0.2185	0.0778	0.0777	1.3280	1.3280	0.0742	0.0742	0.0977	0.0982	0.2088	0.2092
0.3	0.05	0.1649	0.1649	0.1973	0.1974	1.7980	1.7980	0.0777	0.0777	0.1077	0.1080	0.4194	0.4197
	0.1	0.1211	0.1211	0.3971	0.3972	2.5180	2.5180	0.0799	0.0799	0.1369	0.1373	0.7556	0.7561
	0.25	0.0762	0.0762	0.9922	0.9920	4.1450	4.1450	0.0779	0.0779	0.2671	0.2672	1.5316	1.5316
	0	0.4871	0.4871	0.0011	0.0013	0.9835	0.9835	0.0621	0.0621	0.1210	0.1219	0.1268	0.1273
	0.02	0.3673	0.3673	0.0812	0.0813	1.3150	1.3150	0.0806	0.0805	0.1228	0.1236	0.2205	0.2213
0.5	0.05	0.2760	0.2760	0.2008	0.2010	1.7860	1.7870	0.0917	0.0917	0.1305	0.1311	0.4240	0.4253
	0.1	0.2020	0.2020	0.4028	0.4028	2.5260	2.5260	0.0960	0.0960	0.1580	0.1585	0.7584	0.7587
	0.25	0.1270	0.1270	1.0060	1.0060	4.2010	4.2000	0.0900	0.0900	0.2788	0.2791	1.5697	1.5698
	0	0.6851	0.6850	0.0007	0.0005	0.9741	0.9739	0.0518	0.0520	0.1662	0.1689	0.1663	0.1675
	0.02	0.5151	0.5149	0.0807	0.0804	1.3040	1.3040	0.0928	0.0927	0.1675	0.1691	0.2426	0.2428
0.7	0.05	0.3883	0.3882	0.2005	0.2005	1.7790	1.7790	0.1099	0.1098	0.1736	0.1749	0.4392	0.4396
	0.1	0.2836	0.2836	0.4013	0.4016	2.5080	2.5090	0.1153	0.1152	0.1950	0.1961	0.7690	0.7700
	0.25	0.1787	0.1787	1.0020	1.0020	4.1620	4.1620	0.1079	0.1079	0.2979	0.2983	1.5316	1.5315

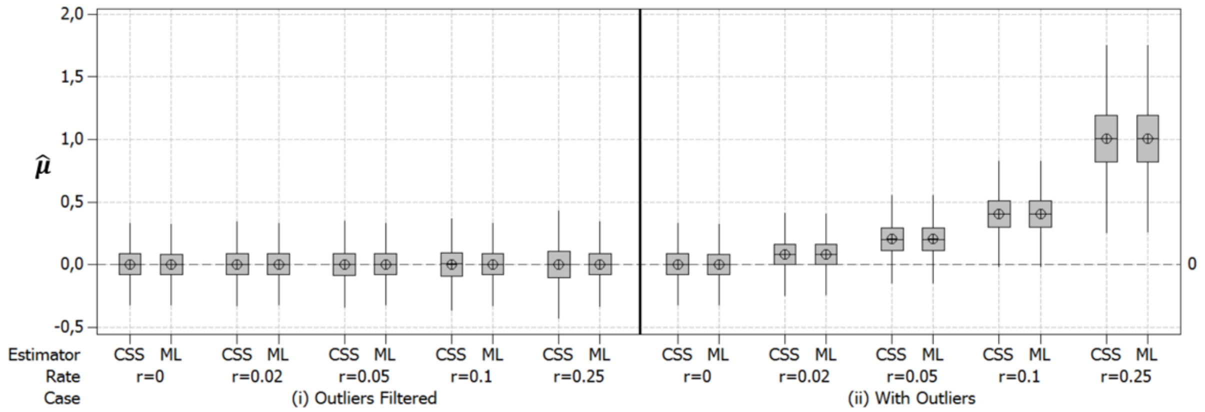
Consider the estimation results for the ***"parameter ϕ".*** The ML and CSS estimators perform nearly equivalently in terms of their means, with some negative bias in all cases, which increases as *"r (outliers' rate)"* increases. In general, the effect of incomplete datasets used in case (i) is slight on the mean of estimates from both estimators for ϕ, while the standard deviations of the CSS estimates get larger than the ones of ML estimates with increasing r. However, with the contaminated datasets used in case (ii), even 2 % outliers in the dataset heavily affect the means of the ML and CSS estimates and a strong negative bias can be observed. This observation is completely analogous to the findings by Weiß and Testik [4] for discrete autoregressive processes.

Now consider the results of estimation corresponding to the ***"parameters μ and σ^2",*** where the observed patterns are the same for both of the parameters. Here, the ML and CSS estimators of the parameters perform nearly equivalently in terms of their means, while slight deviations seem to have no patterns and can be attributable to the simulation error. In terms of the standard deviation of the estimates, CSS estimator is more sensitive to the incomplete datasets of case (i) than the ML estimator. Although the means of the estimators are not significantly affected when incomplete datasets of case (i) are used, bias and standard deviation of estimators significantly increase, as r increases, with the use of contaminated datasets from case (ii).

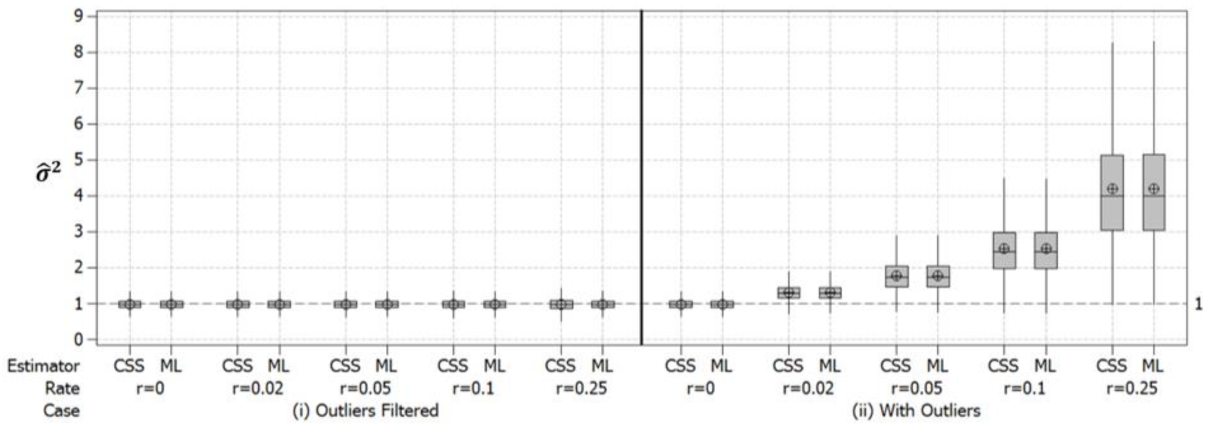
All these results are also represented with the boxplots to visualize the given conclusions above. Note that, only the boxplots of data length $T = 200$ and $\phi = 0.5$ are provided in here due to similarity of conclusions and maintaining the integrity of the text, but complete boxplots for the remaining autocorrelation parameters $\phi = 0.3$ and 0.5 are available in Appendix B.



a. Box plot of estimates for the autocorrelation parameter ϕ.



b. Box plot of estimates for the mean μ.



c. Box plot of estimates for the variance σ^2.

Figure 7. Box plots of the parameter estimates from simulations – $T = 200, \phi = 0.5$.

As stated at the beginning of this section, estimation results for T = 200, 100 and 50 are similar in terms of patterns. The only difference is that, T = 200 results have precise closer means of the estimates to the actual value and smaller standard deviations compared to the T = 100, and 50. The parameter estimation results of T = 100 and 50, estimated from 10,000 repetitions for Case (i) and for Case (ii), are given in the Tables 5-8, respectively.

Table 5. Means and standard deviations of parameter estimates obtained from 10,000 replications with $T = 100$: The case of outliers being filtered from the dataset before estimation.

		Mean of Parameter Estimates						**Standard Deviation of Parameter Estimates**					
$\boldsymbol{\phi}$	$\boldsymbol{r}$	$\hat{\boldsymbol{\phi}}_{\text{ML}}$	$\hat{\boldsymbol{\phi}}_{\text{CSS}}$	$\hat{\boldsymbol{\mu}}_{\text{ML}}$	$\hat{\boldsymbol{\mu}}_{\text{CSS}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{ML}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{CSS}}$	$\hat{\boldsymbol{\phi}}_{\text{ML}}$	$\hat{\boldsymbol{\phi}}_{\text{CSS}}$	$\hat{\boldsymbol{\mu}}_{\text{ML}}$	$\hat{\boldsymbol{\mu}}_{\text{CSS}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{ML}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{CSS}}$
0.3	0	0.2804	0.2804	-0.0011	-0.0013	0.9808	0.9808	0.0963	0.0964	0.1365	0.1380	0.1520	0.1528
	0.02	0.2801	0.2804	-0.0011	-0.0011	0.9807	0.9808	0.0980	0.0984	0.1372	0.1406	0.1533	0.1558
	0.05	0.2795	0.2799	-0.0010	-0.0015	0.9797	0.9793	0.1006	0.1017	0.1382	0.1460	0.1558	0.1622
	0.1	0.2790	0.2798	-0.0016	-0.0011	0.9803	0.9793	0.1045	0.1070	0.1400	0.1543	0.1581	0.1698
	0.25	0.2727	0.2749	-0.0014	-0.0024	0.9777	0.9735	0.1234	0.1297	0.1474	0.1861	0.1713	0.2037
0.5	0	0.4731	0.4732	-0.0019	-0.0024	0.9697	0.9701	0.0892	0.0894	0.1731	0.1755	0.1780	0.1801
	0.02	0.4727	0.4729	-0.0020	-0.0021	0.9692	0.9698	0.0901	0.0914	0.1735	0.1790	0.1789	0.1841
	0.05	0.4724	0.4724	-0.0018	-0.0030	0.9693	0.9697	0.0918	0.0945	0.1742	0.1852	0.1805	0.1918
	0.1	0.4712	0.4718	-0.0014	-0.0022	0.9687	0.9695	0.0948	0.1002	0.1751	0.1964	0.1829	0.2015
	0.25	0.4666	0.4672	-0.0015	0.0008	0.9650	0.9647	0.1056	0.1200	0.1795	0.2352	0.1919	0.2473
0.7	0	0.6695	0.6690	-0.0001	-0.0005	0.9485	0.9470	0.0762	0.0766	0.2338	0.2403	0.2308	0.2317
	0.02	0.6693	0.6687	0.0000	0.0002	0.9483	0.9477	0.0766	0.0781	0.2339	0.2444	0.2312	0.2373
	0.05	0.6689	0.6683	-0.0002	-0.0006	0.9484	0.9482	0.0774	0.0804	0.2341	0.2533	0.2321	0.2464
	0.1	0.6681	0.6676	0.0002	-0.0008	0.9478	0.9515	0.0789	0.0852	0.2348	0.2679	0.2327	0.2623
	0.25	0.6654	0.6651	0.0006	-0.0014	0.9460	0.9719	0.0848	0.1043	0.2367	0.3413	0.2398	0.4457

Table 6. Means and standard deviations of parameter estimates obtained from 10,000 replications with $T = 100$: The case of outliers being in the dataset while estimation.

		Mean of Parameter Estimates						**Standard Deviation of Parameter Estimates**					
$\boldsymbol{\phi}$	$\boldsymbol{r}$	$\hat{\boldsymbol{\phi}}_{\text{ML}}$	$\hat{\boldsymbol{\phi}}_{\text{CSS}}$	$\hat{\boldsymbol{\mu}}_{\text{ML}}$	$\hat{\boldsymbol{\mu}}_{\text{CSS}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{ML}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{CSS}}$	$\hat{\boldsymbol{\phi}}_{\text{ML}}$	$\hat{\boldsymbol{\phi}}_{\text{CSS}}$	$\hat{\boldsymbol{\mu}}_{\text{ML}}$	$\hat{\boldsymbol{\mu}}_{\text{CSS}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{ML}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{CSS}}$
0.3	0	0.2804	0.2804	-0.0011	-0.0013	0.9808	0.9808	0.0963	0.0964	0.1365	0.1380	0.1520	0.1528
	0.02	0.2094	0.2093	0.0790	0.0788	1.3130	1.3130	0.0998	0.0997	0.1381	0.1395	0.2500	0.2517
	0.05	0.1561	0.1561	0.1985	0.1984	1.7850	1.7850	0.1055	0.1054	0.1457	0.1470	0.4511	0.4540
	0.1	0.1126	0.1125	0.3995	0.3993	2.5190	2.5190	0.1062	0.1061	0.1696	0.1710	0.7951	0.7977
	0.25	0.0663	0.0663	1.0000	1.0000	4.1790	4.1780	0.1054	0.1054	0.2850	0.2859	1.5762	1.5768
0.5	0	0.4731	0.4732	-0.0019	-0.0024	0.9697	0.9701	0.0892	0.0894	0.1731	0.1755	0.1780	0.1801
	0.02	0.3536	0.3534	0.0782	0.0777	1.3040	1.3040	0.1055	0.1054	0.1740	0.1761	0.2649	0.2674
	0.05	0.2652	0.2652	0.1978	0.1974	1.7770	1.7770	0.1163	0.1164	0.1809	0.1826	0.4643	0.4666
	0.1	0.1952	0.1952	0.3952	0.3948	2.4760	2.4770	0.1199	0.1198	0.2009	0.2028	0.7859	0.7882
	0.25	0.1166	0.1166	0.9999	1.0000	4.1640	4.1640	0.1157	0.1158	0.3055	0.3066	1.5757	1.5767
0.7	0	0.6695	0.6690	-0.0001	-0.0005	0.9485	0.9470	0.0762	0.0766	0.2338	0.2403	0.2308	0.2317
	0.02	0.6693	0.6687	0.0800	0.0795	1.2790	1.2770	0.1163	0.1160	0.2352	0.2390	0.3019	0.3037
	0.05	0.6689	0.6683	0.2005	0.2001	1.7600	1.7590	0.1326	0.1324	0.2400	0.2431	0.4862	0.4895
	0.1	0.6681	0.6676	0.4020	0.4016	2.4920	2.4900	0.1374	0.1373	0.2551	0.2574	0.8012	0.8022
	0.25	0.6654	0.6651	0.9989	0.9984	4.1210	4.1200	0.1308	0.1307	0.3443	0.3463	1.5640	1.5655

Table 7. Means and standard deviations of parameter estimates obtained from 10,000 replications with $T = 50$: The case of outliers being filtered from the dataset before estimation.

		Mean of Parameter Estimates						Standard Deviation of Parameter Estimates					
$\boldsymbol{\phi}$	$\boldsymbol{r}$	$\hat{\phi}_{\text{ML}}$	$\hat{\phi}_{\text{CSS}}$	$\hat{\mu}_{\text{ML}}$	$\hat{\mu}_{\text{CSS}}$	$\hat{\sigma}^2_{\text{ML}}$	$\hat{\sigma}^2_{\text{CSS}}$	$\hat{\phi}_{\text{ML}}$	$\hat{\phi}_{\text{CSS}}$	$\hat{\mu}_{\text{ML}}$	$\hat{\mu}_{\text{CSS}}$	$\hat{\sigma}^2_{\text{ML}}$	$\hat{\sigma}^2_{\text{CSS}}$
	0	0.2625	0.2623	0.0009	0.0010	0.9665	0.9655	0.1353	0.1354	0.1912	0.1951	0.2127	0.2153
	0.02	0.2615	0.2617	0.0008	0.0003	0.9658	0.9642	0.1377	0.1384	0.1922	0.1993	0.2153	0.2200
0.3	0.05	0.2608	0.2612	0.0007	0.0009	0.9656	0.9642	0.1403	0.1415	0.1933	0.2044	0.2165	0.2263
	0.1	0.2578	0.2593	0.0005	0.0000	0.9637	0.9603	0.1489	0.1522	0.1967	0.2182	0.2228	0.2396
	0.25	0.2474	0.2531	0.0011	0.0022	0.9596	0.9537	0.1761	0.1832	0.2058	0.2659	0.2361	0.3462
	0	0.4511	0.4503	-0.0018	-0.0016	0.9461	0.944	0.1266	0.1269	0.2405	0.2471	0.2469	0.2505
	0.02	0.4502	0.4496	-0.0016	-0.0022	0.9462	0.9437	0.1284	0.1300	0.2410	0.2526	0.2482	0.2559
0.5	0.05	0.4495	0.4489	-0.0021	-0.0007	0.9456	0.9429	0.1295	0.1325	0.2416	0.2580	0.2491	0.2615
	0.1	0.4474	0.4479	-0.0021	-0.0012	0.9441	0.9422	0.1348	0.1411	0.2439	0.2771	0.2536	0.2842
	0.25	0.4402	0.4445	-0.0021	0.0035	0.9404	0.9532	0.1539	0.1707	0.2494	0.3570	0.2679	0.5081
	0	0.6375	0.6362	0.0006	0.0008	0.9007	0.8986	0.1140	0.1152	0.3252	0.3437	0.3065	0.3139
	0.02	0.6370	0.6357	0.0004	0.0019	0.9006	0.9005	0.1148	0.1174	0.3256	0.3508	0.3071	0.3237
0.7	0.05	0.6364	0.6353	0.0005	0.0008	0.9004	0.9026	0.1159	0.1201	0.3256	0.3595	0.3085	0.3391
	0.1	0.6348	0.6335	0.0004	-0.0024	0.8996	0.9069	0.1189	0.1288	0.3271	0.3882	0.3122	0.3803
	0.25	0.6281	0.6276	0.0000	0.0029	0.8944	0.9599	0.1298	0.1534	0.3307	0.6482	0.3189	1.0418

Table 8. Means and standard deviations of parameter estimates obtained from 10,000 replications with $T = 50$: The case of outliers being in the dataset while estimation.

		Mean of Parameter Estimates						Standard Deviation of Parameter Estimates					
$\boldsymbol{\phi}$	$\boldsymbol{r}$	$\hat{\phi}_{\text{ML}}$	$\hat{\phi}_{\text{CSS}}$	$\hat{\mu}_{\text{ML}}$	$\hat{\mu}_{\text{CSS}}$	$\hat{\sigma}^2_{\text{ML}}$	$\hat{\sigma}^2_{\text{CSS}}$	$\hat{\phi}_{\text{ML}}$	$\hat{\phi}_{\text{CSS}}$	$\hat{\mu}_{\text{ML}}$	$\hat{\mu}_{\text{CSS}}$	$\hat{\sigma}^2_{\text{ML}}$	$\hat{\sigma}^2_{\text{CSS}}$
	0	0.2625	0.2623	0.0009	0.0010	0.9665	0.9655	0.1353	0.1354	0.1912	0.1951	0.2127	0.2153
	0.02	0.1941	0.1936	0.0814	0.0813	1.3020	1.3000	0.1386	0.1381	0.1924	0.1959	0.3146	0.3211
0.3	0.05	0.1534	0.1530	0.1621	0.1619	1.6270	1.6250	0.1417	0.1414	0.1955	0.2000	0.4399	0.4472
	0.1	0.0974	0.0972	0.4022	0.4019	2.5070	2.5030	0.1443	0.1440	0.2167	0.2214	0.8492	0.8552
	0.25	0.0565	0.0564	0.9579	0.9583	4.0510	4.0500	0.1442	0.1441	0.3095	0.3143	1.5785	1.5851
	0	0.4511	0.4503	-0.0018	-0.0016	0.9461	0.944	0.1266	0.1269	0.2405	0.2471	0.2469	0.2505
	0.02	0.3334	0.3323	0.0798	0.0793	1.2810	1.2770	0.1425	0.1416	0.2419	0.2470	0.3397	0.3456
0.5	0.05	0.2640	0.2633	0.1597	0.1596	1.6020	1.5990	0.1492	0.1488	0.2438	0.2484	0.4544	0.4627
	0.1	0.1723	0.1720	0.3981	0.3979	2.4710	2.4690	0.1548	0.1545	0.2617	0.2667	0.8411	0.8502
	0.25	0.1020	0.1019	0.9591	0.9588	4.0470	4.0440	0.1555	0.1555	0.3391	0.3438	1.5818	1.5860
	0	0.6375	0.6362	0.0006	0.0008	0.9007	0.8986	0.1140	0.1152	0.3252	0.3437	0.3065	0.3139
	0.02	0.4603	0.4588	0.0844	0.0841	1.2400	1.2380	0.1538	0.1528	0.3293	0.3458	0.3958	0.4384
0.7	0.05	0.3679	0.3665	0.1634	0.1616	1.5600	1.5530	0.1662	0.1655	0.3308	0.3394	0.5005	0.5050
	0.1	0.2397	0.2393	0.4010	0.4009	2.4310	2.4290	0.1705	0.1702	0.3435	0.3512	0.8567	0.8662
	0.25	0.1438	0.1437	0.9625	0.9618	4.0080	4.0040	0.1659	0.1658	0.4093	0.4166	1.5993	1.6037

As in the case T = 200, all the results for T = 100 and T = 50 are also represented with the boxplots to visualize the given conclusions above. Due to similarity of conclusions and maintaining the integrity of the text, complete boxplots for these two cases are provided in Appendix B.

4.2. Effects of Phase I Analysis on Control Chart Performance in Phase II

Remember that, there are two stages for performance computation in this research: (1) The control chart design with estimated parameters, (2) The computation of the ARL performance using the obtained design.

According to the results provided in the previous section, it was shown that filtering of outliers and use of incomplete datasets do not significantly affect the estimates of control chart parameters with the ML estimator, however, presence of outliers may have severe effects on the estimates of control chart parameters both with ML and CSS estimators.

After obtaining the estimation results, the effects of these parameter estimates on the performance of control charts in Phase II were investigated for the case (i) and (ii). The estimates for ϕ from each replication, which is given in Tables 3-8, were used to determine control limits through ARL_0 evaluations. Remember that a modified Shewhart type control chart is used and the expected ARL_0 performance is 370.4 for the AR(1) model with estimated parameter $\hat{\phi}$. ARL_0 results are obtained through the means of R's spc package [45].

In Tables 9 - 14, the results corresponding to the data lengths T = 200, 100 and 50 are provided for the ML and CSS estimators, respectively Since the simulations are repeated for 10,000 times, parameter estimates are obtained through the mean of 10,000 replications. Here, ARL is a function of the estimators and, hence, itself random. Therefore, mean ARL_0 and median ARL_0 were computed again from the results of 10,000 simulation replications. Note that even in the case of no outliers (r=0), the mean ARL0 deviates from the design value 370.4, which is simply an effect of estimated parameters [7].

Consider the results of T = 200 case, which are given in Table 9 and Table 10. In the presence of outliers, the L values, being determined through $\hat{\phi}$, are increased close to the upper bound 3 (corresponding to the design performance 370.4)[1]. Although these increases may look small at a first glance, they have a strong effect on the resulting *true* ARL_0 performance.

[1] In the i.i.d. case ($\phi = 0$), the ARL_0 value of 370.4 corresponds to the choice $\mathrm{L} = 3$. If $\phi \neq 0$, we have to choose $\mathrm{L} < 3$ to obtain $ARL_0 = 370.4$ [41, p. 119].

These ARL_0 estimates are influenced through all three parameter estimates, $\hat{\phi}, \hat{\mu},$ and $\hat{\sigma}^2$. From the mean ARL_0 performance evaluation, it can be observed that significantly higher ARL_0 values are obtained with increasing rate of outliers. Even for only 2 % outliers, the true ARL_0 is considerably greater than the design ARL_0 value of 370.4, which indicates the importance of filtering outliers (note the analogous conclusion by Weiß and Testik [4] in the discrete case). The filtering of outliers, in contrast, results in incomplete data, but as the rate r increases, mean ARL_0 values increase only very moderately, with a somehow stronger effect with the CSS estimator. Interestingly, with increasing rate r for filtering, the *median* ARL_0 values decrease, again with a stronger effect with the CSS estimator. Furthermore, median ARL_0 values are always smaller than the mean ARL_0 values. This positive skewness in the ARL_0 distributions increases with increasing ϕ and decreases with increasing T.

Note that, CSS estimator performs poor compared to the ML estimator when many data are filtered, for example r = 0.25. Therefore, as a recommendation for practitioners, it can be suggested to use ML estimators since CSS estimators may be misleading if many data have to be filtered due to outliers in the Phase I dataset.

Table 9. ARL_0 performance with the ML estimator in Phase I (T = 200).

		Case (i) – Outliers Filtered			**Case (ii) – With Outliers**		
ϕ	***r***	**CL**	**MeanARL$_0$**	**Median ARL$_0$**	**CL**	**Mean ARL$_0$**	**Median ARL$_0$**
0.3	0	2.995	395.7	336.4			
	0.02	2.995	396.0	335.3	2.997	2968	1553
	0.05	2.995	396.7	334.4	2.998	127100	9624
	0.1	2.995	398.7	334.7	2.999	26240000	98210
	0.25	2.995	404.8	328.5	2.999	25600000	1110000
0.5	0	2.979	385.4	311.5			
	0.02	2.979	385.8	312.2	2.990	2887	1406
	0.05	2.979	387.2	310.7	2.995	153300	9012
	0.1	2.979	387.0	310.4	2.997	25680000	106700
	0.25	2.979	392.8	308.9	2.999	28320000	1089000
0.7	0	2.930	387.2	285.1			
	0.02	2.930	387.3	284.8	2.973	2996	1333
	0.05	2.930	388.0	283.5	2.987	169700	8562
	0.1	2.930	388.9	284.9	2.994	41740000	96600
	0.25	2.930	392.8	280.9	2.998	36260000	1059000

Table 10. ARL_0 performance with the CSS estimator in Phase I (T = 200).

		Case (i) – Outliers Filtered			Case (ii) – With Outliers		
ϕ	***r***	**CL**	**MeanARL$_0$**	**Median ARL$_0$**	**CL**	**Mean ARL$_0$**	**Median ARL$_0$**
0.3	0	2.995	395.6	336.5			
	0.02	2.995	397.0	333.8	2.997	2976	1545
	0.05	2.995	398.4	331.4	2.998	121100	9725
	0.1	2.995	402.0	329.1	2.999	24980000	97570
	0.25	2.994	429.1	315.3	2.999	25170000	1104000
0.5	0	2.979	385.8	311.0			
	0.02	2.979	387.9	309.5	2.990	2898	1398
	0.05	2.979	391.6	308.1	2.995	159600	8970
	0.1	2.979	395.1	302.0	2.997	28820000	106100
	0.25	2.978	432.5	287.5	2.999	28780000	1082000
0.7	0	2.930	387.1	284.2			
	0.02	2.930	391.6	282.3	2.973	2996	1317
	0.05	2.930	395.2	277.1	2.987	169500	8618
	0.1	2.930	410.0	275.2	2.994	41510000	97650
	0.25	2.928	546.1	255.8	2.998	36530000	1058000

The only differences between the case T = 200, and cases T = 100 and T = 50 is that T = 200 case has more precise results since the parameter estimates are better for larger data lengths, as expected (see Tables 11-14). Remember that T = 200 estimates are more accurate in terms of mean, and have smaller standard deviations. ARL results show the same pattern that estimates have. Therefore, T = 200 have smaller and better ARL values compared to the smaller data lengths T = 100 and T = 50. Note that, CSS estimator performs poor when many data need to be filtered. The only difference, this performing effect is more clear for the case T = 50, since the remaining data quantity after filtering is very small.

Table 11. ARL_0 performance with the ML estimator in Phase I (T = 100).

		Case (i) – Outliers Filtered			Case (ii) – With Outliers		
ϕ	***r***	**CL**	**MeanARL$_0$**	**Median ARL$_0$**	**CL**	**Mean ARL$_0$**	**Median ARL$_0$**
0.3	0	2.995	414.4	297.5			
	0.02	2.995	415.9	297.0	2.997	3432	1320
	0.05	2.995	417.4	293.8	2.998	197800	8690
	0.1	2.995	421.9	294.4	2.999	45470000	93380
	0.25	2.994	436.5	287.3	2.999	30740000	965300
0.5	0	2.980	420.9	267.9			
	0.02	2.980	421.8	268.1	2.99	3604	1219
	0.05	2.980	420.4	267.4	2.995	311800	8132
	0.1	2.979	424.7	266.3	2.997	28830000	81060
	0.25	2.979	437.5	260.3	2.999	36880000	950200
0.7	0	2.933	434.5	225.6			
	0.02	2.933	435.2	225.3	2.974	4754	1028
	0.05	2.933	436.7	225.2	2.988	408300	7173
	0.1	2.933	436.9	225.0	2.994	43960000	85500
	0.25	2.933	450.7	221.5	2.997	56110000	930800

Table 12. ARL_0 performance with the CSS estimator in Phase I (T = 100).

		Case (i) – Outliers Filtered			Case (ii) – With Outliers		
ϕ	r	CL	MeanARL₀	Median ARL₀	CL	Mean ARL₀	Median ARL₀
0.3	0	2.995	415	296.3			
	0.02	2.995	419.2	294.8	2.997	3482	1304
	0.05	2.995	423.9	289	2.998	215200	8675
	0.1	2.994	435.6	286	2.999	42430000	93950
	0.25	2.994	484.9	260.8	2.999	30730000	979200
0.5	0	2.98	425.8	268.7			
	0.02	2.979	435.3	264.7	2.99	3664	1217
	0.05	2.979	439.6	260.5	2.995	339100	8120
	0.1	2.979	459.8	254.8	2.997	30890000	80770
	0.25	2.978	613.1	224.8	2.999	36070000	957800
0.7	0	2.933	426	223.9			
	0.02	2.933	439.3	222.6	2.974	4648	1016
	0.05	2.933	463.9	217.8	2.988	433500	7020
	0.1	2.932	495	211.7	2.994	44890000	83920
	0.25	2.928	3276	185.3	2.997	58610000	923900

Table 13. ARL_0 performance with the ML estimator in Phase I (T = 50)

		Case (i) – Outliers Filtered			Case (ii) – With Outliers		
ϕ	r	CL	MeanARL₀	Median ARL₀	CL	Mean ARL₀	Median ARL₀
0.3	0	2.994	513.5	249.2			
	0.02	2.994	516.7	247.2	2.997	5880	1141
	0.05	2.994	525.6	246.6	2.997	245800	4322
	0.1	2.994	547.1	238.2	2.998	83830000	82030
	0.25	2.993	570.4	229.0	2.999	51010000	782900
0.5	0	2.98	522.3	210.7			
	0.02	2.98	524.6	209.9	2.99	7011	928.8
	0.05	2.98	524.0	208.8	2.994	122300	3478
	0.1	2.98	543.3	205.2	2.997	82730000	70260
	0.25	2.979	601.5	199.8	2.998	72480000	821600
0.7	0	2.938	621.8	152.0			
	0.02	2.938	624.3	151.8	2.975	19030	678.4
	0.05	2.938	627.2	153.2	2.985	230600	2463
	0.1	2.938	678.8	150.8	2.994	136600000	51830
	0.25	2.939	689.3	146.0	2.997	158200000	667700

Table 14. ARL_0 performance with the CSS estimator in Phase I (T = 50).

		Case (i) – Outliers Filtered			Case (ii) – With Outliers		
ϕ	r	CL	MeanARL₀	Median ARL₀	CL	Mean ARL₀	Median ARL₀
0.3	0	2.994	519.8	245.5			
	0.02	2.994	524.4	241.3	2.997	6291	1123
	0.05	2.994	569.1	238.5	2.997	295800	4160
	0.1	2.994	607.3	224.8	2.998	70850000	80500
	0.25	2.992	845	188.8	2.999	51100000	786700
0.5	0	2.98	523.5	205.9			
	0.02	2.98	541.8	201.9	2.99	7593	899.3
	0.05	2.98	562.6	197.4	2.994	157300	3388
	0.1	2.979	753.2	187.2	2.997	77810000	68040
	0.25	2.976	141300	158	2.998	76210000	787900
0.7	0	2.939	688.6	144.5			
	0.02	2.938	824.4	142.5	2.976	17800	647.1
	0.05	2.937	10010	140.8	2.986	236400	2310
	0.1	2.936	2212	129.4	2.994	157700000	50890
	0.25	2.929	264500	106.8	2.997	167000000	644700

5. ILLUSTRATIVE EXAMPLE

The assay data from Table 3 in Mukundam et. al. [46] constitutes a time series of length $T = 53$, which exhibits an AR(1)-like autocorrelation structure. Please see Table 15 for the data provided in Table 3 in Mukundam et. al. [46].

Table 15. The assay data

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
100.8	100.8	101.5	101.5	101.5	101.3	100.9	100.6	100.6	100.4	100.8	100.1	100.7	100.4
15	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
100.4	100.3	98.9	100.3	100.3	100.3	100.3	100.2	100.8	100.8	99.5	100.7	100.4	100.4
29	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**
100.8	100.7	100.7	100.4	100	100.3	100.7	100.7	100.7	100.6	100	100.7	100.8	100.3
43	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**			
100.3	100.9	100.5	100.5	101.2	100.6	100.7	100.6	101.1	101	101.1			

Fitting an AR(1) model to the data via ML estimation, the initial estimates $\hat{\phi} = 0.387$, $\hat{\mu} = 100.60$, and $\hat{\sigma}^2 = 0.202$ are obtained. Remember the equations given in Section 2.2.2. to calculate control limits for Shewhart Type Control Charts [3]:

$$UCL = \mu_w + L\sigma_w$$
$$CL = \mu_w$$
$$LCL = \mu_w - L\sigma_w$$

Based on these estimates and aiming at $ARL_0 = 100$ (which seems a reasonable level in view of $T = 53$), we compute $L = 2.557$ and, hence, the upper and lower control limits $LCL = 99.45$ and $UCL = 101.75$, respectively:

$$UCL = 100.60 + 2.557 \,.\, \sqrt{0.202}$$
$$CL = 100.60$$
$$LCL = 100.60 - 2.557 \,.\, \sqrt{0.202}$$

Plotting the data on this chart, a signal is triggered at time $t = 17$ (see Figure 8) corresponding observation $x_{17} = 98.9$ is much below the LCL. Hence, this is treated as an outlier and excluded from the data set (i.e., we replace the observation 98.9 at time 17 by the value “NA”).

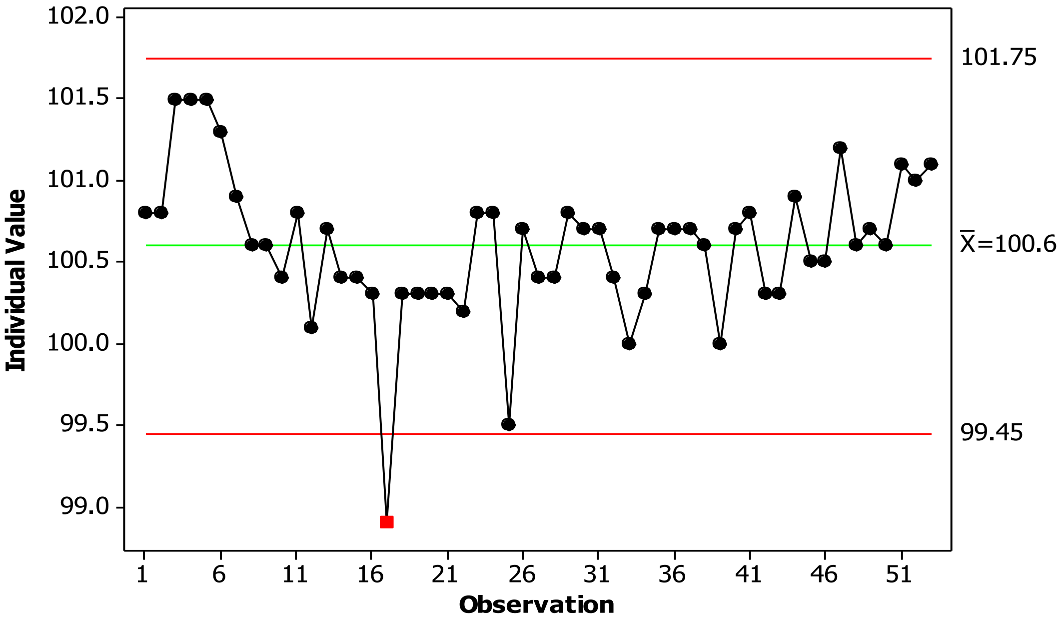


Figure 8. Shewhart type control chart for the assay data

Next, we use the filtered data set and obtain the revised estimates as $\hat{\phi} = 0.410$, $\hat{\mu} = 100.63$, and $\hat{\sigma}^2 = 0.150$. The strong deviation between the original and the revised estimates $\hat{\phi}$ and $\hat{\sigma}^2$ illustrates the effect of the outlier at time $t = 17$. Then the revised chart design has L = 2.554, CL = 100.63, LCL = 99.64 and UCL = 101.62. The revised chart has narrower limits, and as a result, a new alarm is triggered, now for time $t = 25$ where $x_{25} = 99.5 <$ LCL (see Figure 9). We exclude this observation similarly and continue with the third iteration of our Phase I analysis.

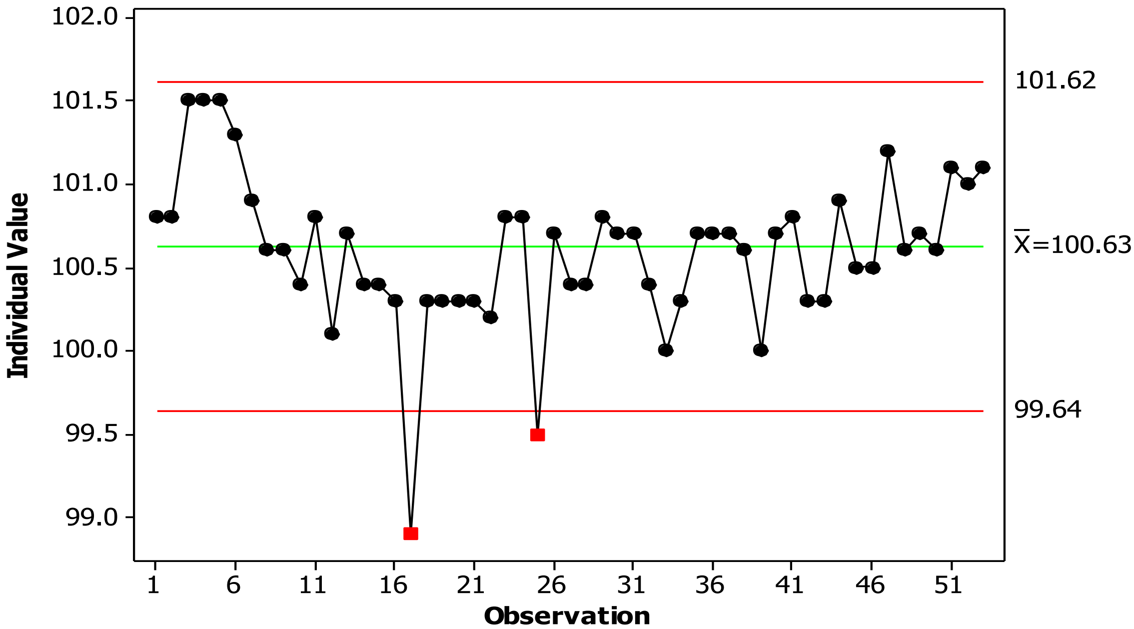


Figure 9. Revised Shewhart type control chart for assay data (x_{17} was filtered)

Using the data after filtering the outliers at times $t = 17$ and $t = 25$ (i.e., with "NA" at times 17 and 25), revised estimates are now $\hat{\phi} = 0.535$, $\hat{\mu} = 100.66$, and $\hat{\sigma}^2 = 0.127$ (note the strong increase in $\hat{\phi}$), where the revised chart design has $L = 2.5239$, $CL = 100.66$, $LCL = 99.76$ and $UCL = 101.56$. Now all points are plotted between the control limits (see Figure 10), i.e., the Phase I analysis stops and gives the AR(1) model with the above parameter values as the in-control model.

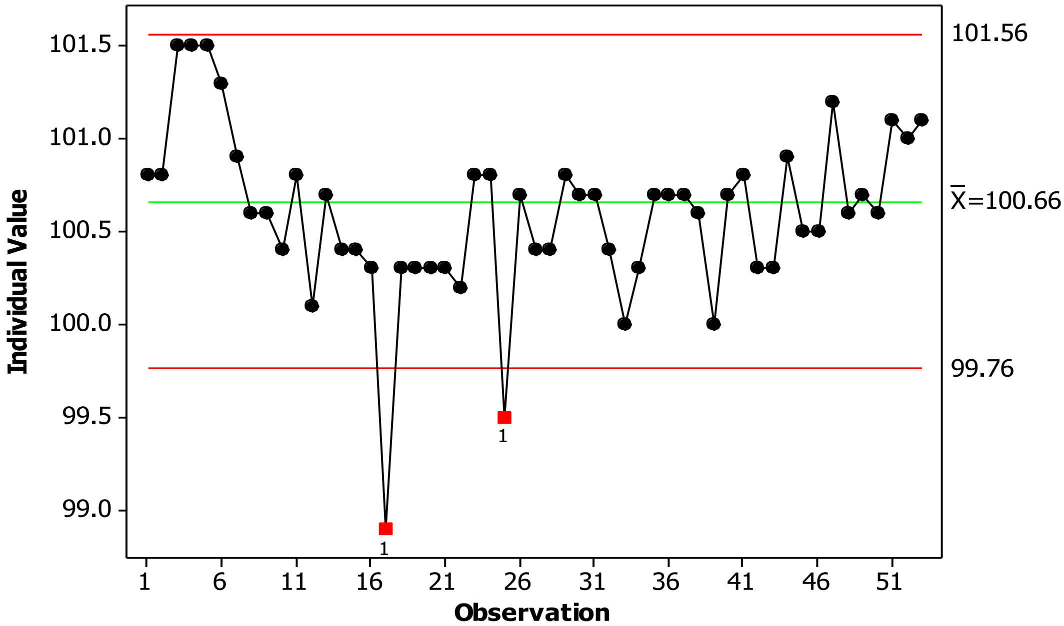


Figure 10. Revised Shewhart type control chart for assay data (x_{17} and x_{25} were filtered)

The whole Phase I process is summarized in Table 16. In fact, comparing with the analysis in Mukundam et al. [46], the observations $t = 17$ and $t = 25$ were also declared as outliers there, although the authors worked with a different monitoring approach and ignored the apparent autocorrelation in the data. It supports our observation from Figure 1 that especially the estimation of autocorrelation parameter and variance is severely affected by even only few outliers."

Table 16. Summary of Phase I analysis for assay data.

Iteration	**Identified outliers**	$\hat{\phi}$	$\hat{\mu}$	$\hat{\sigma}^2$	**LCL**	**UCL**	**New outliers**
1	–	0.387	100.60	0.202	99.45	101.75	$t = 17$
2	$t = 17$	0.410	100.63	0.150	99.64	101.62	$t = 25$
3	$t = 17, 25$	0.535	100.66	0.127	99.76	101.56	–

6. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this thesis, as an alternative but more realistic approach to study the effects of Phase I analysis on the performance of control charts in Phase II, availability of contaminated or, clean but incomplete datasets are considered for parameter estimation. It was shown that filtering of outliers and use of incomplete data sets do not significantly affect the estimates of control chart parameters when the ML estimators are used. On the other hand, presence of outliers in the Phase I data set may have severe effects on the estimates of control chart parameters both with ML and CSS estimators. Since the main intention of a control chart is online process monitoring, study of control chart designs from Phase I on the ARL performance in Phase II indicate that, ARL_0 performances are influenced by the parameter estimates $\hat{\phi}, \hat{\mu},$ and $\hat{\sigma}^2$, used with the AR(1) process model. It is observed that significantly higher ARL_0 values are obtained with increasing rate of outliers. On the other hand, filtering of outliers results in incomplete data, but as the rate increases, mean ARL_0 values increase only very moderately, with a more robust performance with the ML estimator. As a recommendation for practitioners, it is suggested to use ML estimators since CSS estimators may be misleading if many data have to be filtered due to outliers in the Phase I data set.

For future research, a more refined study of the Phase I process by also considering the efficiency in detecting outliers may be considered. In this context, also the use of robust estimators seems to be an interesting option. There are enumerable proposals available in the literature on how to robustly estimate the AR(1) parameters, see the survey by Dürre et al. [47]. While many of these proposals are computationally demanding, a relatively quick and intuitive way of robust parameter estimation is the approach of Ma and Genton [48], which is somehow similar to the method of moments. There, autocorrelations (and hence ϕ) are estimated based on the Q_n estimator, the latter being computed from absolute pairwise differences. To get an idea about the performance of such a robust approach, the above simulations in this thesis were extended and Q_n was used to estimate ϕ and σ^2 [48, pp. 665-666], while μ was estimated by using the median. Some results are summarized in Table 17. Comparing with the corresponding ML estimates in Table 4, it can be seen that the robust estimates have higher standard deviation if there are no outliers, but they show less bias and standard deviation if the data are contaminated by outliers. Note that even only 2 % outliers have a visible effect on the robust estimates (although the effect is less worse than for the ML estimates). An analogous observation is made if the ARLs are compared to the ones in Table 9; while for clean data, the ML approach is preferable, the robust approach gives more satisfactory results as long as data are contaminated. A more detailed analysis and a strategy of when and how to use robust estimators will be part of a future research project.

Table 17. Means and standard deviations of robust parameter estimates (10,000 replications with $T = 200$ and $\phi = 0.5$, with outliers), resulting chart design and ARL performance.

	Mean of			**Standard Deviation of**			**Mean of**	
$\boldsymbol{r}$	$\hat{\boldsymbol{\mu}}_{\text{rob}}$	$\hat{\boldsymbol{\phi}}_{\text{rob}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{rob}}$	$\hat{\boldsymbol{\mu}}_{\text{rob}}$	$\hat{\boldsymbol{\phi}}_{\text{rob}}$	$\hat{\boldsymbol{\sigma}}^2_{\text{rob}}$	**CL**	**$\mathbf{ARL_0}$**
0	-0.0004	0.4859	0.9984	0.1342	0.0684	0.1420	2.979	427
0.02	0.0247	0.4698	1.1190	0.1346	0.0696	0.1627	2.981	811
0.05	0.0651	0.4446	1.3220	0.1359	0.0729	0.2093	2.984	2544
0.1	0.1382	0.4011	1.7050	0.1377	0.0794	0.3394	2.988	34740
0.25	0.4233	0.2783	3.1450	0.1465	0.1035	1.1245	2.995	836300000

REFERENCES

[1] W. H. Woodall and D. C. Montgomery, "Research issues and ideas in Statistical Process Control," *Journal of Quality Technology,* vol. 31, no. 4, pp. 376-387, **1999**.

[2] A. Jones-Farmer, W. H. Woodall, S. H. Steiner and C. W. Champ, "An overview of Phase I analysis for process improvement and monitoring," *Journal of Quality Technology,* vol. 46, no. 3, pp. 265-280, **2014**.

[3] D. C. Montgomery, Statistical Quality Control: A Modern Introduction, 7th Edition International Student Version ed., New York: John Wiley & Sons Singapore Pte. Ltd., **2013**.

[4] C. H. Weiß and M. C. Testik, "On the Phase I Analysis for Monitoring Time-Dependent Count Processes," *IIE Transactions*, vol. 47, no. 3, pp. 294-306, **2015**.

[5] S. Chakraborti, S. W. Human and M. A. Graham, "Phase I statistical process control charts: an overview and some results," *Quality Engineering,* vol. 21, no. 1, pp. 52-62, **2008**.

[6] L. A. Jones, "The statistical design of EWMA control charts with estimated parameters," *Journal of Quality Technology,* vol. 34, no. 3, pp. 277-288, **2002**.

[7] W. A. Jensen, L. A. Jones-Farmer, C. W. Champ and W. H. Woodall, "Effects of parameter estimation on control chart properties: a literature review," *Journal of Quality Technology,* vol. 38, no. 4, pp. 349-364, **2006**.

[8] W. H. Woodall and D. C. Montgomery, "Some current directions in the theory and application of statistical process monitoring," *Journal of Quality Technology,* vol. 46, no. 1, pp. 78-94, **2014**.

[9] M. C. Testik, "Model inadequacy and residuals control charts for autocorrelated processes," *Quality and Reliability Engineering International,* vol. 21, no. 2, pp. 115-130, **2005**.

[10] S. Knoth and W. Schmid, "Control charts for time series: A review," in *Frontiers in Statistical Quality Control*, vol. 7, H. Lenz and P. Wilrich, Eds., Heidelberg, Germany, Physica-Verlag, pp. 210-236, **2004**

[11] W. H. Woodall, "Controversies and contradictions in statistical process control," *Journal of Quality Technology,* vol. 32, no. 4, pp. 341-350, **2000**.

[12] W. A. Shewhart, "Quality Control Charts," *Bell System Technical Journal,,* vol. 5, no. 4, pp. 593-603, **1926**.

[13] W. E. Company, Statistical Quality Control Handbook, Indianapolis, Indiana: Western Electric Co, **1956**.

[14] C. W. Champ and W. H. Woodall, "Exact results for Shewhart control charts with supplementary runs rules," *Technometrics,* vol. 29, no. 4, pp. 393-399, **1987**.

[15] D. M. Hawkins and D. H. Olwell, Cumulative sum charts and charting for quality improvement., Springer Science & Business Media, **1998**.

[16] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: properties and enhancements," *Technometrics,* vol. 32, no. 1, pp. 1-12, **1990**.

[17] W. H. Woodall, D. J. Spitzner, D. C. Montgomery and S. Gupta, "Using control charts to monitor process and product quality profiles," *Journal of Quality Technology,* vol. 36, no. 3, pp. 309-320, **2004**.

[18] G. Vining, "Technical Advice: Phase I and phase II control charts," *Quality Engineering,* vol. 21, no. 4, pp. 478-479, **2009**.

[19] E. Dasdemir, C. Weiss, M. Testik and S. Knoth, "Evaluation of Phase I Analysis Scenarios on Phase II Performance of Control Charts For Autocorrelated Observations," *submitted to Quality Engineering,* **2015**

[20] W. H. Woodall, "The statistical design of quality control charts," *The Statistician* , pp. 155-160, **1985**.

[21] C. W. Champ and L. A. Jones, "Designing Phase I—X Charts with Small Sample Sizes," *Quality and Reliability Engineering International,* vol. 20, no. 5, pp. 497-510, **2004**.

[22] B. M. Adams and I.-T. Tseng, "Robustness of forecast-based monitoring schemes," *Journal of Quality Technology,* vol. 30, no. 4, pp. 328-39, **1998**.

[23] R. A. Boyles, "Phase I Analysis for Autocorrelated Processes," *Journal of Quality Technology,* vol. 32, no. 4, pp. 395-409, **2000**.

[24] J. Ledolter, "The effect of additive outliers on the forecasts from ARMA models," *International Journal of Forecasting,* vol. 5, no. 2, pp. 231-240, **1989**.

[25] H. D. Maragah and W. H. Woodall, "The effect of autocorrelation on the retrospective X-chart," *Journal of Statistical Computation and Simulation,* vol. 40, no. 1-2, pp. 29-42, **1992**.

[26] L. C. Alwan and H. V. Roberts, "Time-series modeling for statistical process control," *Journal of Business & Economic Statistics,* vol. 6, no. 1, p. 87–95, **1988**.

[27] D. Montgomery and C. Mastrangelo, "Some statistical process control methods for autocorrelated data," *Journal of Quality Technology,* vol. 23, pp. 179-193, **1991**.

[28] L. C. Alwan, "Effects of autocorrelation on control chart performance," *Communications in Statistics - Theory and Methods,* vol. 21, no. 4, pp. 1025-1049, **1992**.

[29] Harris, Thomas J. and W. H. Ross, "Statistical process control procedures for correlated observations," *The Canadian Journal of Chemical Engineering,* vol. 69, no. 1, pp. 48-57, **1991**.

[30] G. C. Runger and T. R. Willemain. , "Model-based and model-free control of autocorrelated processes," *Journal of Quality Technology,* vol. 27, no. 4, pp. 283-292, **1995**.

[31] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, Time Series Analysis: Forecasting and Control, 4 ed., Hoboken, NJ: John Wiley & Sons, Inc., **2008**.

[32] W. Schmid, "On the run length of a Shewhart chart for correlated data," *Statistical Papers,* vol. 36, no. 1, pp. 111-130, **1995**.

[33] H. Kramer and W. Schmid, "The influence of parameter estimation on the ARL of Shewhart type charts for time series," *Statistical Papers,* vol. 41, no. 2, pp. 173-196, **2000**.

[34] G. Runger , T. Willemain and S. Prabhu, "Average run lengths for Cusum control charts applied to residuals," *Communications in Statistics—Theory and Methods,* vol. 24, pp. 273-282, **1995**.

[35] S. Vander Wiel, "Monitoring processes that wander using integrated moving average models," *Technometrics,* vol. 38, pp. 139-151, **1996**.

[36] D. W. Apley and J. Shi, "The GLRT for statistical process control of autocorrelated processes," *IIE Transactions,* vol. 31, no. 12, pp. 1123-1134, **1999**.

[37] G. C. Runger, "Assignable causes and autocorrelation: Control charts for observations or residuals?," *Journal of Quality Technology,* vol. 34, no. 2, pp. 165-170, **2002**.

[38] C.-W. Lu and M. R. Reynolds, "Control charts for monitoring the mean and variance of autocorrelated processes," *Journal of Quality Technology,* vol. 31, no. 3, pp. 259-274, **1999**.

[39] H. Kramer and W. Schmid, "Control charts for time series," *Nonlinear Analysis: Theory, Methods & Applications,* vol. 30, no. 7, pp. 4007- 40016, **1997**.

[40] W. Jiang, "Multivariate control charts for monitoring autocorrelated processes AA," *Journal of Quality Technology,* vol. 36, no. 4, pp. 367-379, **2004**.

[41] W. Schmid, "On the run length of a Shewhart chart for correlated data," *Statistical Papers,* vol. 36, no. 1, pp. 111-130, **1995**.

[42] R. H. Jones, "Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations," *Technometrics,* vol. 22, no. 3, pp. 389-395, **1980**.

[43] R Core Team, "R: A Language and Environment for Statistical," R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, **2014**.

[44] A. Fox, "Outliers in Time Series," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 34, no. 3, pp. 350-363, **1972**.

[45] S. Knoth, "spc: Statistical Process Control – Collection of Some Useful Functions," R Foundation for Statistical Computing, http://cran.r-project.org/web/packages/spc/index.html, **2014**.

[46] K. Mukundam, D. R. Varma, G. R. Deshpande, V. Dahanukar and A. K. Roy, "I-MR Control Chart: A Tool for Judging the Health of the Current Manufacturing Process of an API and for Setting the Trial Control Limits in Phase I of the Process Improvement," *Organic Process Research & Development*, vol. 17, pp. 1002-1009, **2013**.

[47] A. Dürre, R. Fried and T. Liboschik, "Robust estimation of (partial) autocorrelation," SFB 823 Discussion Paper 12/2014, **2014**.

[48] Y. Ma and M. G. Genton, "Highly robust estimation of the autocovariance function," *Journal of Time Series Analysis*, vol. 21, no. 6, pp. 663-684, **2000**.

[49] S. V. Crowder, "A simple method for studying run-length distributions of exponentially weighted moving average charts," *Technometrics,* vol. 29, no. 4, pp. 401-407, **1987**.

[50] G. C. Runger, T. R. Willemain and S. Prabhu, "Average run lengths for Cusum control charts applied to residuals," *Communications in Statistics—Theory and Methods*, vol. 24, no. 1, pp. 273-282, **1995**.

[51] S. V. Wiel, "Monitoring processes that wander using integrated moving average models," *Technometrics,* vol. 38, no. 2, pp. 139-151, **1996**.

[52] P. J. Brockwell and R. A. Davis, Introduction to Time Series and Forecasting, New York: Springer, **1996**.

APPENDICES

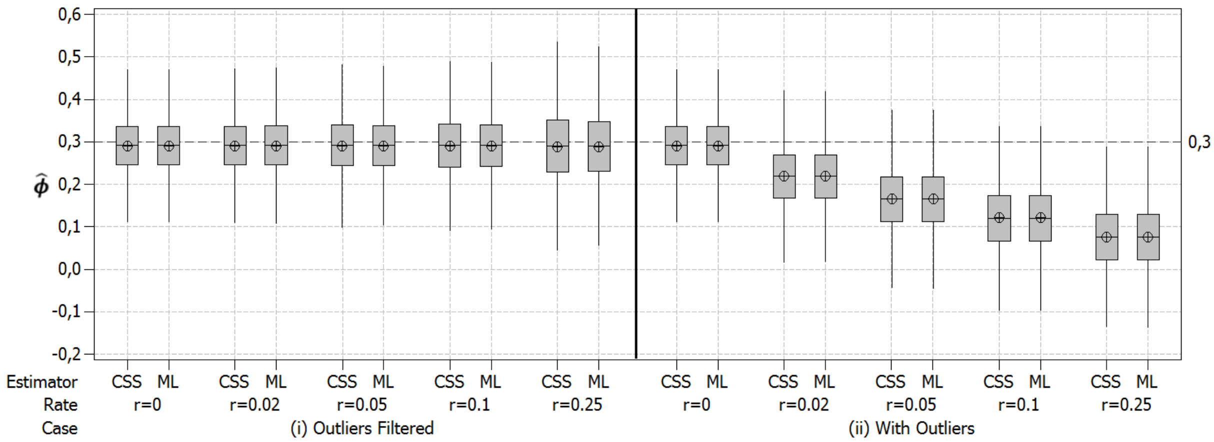
Appendix A - Numerical Calculations of the ARL in the AR(1) Case

To calculate the ARL of the modified AR(1) Shewhart chart, we essentially make use of ideas published in Schmid [41]. Introducing $Z_t = (1 - \phi)X_t$, which turns out to be an EWMA smoothing of the AR(1) residuals, allows to apply numerical routines for computing the ARL values of EWMA charts in the i.i.d. case. For ARL_0, we can directly use the function xewma.arl in the R package spc [45] with smoothing constant $\lambda = 1 - \phi$ and head start value $Z_0 = 0$. To evaluate ARL_1, we have to deal with the special way on how the shift in the mean of X_t, say δ, is mirrored from the residuals. As already mentioned by Schmid [41], we have to combine a random head start with mean δ and the ARL_1 function provided by, e.g., xewma.arl with shift size $(1 - \phi)\delta$. The related integral is approximated by applying the Gauß-Legendre quadrature with 30 nodes. The latter is used as well within the Nyström method implemented in the R function xewma.arl for numerically solving the ARL integral equation of a common i.i.d. EWMA chart following Crowder [49].

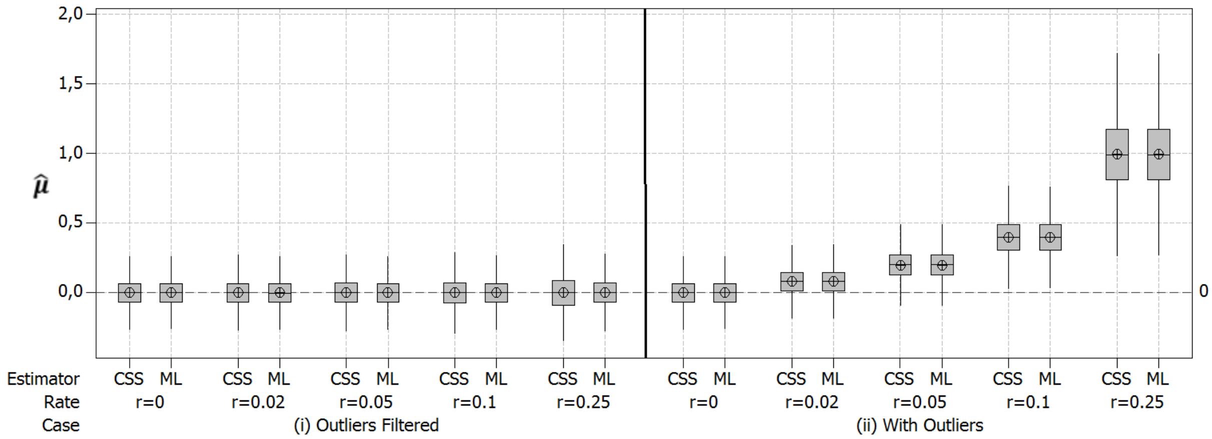
Finally, note that for negative ϕ, the EWMA smoothing constant λ is larger than 1, which is rather remarkable, because the classical upper bound 1 corresponds to the memory-less Shewhart chart. Values for λ larger than 1 would diminish this non-existing memory even further.

Appendix B – Remaining Boxplots for the Parameter Estimates

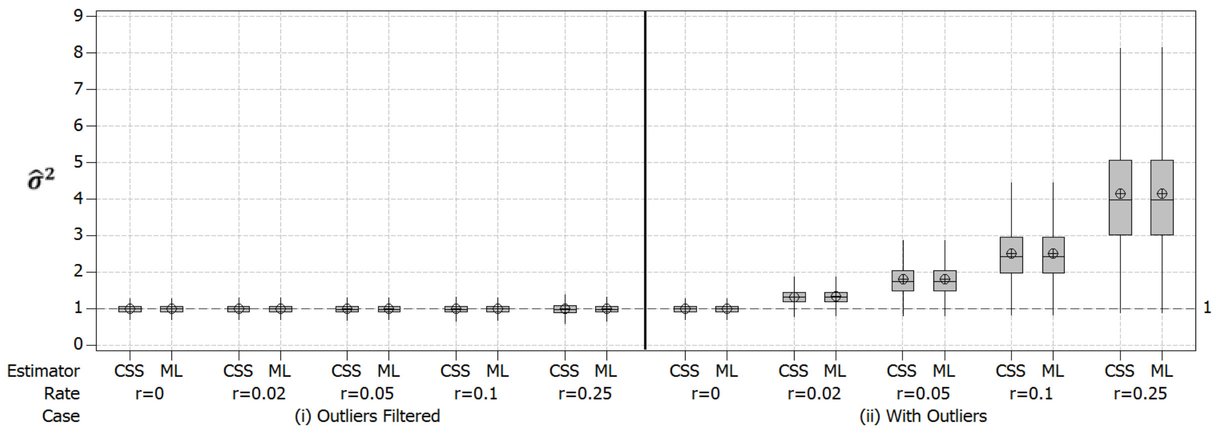
I. T=200



a. Box plot of estimates for the autocorrelation parameter ϕ.

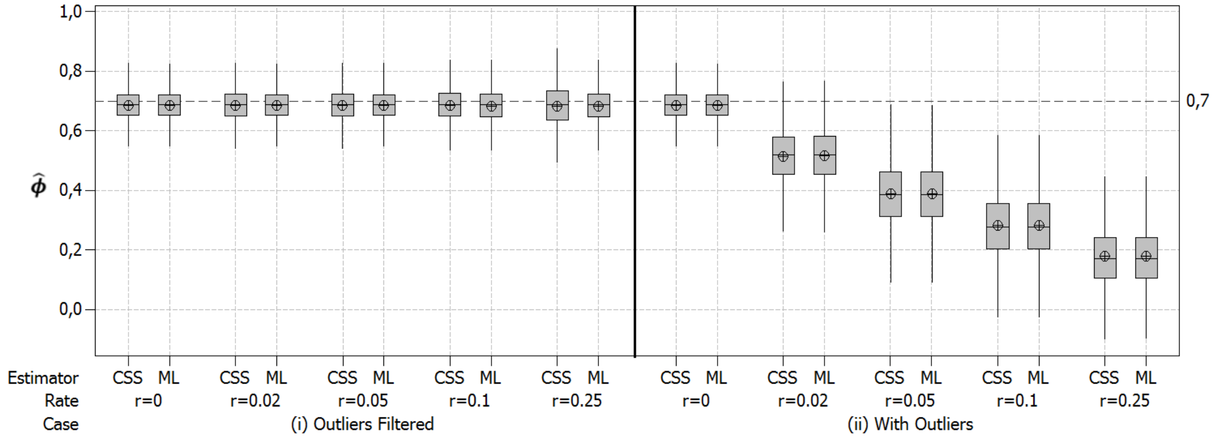


b. Box plot of estimates for the mean μ.

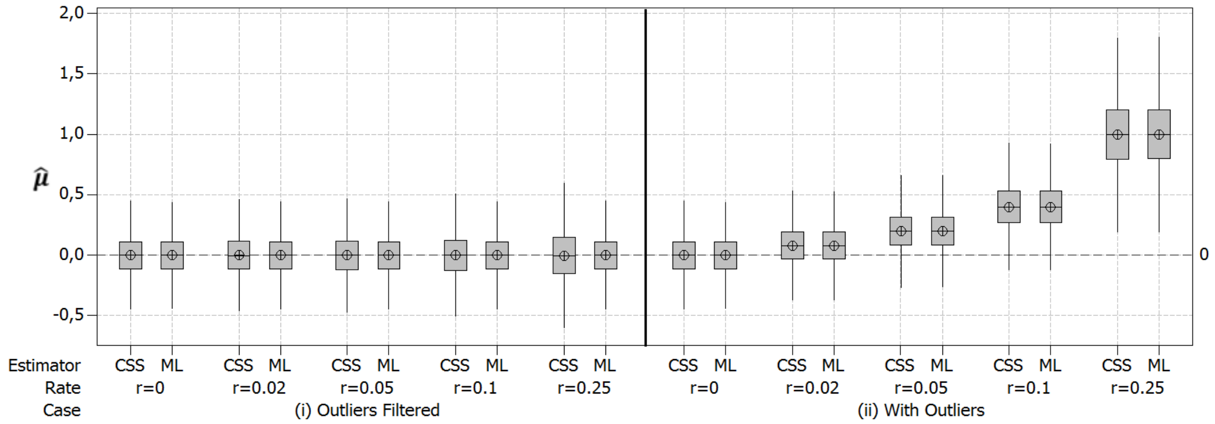


c. Box plot of estimates for the variance σ^2.

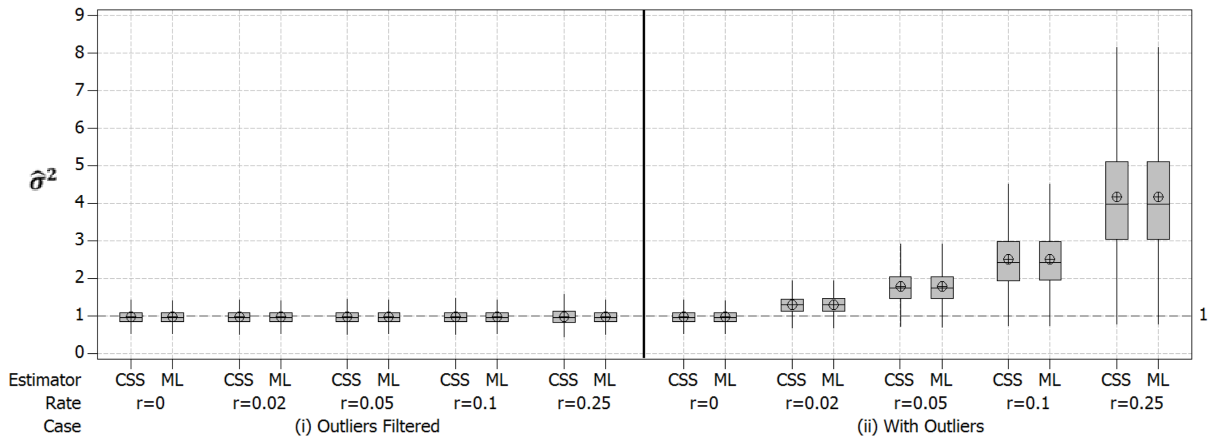
Figure 11. Box plots of the parameter estimates from simulations – $T = 200, \phi = 0.3$.



a. Box plot of estimates for the autocorrelation parameter ϕ.



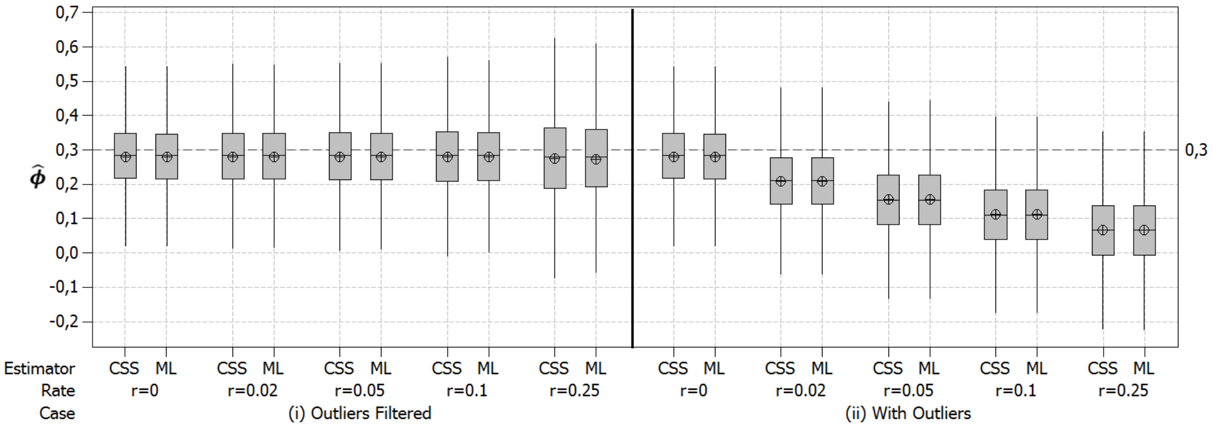
b. Box plot of estimates for the mean μ.



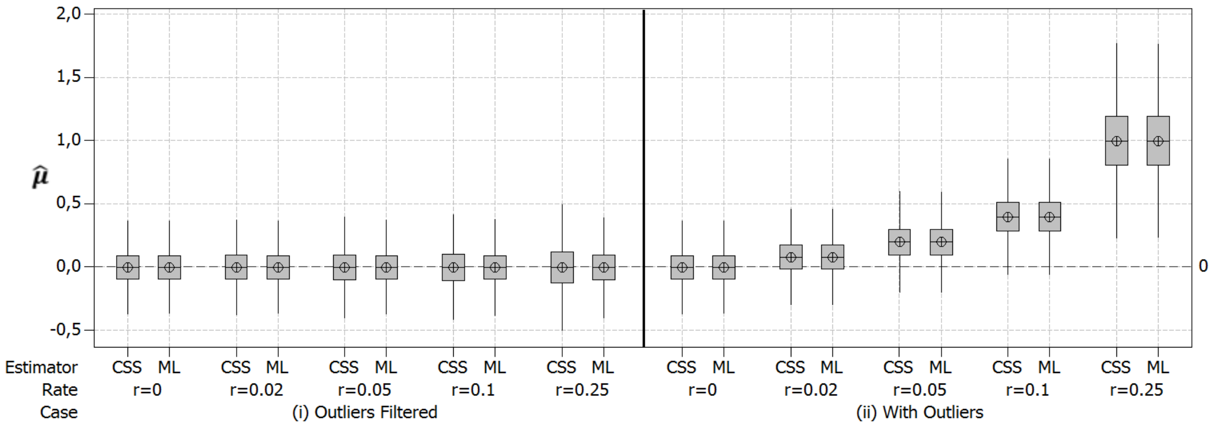
c. Box plot of estimates for the variance σ^2.

Figure 12. Box plots of the parameter estimates from simulations – $T = 200, \phi = 0.7$.

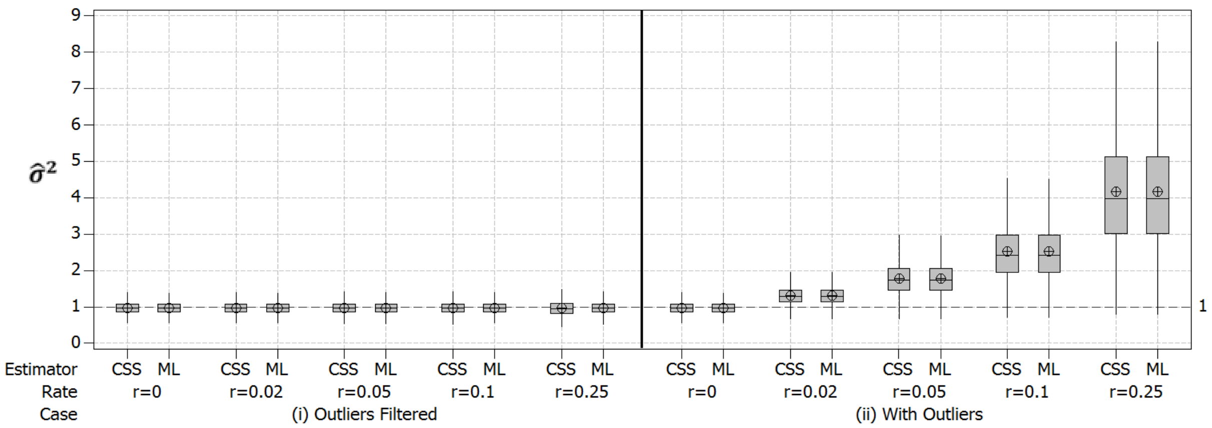
II. T=100



a. Box plot of estimates for the autocorrelation parameter ϕ.

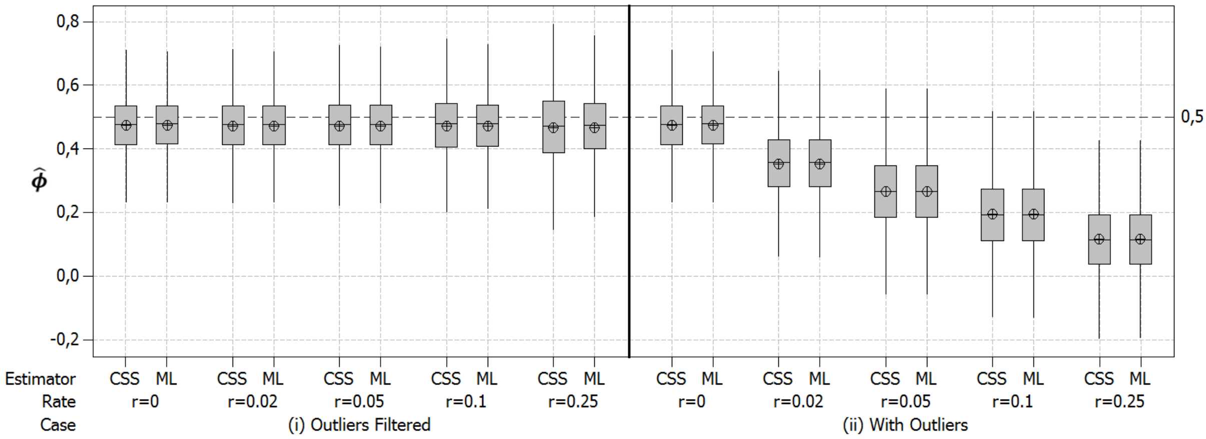


b. Box plot of estimates for the mean μ.

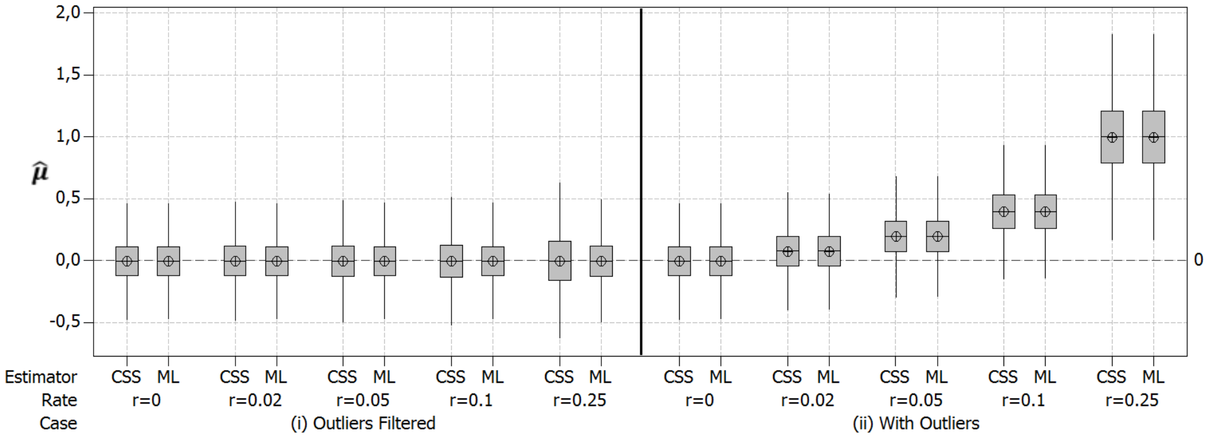


c. Box plot of estimates for the variance σ^2.

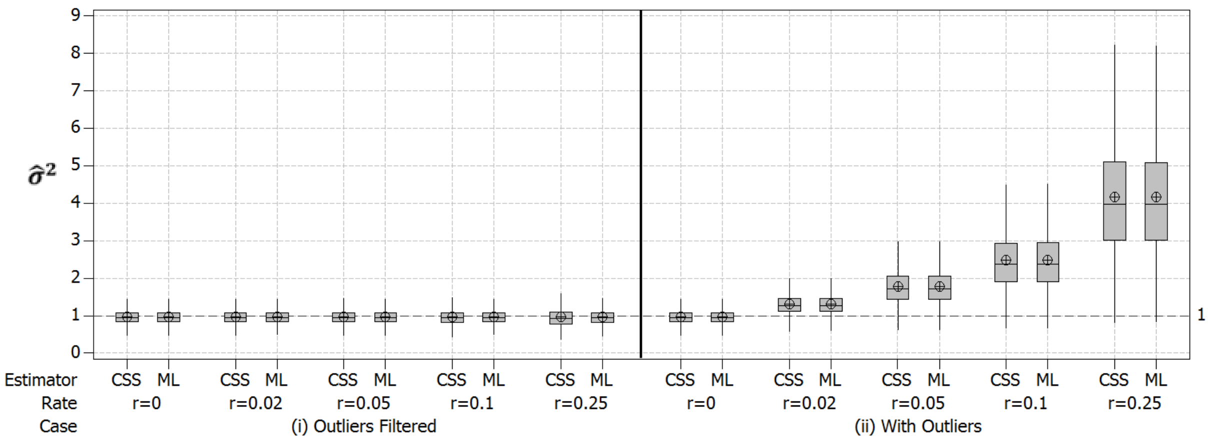
Figure 13. Box plots of the parameter estimates from simulations – $T = 100, \phi = 0.3$.



a. Box plot of estimates for the autocorrelation parameter ϕ.

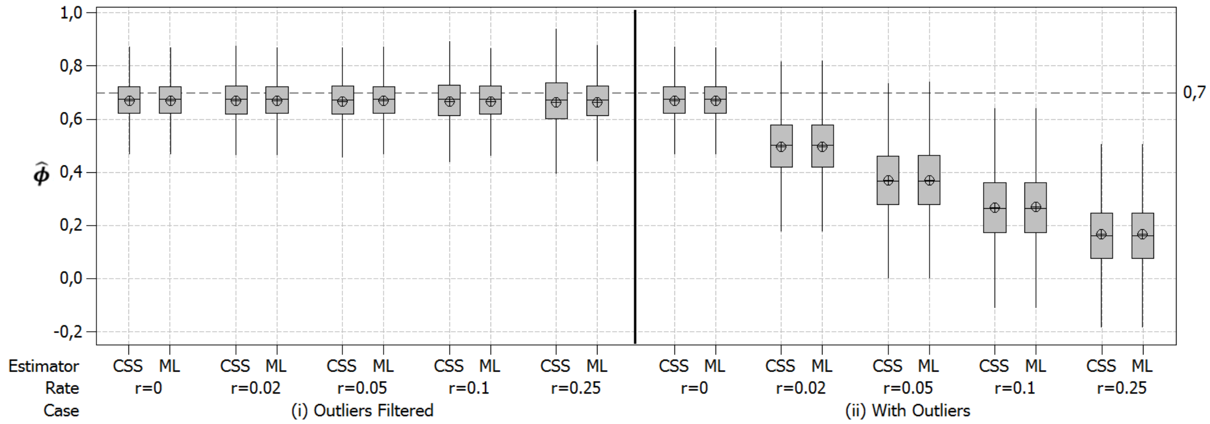


b. Box plot of estimates for the mean μ.

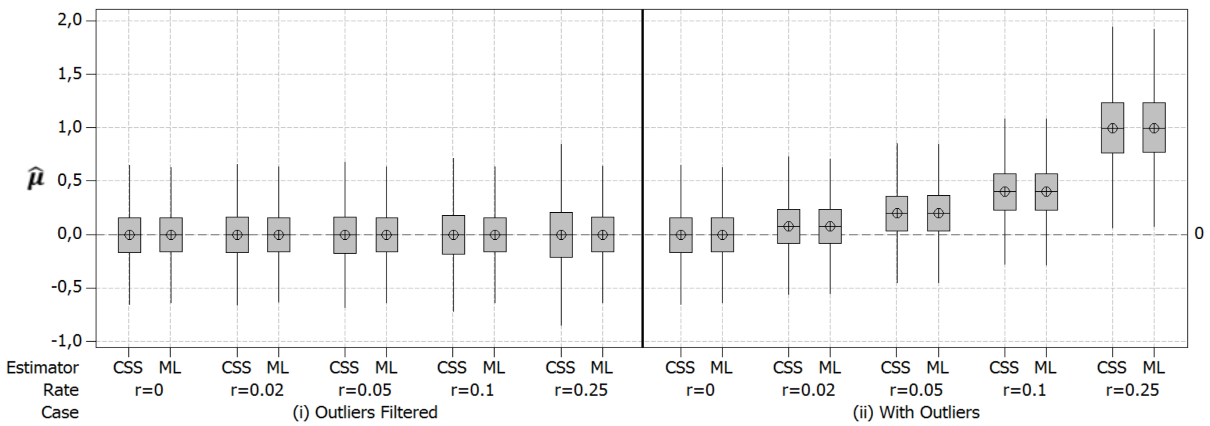


c. Box plot of estimates for the variance σ^2.

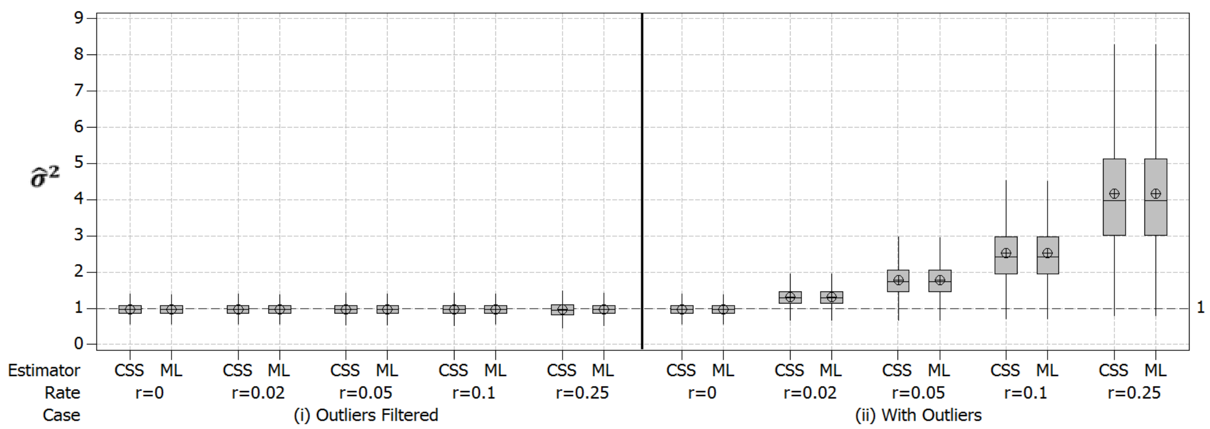
Figure 14. Box plots of the parameter estimates from simulations – $T = 100, \phi = 0.5$.



a. Box plot of estimates for the autocorrelation parameter ϕ.



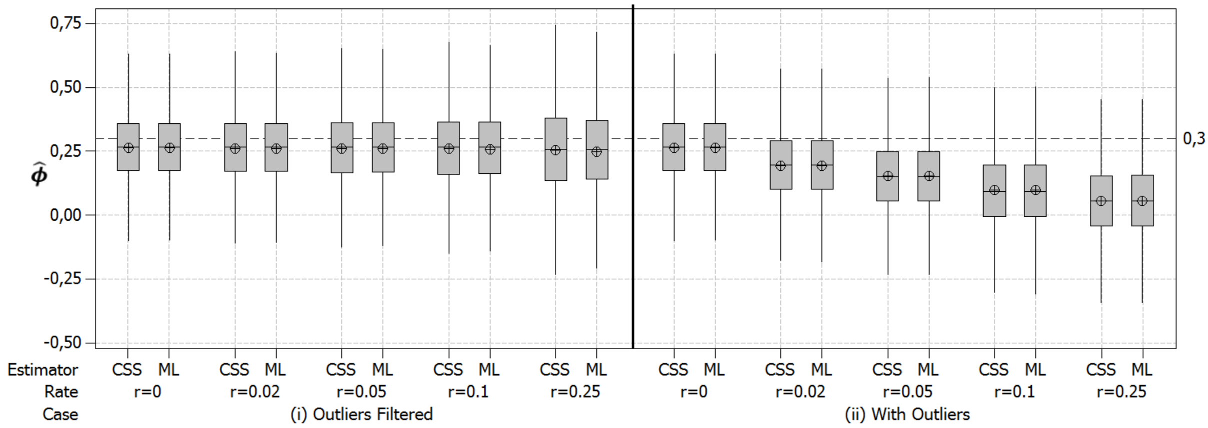
b. Box plot of estimates for the mean μ.



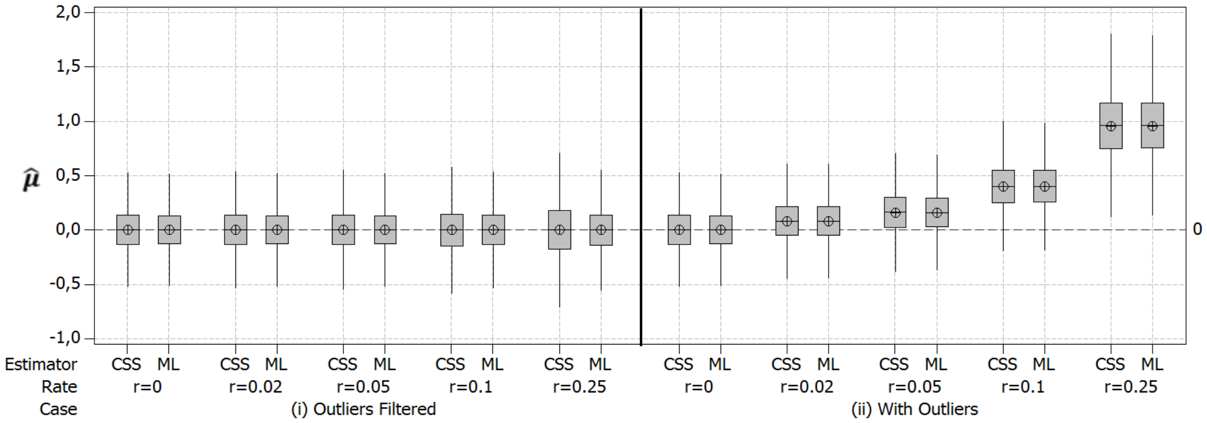
c. Box plot of estimates for the variance σ^2.

Figure 15. Box plots of the parameter estimates from simulations – $T = 100, \phi = 0.7$.

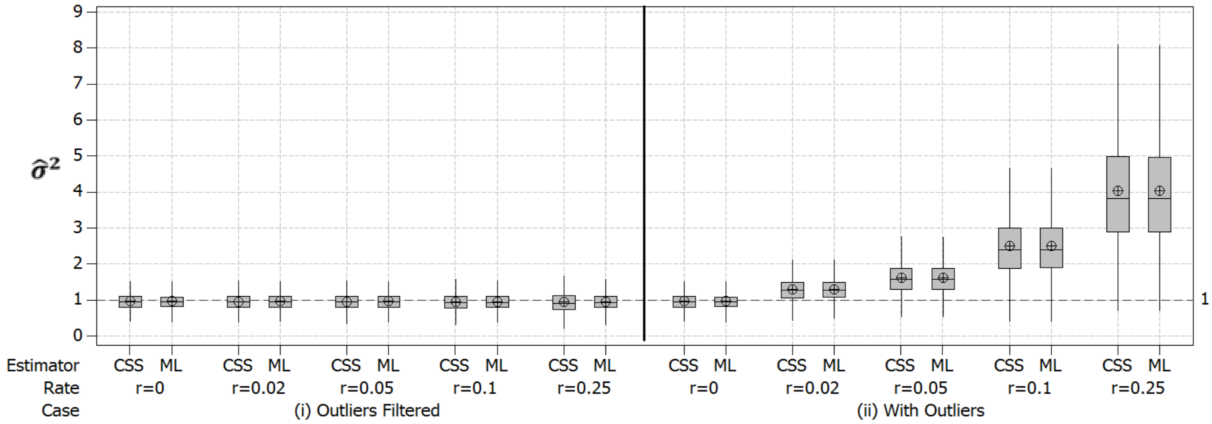
III. T=50



a. Box plot of estimates for the autocorrelation parameter ϕ.

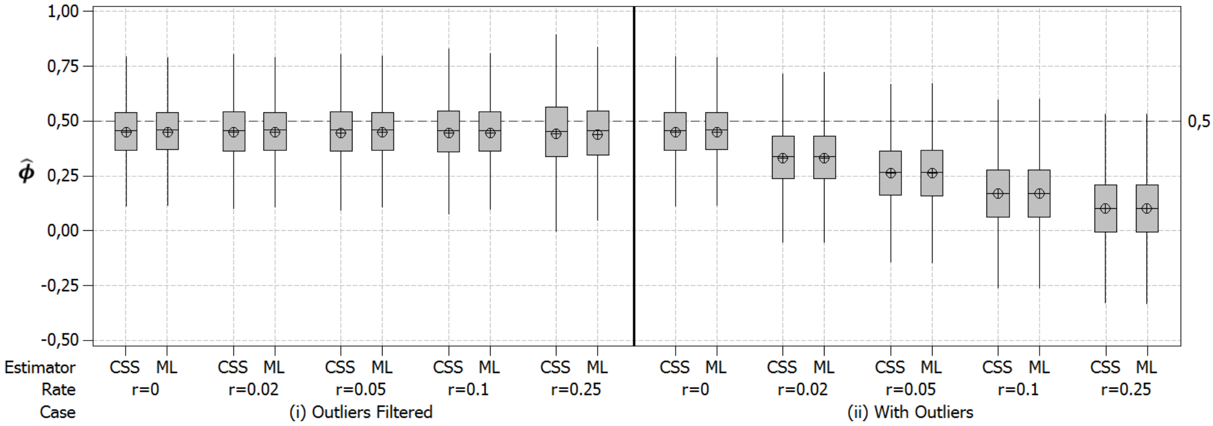


b. Box plot of estimates for the mean μ.

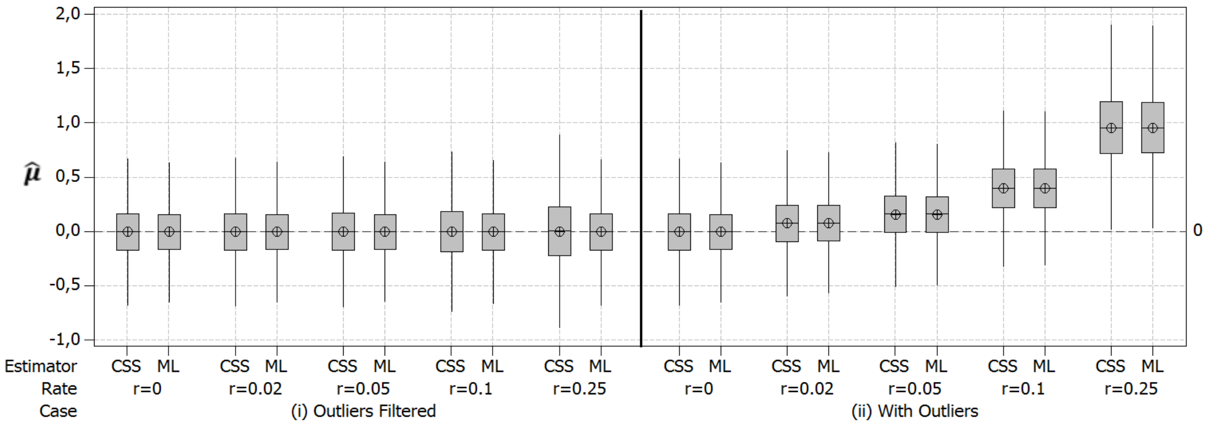


c. Box plot of estimates for the variance σ^2.

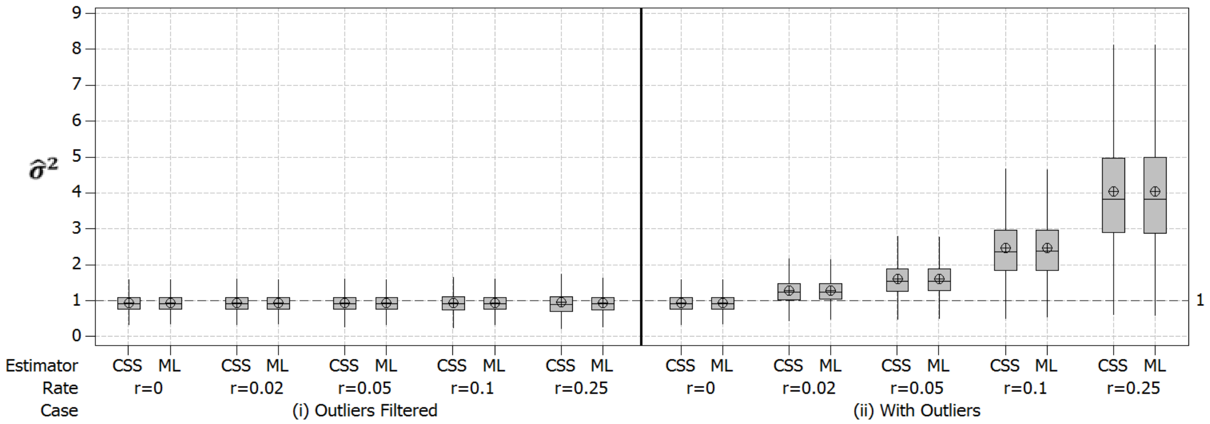
Figure 16. Box plots of the parameter estimates from simulations – $T = 50, \phi = 0.3$.



a. Box plot of estimates for the autocorrelation parameter ϕ.

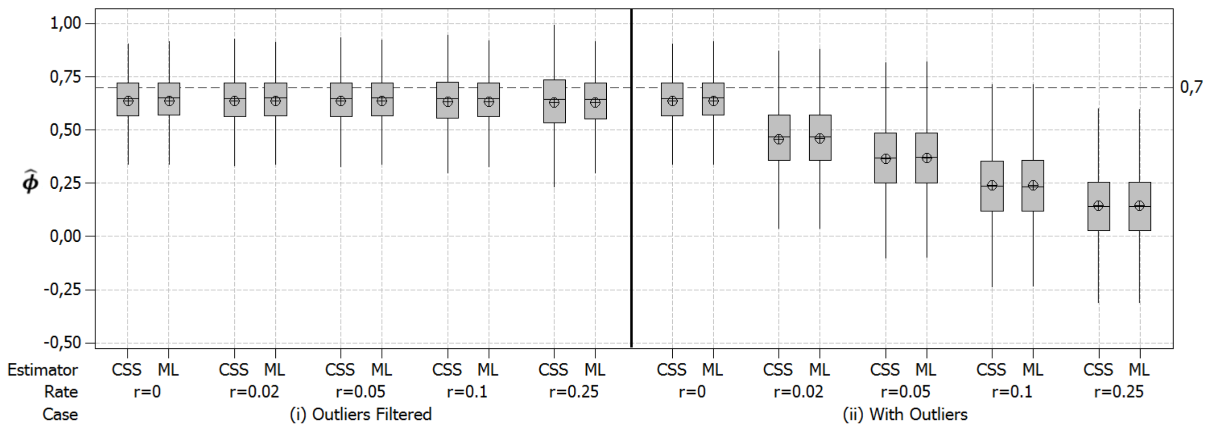


b. Box plot of estimates for the mean μ.

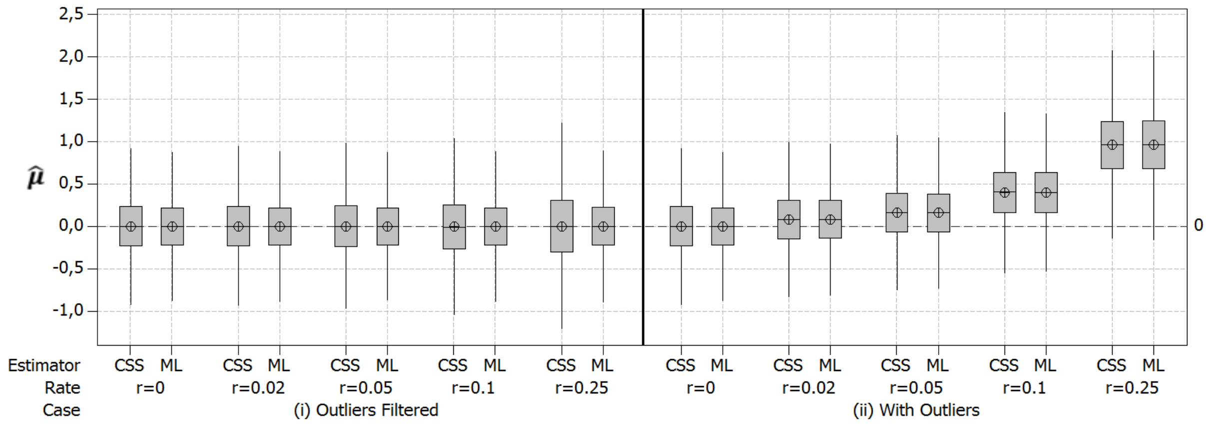


c. Box plot of estimates for the variance σ^2.

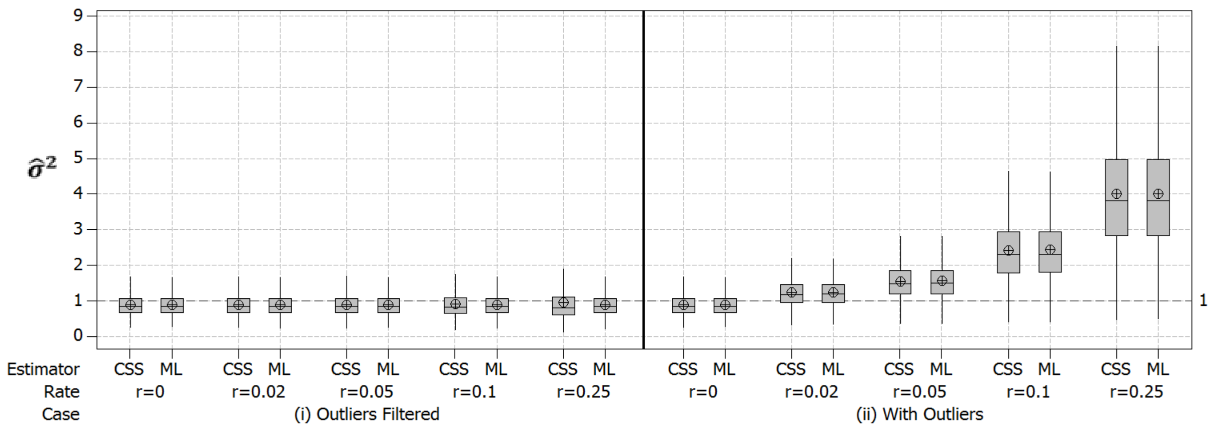
Figure 17. Box plots of the parameter estimates from simulations – $T = 50, \phi = 0.5$.



a. Box plot of estimates for the autocorrelation parameter ϕ.



b. Box plot of estimates for the mean μ.



c. Box plot of estimates for the variance σ^2.

Figure 18. Box plots of the parameter estimates from simulations – $T = 50, \phi = 0.7$.

CURRICULUM VITAE

Credentials

Name, Surname : Erdi Daşdemir

Place of Birth : Kars

Marital Status : Single

E-mail : edasdemir@hacettepe.edu.tr , erdidasdemir@gmail.com

Adress : Hacettepe University, Industrial Engineering Department

Education

M.Sc. : Industrial Engineering, Hacettepe University - Ongoing, (GPA: 4.00/4.00, as of June, 2015)

B.Sc. : Industrial Engineering, Hacettepe University (GPA: 3.61 /4.00, graduated with a rank of top 3 in the department.)

Foreign Languages

English: Advanced

- TOEFL: 109 (May, 2015)
- YDS: 86.25 (April, 2015)

Work Experience

- Hacettepe University, Industrial Engineering Department, Ankara/Turkey, Research Assistant, Date: September 2013 – Ongoing
  - University Research/ Teaching Staff
- BNB Consulting Ltd. Şti. , Cyber Park, Bilkent, Ankara / Turkey, Business Development, Date: October 2012 – August 2013
  - Improving BNB operations to becoming an institutionalized and well-qualified company at international level in terms of project design and management.
- Aselsan, Ankara / Turkey, IT Assistant, Date: June 2012 – July 2012 / 20 Workdays
  - Had responsibilities at Information Management department of Aselsan IT.
  - Had a place in the project of drawing information flow of Aselsan IT.
- Hacı Ömer Sabancı Holding, İstanbul / Turkey, HR Assistant, Date: August 2011 – September 2011 / 20 Workdays

    - Had a place in the Project of Sabancı 2012 Golden Collar Awards as a research team.
    - Researched and collected KPI information for project benchmarks, made a market research, contacted with consultant firms and other competitive firms.
- Pimeks Aluminum PVC & Metal Industry Co., Ankara/Turkey, Production Engineering & Business Development, Date: July 2011 – August 2011 / 20 Workdays
    - Was Responsible for applying 5S production system to the Steel Workshop of the Factory.
- Clementon Park and Splash World, New Jersey / USA , Lifeguard, Date: June 2010 – September 2010

Participated in international working and cultural exchange program in United States and worked in the Clementon Amusement park as a lifeguard.

Areas of Experience

Quality Engineering, Statistical Quality Control, Healthcare Systems Engineering, Lean Systems, Six Sigma Methodology, Process Improvement, Simulation Applications

Projects and Budgets

- Minimizing The Financial Losses Of Hacettepe University Hospitals: Improving Hospital Billing Process and Reducing Billing Errors (TUBITAK, 2.000 TL)

Publications

- Dasdemir, E., Weiss, C., Testik, M. and Knoth, S. . “Evaluation of Phase I Analysis Scenarios on Phase II Performance of Control Charts for Autocorrelated Observations” - *submitted*.
- Dasdemir, E., Weiss, C., Testik, M., and Knoth, S. (21-25 September 2014), “Phase I Analysis of Autocorrelated Time Series Data”, *Proceedings of European Network for Business and Industrial Statistics (ENBIS) 14 Conference*, Linz/Austria.
- Aydın, O., Shaygan, A. and Dasdemir, E. (21-25 September 2014), “A Cause and Effect Diagram and AHP Based Methodology for Selection of Quality Improvement Projects", *Proceedings of European Network for Business and Industrial Statistics (ENBIS) 14 Conference*, Linz/Austria.

- Sonmez, V., Dasdemir, E., Testik, M. and Aydin, O. (5-8 August, 2014), "A Simulation Based Decision Support Tool for Hospital Bed Capacity Planning", *Proceedings of Global Conference on Healthcare Systems Engineering (GCHSE) 2014*, Istanbul/Turkey
- Dasdemir, E., Oguz, M., Atalay, M., Bilgin, V., Testik, M. and Soydan, G. (15-19 September 2013), "Improving Hospital Billing Processes for Reducing Costs of Billing Errors", *Proceedings of European Network for Business and Industrial Statistics (ENBIS) 13 Conference,* Ankara/Turkey.

Oral and Poster Presentations

- "Phase I Analysis of Autocorrelated Time Series Data", *Oral Presentation in European Network for Business and Industrial Statistics (ENBIS) 14 Conference,* Linz/Austria.
- "A Simulation Based Decision Support Tool for Hospital Bed Capacity Planning" *Oral Presentation in Global Conference on Healthcare Systems Engineering (GCHSE) 2014,* Istanbul/Turkey
- "Improving Hospital Billing Processes for Reducing Costs of Billing Errors", *Oral Presentation in European Network for Business and Industrial Statistics (ENBIS) 13 Conference*, Ankara/Turkey.

ARKA BOŞ SAYFA

TEZ CİLDİ ARKA KAPAĞI