Statistical Process Control for Serially Correlated Data

Published by: Labyrint Publication
P.O. Box 662
2900 AR Capelle a/d IJssel
The Netherlands
fax +31 (0)10 2847382

Printed by: Ridderprint, Ridderkerk

ISBN 90-72591-63-1

Rijksuniversiteit Groningen

Statistical Process Control

for Serially Correlated Data

Proefschrift

ter verkrijging van het doctoraat in de
Economische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. D. F. J. Bosscher,
in het openbaar te verdedigen op
donderdag 11 februari 1999
om 14.15 uur

door

Jakob Edo Wieringa

geboren op 30 maart 1970
te Leens

Promotor: Prof. dr. drs. A. G. M. Steerneman

Aan Nathalie

Preface

This thesis is the result of four years of scientific research at the department of Econometrics of the University of Groningen. It was a pleasure for me to work on problems in the field of industrial statistics that are both interesting from a scientific point of view and relevant in practice. At this place, I would like to thank the people and organizations that have supported me in writing this thesis.

First of all, I would like to thank my promotor, Ton Steerneman. He contributed a lot to this thesis, by providing me with many useful suggestions and comments. Thank you, Ton, for your enthusiasm, for your constructive criticism, and for being such a fine colleague. I really enjoyed working with you, and I was proud to be 'your' Ph.D.-student.

As a part of my research, I had the opportunity to work for six months at Philips Semiconductors Stadskanaal. I am very grateful to the people who supported me there. Among them are my colleagues from the Industrial Support Group. They showed great interest in the problems I was working on, and provided me with help whenever needed. In particular I would like to thank Albert Trip and Klaas Huisman for sharing their wit and wisdom in the numerous discussions we have had. I experienced that practical problems are a stimulus to scientific research.

I am thankful for the financial support that I received from the Graduate School/Research Institute Systems, Organizations, and Management (SOM) during the four years of the research project. In addition, I thank my new employer, the Institute for Business and Industrial Statistics (IBIS UvA BV), for providing additional financial support for the completion of my thesis.

I would also like to thank the members of the promotion committee, the professors Wim Albers, Paul Bekker and Ronald Does. They are gratefully acknowledged for their interest in my work, their willingness to read the manuscript, and their useful suggestions for improvements.

Furthermore, I am grateful to Sylvia Peekstok for proofreading this dissertation and for correcting my English. I thank my roommate Arjan

for his pleasant company during the last few years and for letting me use his computer for the many simulation studies of this thesis. I really enjoyed my lunch-talks with Suwarni, Jacobien and Sytze about (in chronological order) cats, dogs and kids. However, I feel I must apologize to my other colleagues, for bringing these subjects up far too often.

I would like to thank my parents, and the rest of my family. Though you did not understand most of what I was working on, you always showed great interest in my work.

Finally, a word of thanks to my 'two women'. Thank you Nathalie, you are responsible for the most beautiful moments of my life. Thank you for your support and your patience, especially in the last few months before the completion of this thesis. Thank you Tineke, your smile just makes my day.

Zoutkamp, december 1998

Jaap Wieringa.

Contents

Chapter 1

Introduction

There has always been a great deal of attention for the quality of manufactured products and services. In ancient times, the quality level was sometimes maintained in brute ways. For example, to quote the code of Hammurabi which was dated to 2150 B.C.:

> *If a builder has built a house, and his work is not strong, and the house falls in and kills the householder, that builder shall be slain.*

A second example concerns Phoenician inspectors who eliminated any repeated violations of quality standards by chopping off the hand of the maker of the defective product.

A more subtle approach towards monitoring and improving quality is Statistical Process Control (SPC), that aims at quality improvement through reduction of variation. One of the primary techniques of SPC is the control chart. After its introduction by Walter A. Shewhart in the twenties, it was W. Edwards Deming who extended his ideas to a quality improvement strategy that is not only applicable in a manufacturing environment, but in all areas of an organization, from administration to sales. This philosophy is known as *Total Quality Management* (TQM). Shortly after the second world war, he succeeded in convincing Japanese topmanagers of the usefulness of this approach. In those years, while the Western countries were only interested in quantity of production, the Japanese built up a head start by focusing on quality of production output. It was not until the late seventies that the Western world realized the necessity of such quality improvement programs. Won over by the success of the Japanese, Western manufacturers became interested in Total Quality Management.

Nowadays, 'quality' is in the center of attention worldwide. Over the years, the control chart has proved its effectiveness in practice as a tool for reducing the variation in process outcomes.

Traditionally, in the development of manufacturing processes, a deterministic viewpoint prevailed. Much engineering effort was put into action to attain a situation where important quality characteristics of every product comply exactly with predetermined target values.

The success of Shewhart's approach is based on the idea that in any production process, no matter how well designed it is, there exists a certain amount of natural variability in output measurements. Such variation is to some extent unavoidable without a profound revision of the process. This type of variation is called *variation due to common causes*. Trying to counteract the effect of the variation due to common causes is in many cases like intervening in a stable system, thereby increasing instead of reducing the variation around the target.

If the variation due to common causes is small compared to the requirements of customers, and if the process is 'on target', this poses no problems. The process is *capable* to meet the demands. However, the process may be affected by external sources of variation, that are upsetting the normal functioning of the process. Such causes are called *special causes of variation*. The presence of special causes may lead to excessive variation in process outcomes, resulting in malfunctioning of the product and customer complaints. In such cases, quality improvement is possible by detection and removal of special causes of variation.

Shewhart developed the *control chart* to detect the presence of special causes of variation. In its basic form, a control chart is a plot of (a function of) observations of a production process against time. The points that are plotted on the graph are compared to a pair of so-called *control limits*. One or more points outside the bandwidth of these limits are called *out-of-control signals* and indicate the presence of special causes of variation, that are upsetting the process.

Traditionally, two fundamental assumptions are made for the development of control charts. Firstly, it is assumed that the distribution function underlying the observations of a quality characteristic of interest, is normal. Secondly, it is assumed that the process data is independently distributed. One or both assumptions are frequently violated in practice. In this thesis, we will concentrate on the case where independence of successive observations cannot be assumed. We will however make the assumption that the data is normally distributed.

For the case of non-normal data, we refer to recent articles by Chou, Polansky and Mason (1998) and Shore (1998) and the references cited therein. Other references are Burr (1967), Schilling and Nelson (1976) and Haridy and El-Shabrawy (1996).

The purpose of this chapter is to provide an introduction to the subject. A precise definition of the terminology that is used can be found in subsequent chapters. We start the chapter with some motivating practical examples in Section 1.1. In Section 1.2, we discuss what violations of the independence assumption are allowed in subsequent chapters. In Section 1.3, the subject of the thesis is discussed. This chapter is concluded with an overview of the thesis in Section 1.4

1.1 Some motivating examples

The most commonly reported effect on control charts of violating one or both of the fundamental assumptions is the erroneous placement of the control limits. Examples of this phenomenon are numerous. In the last subsection of this section we discuss an investigation conducted by Alwan (1989), see also Alwan and Roberts (1995), where, in a sample of 235 control chart applications, it was discovered that about eighty-five percent displayed incorrect control limits. More than half of these displacements were due to violation of the independence assumption. In the following subsections, we will discuss four applications of control charts where violation of the independence assumption leads to misplaced control limits.

1.1.1 Monitoring electrical resistance of insulation material

In Table 2 on page 20 of his pioneering book "Economic Control of Quality of Manufactured Product", Shewhart (1931) presents a data set of measurements on electrical resistance of insulation material. The data is reprinted in Table 6.1 of Chapter 6 of this thesis.

The data set consists of 204 observations, which Shewhart grouped into 51 subsamples of size four. Eight of the sample means exceeded the control limits that were computed for this data. In Chapter 6, the data set is reanalyzed. It turns out that the independence assumption is violated, so that the standard approach for computing control limits is not valid in this case. In Chapter 6, some of the control charts that are discussed in Chapter 3, 4, and 5 which take the effect of serial correlation into account, are applied to the individual observations. These charts indicate that the series of 204 individual observations contain only two (perhaps three) outliers.

In this case, misplacement of control limits is due to serial correlation in the data. The standard control chart methodology suggests the presence of process upsets which cannot be linked to a responsible special cause of variation. In such cases, a search for a cause for the suspected process upsets will be initiated while the process is functioning normally.

In later chapters of this thesis more appropriate control charts for serially correlated data are discussed.

1.1.2 Monitoring the large outside diameter of a flywheel

In an article by McCoun (1949), which was reprinted in the October 1974 issue of *Quality Progress*, a quality problem at International Harvester's truck engine works is presented. A great deal of trouble had been experienced in maintaining the size on the large outside diameter of a flywheel. Samples of size five were taken at approximately one hour intervals, and control charts for the mean and the range were set up. The specifications on the diameter of the flywheel were such that for an in-control process, individual measurements should fall amply within the specification limits.

The resulting control charts indicated no out-of-control situations. However, a large portion of the production did not even comply with the specification limits. By considering a series of consecutive measurements, the cause of this mystery was found. It turned out that the measurements exhibited systematic cyclic behavior. Unknowingly, the samples were taken in phase with two periods of the cycle, so that all the sampled observations came from the same place in the cycle (the bottom). As a result, only a part of the variation in the process outcomes was monitored in the control charts. The out-of-spec products were produced in the part of the production cycle that was not sampled. In this case, misplacement of the control limits occurred due to the fact that the variation observed in the samples did not represent the variation in the process.

This problem was solved by reworking a part of the mechanism of the machine, so that the systematic behavior was removed.

1.1.3 Monitoring viscosity measurements

Montgomery (1996) discusses a data set of one hundred viscosity measurement, taken from a certain chemical process. Based on physical arguments which follow shortly, and from visual appearance of the data set, serial correlation is suspected. Montgomery first sets up a standard control chart for the mean, thereby ignoring these hints. Out of the hundred observations,

nineteen appear to be out of control. However, if the data is monitored with a control chart that accounts for serial correlation, there is no reason to suspect the presence of special causes of variation. Also in this case, violation of the independence assumption leads to too many false out-of-control signals.

The data set is used to illustrate that in some cases it is possible to relate autocorrelation in process measurements to process inertia in an analytical way. Montgomery considers a tank, that is filled up constantly with a certain liquid from the top of the tank. With the same flow rate, the material is flowing out again at the bottom of the tank. The concentration of the liquid entering the tank may vary over time. The concentration of the liquid in the tank is a mix of previous concentrations. Montgomery uses laws of physics to demonstrate that successive concentration measurements taken at equally spaced time intervals on the outflow side of the tank are serially correlated.

In this case, serial correlation cannot be removed without redesigning the process. It must be considered as a part of the process. In other words, the variation due to serial correlation is a part of the common-cause variation. Monitoring data from such a process with control charts that assume independence will result in many out-of-control signals, indicating the presence of serial correlation. When it is known that process data exhibits serial correlation, control charts that account for serial correlation must be used.

1.1.4 Monitoring the composition of diode tin/lead layers

In Chapter 8 of this thesis, we present a case study of a quality improvement project that was conducted at Philips Semiconductors Stadskanaal, the Netherlands, a leading manufacturer of diodes. The problem that will be considered concerns the quality of a tin/lead layer that is applied to diodes. A diode is an electrical component that will be soldered on printed circuit boards. A bad tin/lead layer obstructs soldering of the diodes.

We will discuss several aspects of the quality improvement project that are within the framework of this thesis in Chapter 8. A more complete report can be found in Wieringa (1997). At a certain point in our investigation, it was decided that the ratio of the amount of tin to the amount of lead in the layers needed to be monitored with a control chart. These measurements exhibited serial correlation, so that application of a standard control chart resulted in many false out-of-control signals, see Figure 8.14. However, in this case, the charts that are discussed in subsequent chapters

are not directly applicable either. The data shows *non-stationary* behavior (see Chapter 2 for a definition). A characteristic of non-stationary processes is that the level of the process may wander away from a predetermined target, without ever returning to it. For such processes, it is necessary to apply a kind of active control that keeps the process 'on-target', before the output can be monitored with control charts.

In the case study, we decided to adjust the process once a week to prevent the mean of the process drifting away from its target. In the meantime, the output of the process is monitored with widened control limits to allow for a slight wandering of the mean.

1.1.5 More examples

Other examples of misplaced control limits can be found among the 235 "expert control chart applications" that are studied by Alwan (1989). The author only considered 'real life' data sets, encountered in quality control text books and manuals, advertisements and brochures of quality control software vendors, and articles from a prominent quality control journal.

In over eighty-five percent of the cases, one or both of the assumptions that underlie the standard implementation of control charts are violated, leading to control limits which are not appropriate for the data under consideration. Alwan classified the violations into four categories: non-i.i.d. behavior, purely described by time series models, non-i.i.d. behavior described by models combining time series models and deterministic variables, non-i.i.d. behavior described by deterministic variables, and violation of the normality assumption. Almost half of the observed misplacements are due to a violation of the first category.

To illustrate the consequences of the misplacements, Alwan (1989) determined the total number of out-of-control signals that occurred in the 235 data sets when the standard three-sigma methodology (see Chapter 2) was routinely applied. The total number of out-of-control signals thus identified across all 235 applications was 674. In contrast, the number of out-of-control signals when the three-sigma methodology was applied to residuals of fitted time series models was found to be 90. Alwan (1989) believes that the large disparity (674 versus 90) reflects the fact that due to positive autocorrelation, false out-of-control signals occur more frequently.

Moreover, Alwan (1989) found that the two approaches agreed in the detection of only 57 out-of-control situations. Therefore, the large majority of the out-of-control signals identified by the standard methodology turned out not to be out-of-control signals from the perspective of time series

modelling. In addition, 33 of the 90 out-of-control signals identified by the time series approach were not detected by routinely applying the standard approach.

Hence, by ignoring serial correlation there is a great danger of misinterpretation of the information present in the data. In a substantial number of cases, special causes are not detected when they do, in fact, truly exist, and in even more cases, special causes are signalled that do not exist. As a result, most of the actions that are undertaken on the basis of the out-of-control signals will not produce any effect. The fruitless search for the cause of the apparent, but nonexisting, process upsets will have a detrimental effect on the popularity of the control chart and may obstruct other applications of SPC methods.

1.2 Violation of the independence assumption

When there are only common causes of variation present, observations of a production process may look like the observations in Figure 1.1 (a), (b), (c) or (d).

Figure 1.1(a) represents the case where successive observations are uncorrelated. There is no memory in the data: previous observations do not influence future observations. The process fluctuates randomly around the mean. In most applications of control charts, it is assumed that the data exhibits this kind of behavior. However, in practice, most data sets show some form of serial correlation.

In Figure 1.1(b), successive data points are negatively correlated. An observation below the mean tends to be followed by an observation that is larger than the mean value, and vice versa. The sequence of observations exhibits alternating behavior.

The data in Figure 1.1(c) is positively autocorrelated. If the current observation is on one side of the mean, the next observation will most likely be found on the same side of the mean. Positively correlated data is characterized by runs above and below the mean. This type of serial correlation is more often encountered in practice than negative autocorrelation. It may occur due to mixing of raw materials in a tank as discussed in Subsection 1.1.3.

In Figures 1.1(a), (b), and (c), the observations are *mean reverting*. The process may wander away from the mean in a non-random manner, but will eventually return to the mean. The underlying processes are *stationary* (see Chapter 2). In Figure 1.1(d), realizations of a *nonstationary* process are

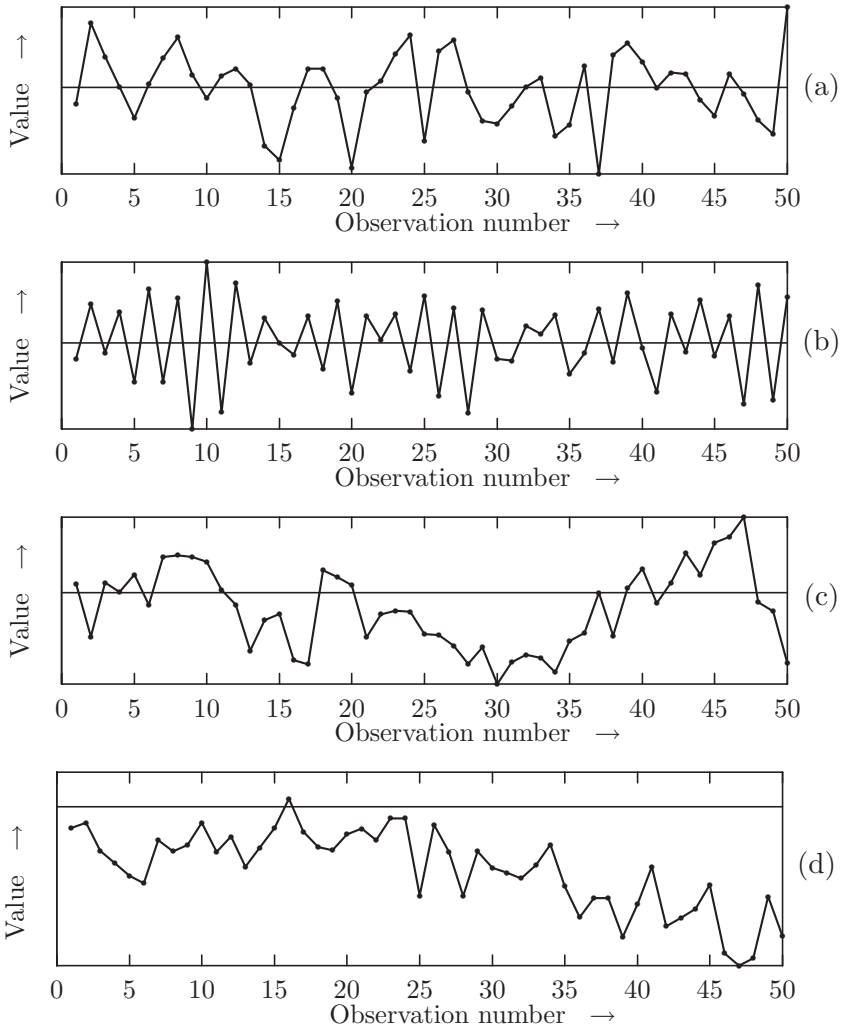


Figure 1.1: Four different kinds of process observations.

drawn. If a non-stationary process is left alone, it may wander away from the mean without ever returning to it.

In a manufacturing environment where target values are set for the quality characteristics, monitoring process measurements of Figures 1.1(a), (b), and (c) is fundamentally different from monitoring a process underlying Figure 1.1(d). The last must be adjusted regularly, otherwise the process will wander away from the target value. In this case, some form of active control is necessary to ensure that the process stays on target. Thereafter, control charts can be used to monitor the adjusted process for special causes of variation.

Also in the case of stationary autocorrelated data (Figures 1.1(b) and (c)), a combination of active control and control charts can be used in some cases. When possible, active control can be applied to the observations to remove the extra variation due to serial correlation. Such procedures are sometimes collectively labeled as *Automated Process Control* (APC) or *Engineering Process Control* (EPC). The purpose of these techniques is to filter out serial correlation, so that observations from the adjusted process are approximately uncorrelated. This has two beneficial effects. Firstly, the variation in measurements is reduced, since the systematic variation due to serial correlation is removed. Secondly, the independence assumption is no longer violated, so that the standard control charts can be applied to monitor the adjusted process for special causes of variation.

Several authors advocate this integration of APC (or EPC) and SPC (EPC) methods. A classical article is Box and Kramer (1992), other references are MacGregor (1988), Part II of Keats and Hubele (1989), Montgomery, Keats, Runger, and Messina (1994), Box, Coleman, and Baxley (1997), Box and Luceño (1997a,1997b), and Janakiram and Keats (1998).

Active control of processes requires a thorough knowledge of the underlying process. Such knowledge is not always present. Furthermore, it is required that there are possibilities to adjust the level of the process frequently in a highly accurate manner. This may not be very cost-efficient. In other cases, it is (given the process) not possible to remove the serial correlation at all (see the example discussed in Section 1.1.3). Therefore, in some cases, serial correlation cannot be removed before control charts are applied to monitor process outcomes, and this systematic variance component must be viewed as part of the common-cause variation of the process. This induces the need for the development of control charts which allow for violations of the independence assumption. This is the subject of this thesis.

1.3 Scope of the thesis

It is the purpose of this thesis to extend the application of control charts to cases where serially correlated observations are allowed. This extension is interesting from a theoretical point of view. However, the main motivation stems from practical situations such as the examples from Subsection 1.1, which illustrate that the independence assumption is often not fulfilled for many real life data sets. The question that is raised in this thesis is how a sequence of serially correlated observations is properly monitored for the presence of special causes of variation.

This question is treated as follows. Firstly, it will be investigated what happens if the serial correlation in the data is not recognized or ignored, and standard control chart procedures are applied to the observations. This is what would happen in most practical cases. An operator or a process engineer might not be aware of the problems that are caused by serial correlation in the data, and consequently apply control charts as if the data were uncorrelated.

Secondly, a survey of existing procedures that take serial correlation into account will be conducted. In addition, if necessary, alternative procedures will be developed that take serial correlation into account.

And thirdly, a comparison of these procedures is conducted on the basis of *Average Run Length* (ARL) behavior, a notion that will be defined in Chapter 3. We are aware of, and do agree with, the objection that the ARL does only reflect a small part of the behavior of the control chart. Indeed, it would be better to discuss the distribution of the run length. On the other hand, the ARL is widely used in the literature to compare different control charts.

We will restrict ourselves to stationary data, which is assumed to be generated by an ARMA model with normally distributed disturbances (see Chapter 2). The class of ARMA models can be used to capture a wide range of serial dependence in the data. Among these models, the model for first-order autocorrelation (the AR(1) model) is commonly reported as being most frequently encountered in practice. Therefore, we will only pay attention to the AR(1) case. Results for other models are derived in a similar manner and are not discussed in the text, but only displayed in an appendix. We will not go into detail concerning the identification and estimation of an appropriate time series model. We will assume that enough data is available (for example as a result from a so-called prerun) to correctly identify and estimate an underlying process model. In addition, we will not discuss the possible integration with APC (or EPC) techniques.

We refer to the references cited earlier.

1.4 Outline of the thesis

Chapter 2 introduces the reader to terminology that is used in the remainder of the thesis. The chapter starts out with a review of the philosophy of SPC, in particularly the control chart. Secondly, some theory concerning time series models is briefly discussed. The chapter ends with a discussion of the problems that arise when the variance of serially correlated data is estimated as if the observations were independent. References to relevant literature are made where appropriate estimators are discussed. For the remainder of the thesis, the problems concerning the estimation of the variance are acknowledged, but assumed to be nonexistent, since there are good methods to obtain estimates of the process parameters.

In Chapter 3, we will discuss and evaluate the ARL behavior of two types of Shewhart-type control charts that are suggested in the literature. It turns out that both control charts do not exactly comply with the intended ARL behavior. Apart from the estimation problems of Chapter 2, the standard control chart theory with appropriate variance estimates cannot be routinely applied to serially correlated data. Especially for large, positive autocorrelation, we will find that shifts in the mean of the process may go undetected much longer than expected, compared to the case of uncorrelated observations. For this reason, a third Shewhart-type control chart is developed that has improved ARL behavior for first-order autocorrelation.

For each of these three approaches, an *Exponentially Weighted Moving Average* (EWMA) or a *CUmulative SUM* (CUSUM) can be computed. The resulting sample statistics can be drawn in corresponding EWMA or CUSUM control charts. In Chapter 4, the ARL properties of these EWMA-type control charts are evaluated. In Chapter 5 the CUSUM-type control charts are considered. As a result, Chapters 3, 4, and 5 have basically the same design. In Chapter 6 two examples are worked out that illustrate the theory in previous chapters.

In Chapters 3, 4, and 5, control charts are developed for the case of individual observations. These individual data points may be means of samples or individual measurements. In Chapter 7, we consider the case where subgroups of observations are sampled (the case $n > 1$). This poses new problems for the control chart for the mean of the process, but also allows us to consider control charts for the spread of the process. In the case

of individual observations, which is considered in the Chapters 3, 4, and 5, it is hard to find a suitable measure of the within-sample variation, because the 'sample' consists of only one observation. Therefore, the focus in those chapters is on control charts for the mean. However, with subgroups of size $n > 1$ in Chapter 7, it is possible to consider control charts which monitor the process for special causes that affect the standard deviation of the process.

In Chapter 8, a case study is presented that was carried out in the context of this thesis. Several aspects of a quality improvement project that was conducted at Philips Semiconductors Stadskanaal, the Netherlands, are considered. One of the problems we encountered there was how to monitor serially correlated data.

Chapter 2

Statistical Process Control

In this chapter, we will discuss the basics of *Statistical Process Control* (SPC). Concepts that are used in subsequent chapters will be defined. In Section 2.1, the meaning of the word "quality" is explained. In Section 2.2, we will focus on the relation between quality and variation. Two important concepts, namely, "common causes of variation" and "special causes of variation", will be discussed. Furthermore, we will argue that the philosophy behind SPC is not only applicable to manufacturing environments, it is useful in all areas of an organization. In Section 2.3, the Shewhart control chart for independent observations is introduced. In subsequent chapters, we want to study the effect of serial correlation on the behavior of the control chart. In Section 2.4, we review some of the time series models that are used throughout this thesis.

2.1 Quality defined

The word "quality" is used by a lot of people in different contexts. In most companies it is recognized that quality of a service or a product is very important. However, it is very hard to give a perfect definition that discriminates products or services of bad quality from products or services of high quality. Various authors have attempted to formulate such a definition.

Shewhart (1931) argues that there are two aspects of quality. Firstly, there is an *objective* concept of quality, resulting in quantatively measurable physical characteristics, that are independent of a second, *subjective*, aspect of quality. The latter has to do with what we think, feel or sense. Shewhart recognizes that the subjective side of quality is commercially interesting,

but that it is necessary to establish standards of quality in a quantitative (objective) manner. Deming (1982) also stresses the subjective side of quality:

> *Quality can be defined only in terms of the agent (p. 168).*

Furthermore, he emphasizes that impressions of quality are not static. They change over time. This creates problems in defining quality, since it is difficult to translate future needs of the user into measurable characteristics.

In a vivid case history starting on page 41, Crosby (1979) defines quality as "conformance to requirements". In this view, a product is of good quality if it meets its specifications. This definition encloses an important part of what customers perceive as "quality". A manufacturer will most certainly receive a lot of 'quality complaints' if a large part of his production exceeds tolerances that were agreed upon with the customers. On the other hand, the definition is too narrow. For example, the specifications themselves are part of what is perceived as "quality". A product that is produced in conformance with specifications that are not popular with customers will not be ranked as a high quality product.

In Juran and Gryna (1988), a broader definition is given: "quality is fitness for use". In this view, "quality" is defined as a relative notion. Different usages of the product will result in different requirements with regard to the product. Let us consider shoes as an example. The requirements of a person looking for high quality jogging shoes will differ from the requirements of the same person looking for high quality elegant shoes. Montgomery (1996) considers 'conformance to requirements' as one of two aspects of 'fitness for use'. The other aspect is 'quality of design', which emphasizes the intentional design differences between types of a product. The conformance aspect is how well the product conforms to the specifications required by the design.

Nowadays, these definitions of "quality" are labeled as traditional. It is recognized that the quality of a product or a service is not a single, identifiable characteristic. Garvin (1987) distinguishes the following eight dimensions of quality, which, when taken together, incorporate more aspects than the 'traditional' definitions of quality.

1. **Performance** is one of the traditional measures of quality. It refers to the basic functioning of a product. For example, for an automobile, performance would include acceleration, handling, cruising speed, and comfort;

2. **Features** that are added to the basic functioning of a product attribute to a higher quality;

3. **Reliability** is another traditional measure of quality. A reliable product rarely fails. This aspect, which is sometimes reformulated as 'being free of deficiencies' is a very important dimension of quality;

4. **Conformance** is related to reliability. It refers to the degree to which a product meets pre-established requirements. This dimension of quality is very important in situations where products are used as the components in a more complex assembly. Specifications on the individual components are usually expressed as a target and a tolerance. If each of the components is just slightly too big or too small, a tight fit is unlikely, and the final product may not perform as intended by the designer, or may wear out early;

5. **Durability** is a measure of product life, either economically (expected cost of repair exceeds current product value) or physically (repair is impossible). A product that lasts longer is usually viewed as being of higher quality;

6. **Serviceability** relates to the time and effort that is needed to repair a product. The breaking down of a product is usually viewed as an annoyance, but a prompt repair may relieve part of the irritation;

7. **Aesthetics** is a subjective dimension of quality. It refers to the look, feel, sound, taste or smell of a product. It is greatly influenced by the preferences of the individual customer. On this dimension of quality it is usually not possible to meet the needs of every customer;

8. **Perceived quality** is also a very subjective dimension. When customers do not have full information about a product they may base their quality image on past experiences, the reputation of the manufacturer, the quality of other products from the same manufacturer, or the name of the product.

Realizing that quality is not a one-dimensional characteristic of a product and that quality is determined at various levels in the production process,

Garvin stresses that manufacturers should not strive to be first on all eight dimensions of quality. Rather, he should select a number of dimensions on which to compete.

Traditionally, the quality control departments in factories compete with regard to the conformance dimension of quality. It is their responsibility to ensure that requirements set on a quality characteristic are met. Such requirements are usually stated in the form of specification limits. All parts within limits are classified as conforming. The objective then is to produce "zero defects". Sullivan (1984), argued that this conformance-to-specification-limits approach effectively prevents ongoing quality improvement. As long as all outcomes of a production process are within specification limits, a process engineer would have great difficulty in convincing his plant manager to make any investment for the improvement of quality. Sullivan advocates defining quality as "*uniformity around the target*". In this more modern point of view, which was put forward by, among others, Deming and Taguchi, any deviation from target reduces reliability and increases costs, in the form of plant and customer loss. Operational objectives directed towards achieving ongoing quality improvement should not be stated in terms of specification limits such as "zero defects". The attention for quality improvement will diminish as soon as the manufacturing process is able to produce amply within specs. A more continuous drive for ongoing quality improvement will be obtained if the aim is to *reduce variation around the target*.

This relation between "quality" and "variation" is summarized in the following phrase which can be found in Montgomery (1996):

"*Quality is inversely proportional to variability*".

This statement demonstrates the contribution of statistical methods to quality improvement projects. By acknowledging that variation is present in process outcomes, and that they are to some extent uncertain, it becomes necessary to employ methods which take uncertainty explicitly into account.

In this thesis, techniques are described that are aimed at improving quality through reduction of variability. We realize that this only covers one of the many facets of quality, but nevertheless, it can make an important contribution.

2.2 Reducing variability

In the previous section, an attempt was made to define the use of the word "quality". As a result of the close relation between quality and variation, it was natural to use terminology that expresses uncertainty in process outcomes such as "rarely", "degree to which", "probability", and "variation". It was also argued that quality improvement can be obtained by reduction of variability. The approach towards reduction of variability is based on the idea that the phenomena causing variation in process outcomes can be classified in two groups: *common causes of variation* and *special causes of variation*. In Subsection 2.2.1, the difference between these classes is discussed. Moreover, we will concern ourselves with the question how this distinction can be utilized to improve the quality of process outcomes. In many practical cases, the implementation of this approach is limited to manufacturing processes. However, the basic ideas are applicable in all areas of the organization, from production to sales. It is important that the whole organization is aware of the opportunities of quality improvement by reduction of variation. This concept, which is sometimes called "Statistical Thinking", is discussed in Subsection 2.2.2.

The information concerning the presence of one or both types of variation is in practice based on samples from a process. These samples should be taken with care. This is discussed in Subsection 2.3.1.

2.2.1 Common causes and special causes of variation

Dr. Walter A. Shewhart is unmistakably the founding father of Statistical Process Control. The concepts he developed in the twenties have been written down in his pioneering work, "Economic Control of Quality of Manufactured Product" (1931). His ideas are still very relevant and the most important tool, the control chart (which will be discussed in the next section and, in fact, throughout this thesis), is nowadays used world wide in a virtually unaltered form.

Shewhart compares a manufacturer that tries to make all products conform with a certain target to a marksman, aiming at a bull's-eye. Just as it is impossible for the marksman to hit the bull's-eye with every shot, it is not to be expected that every product will exactly comply with the target. This is not a problem in itself, since 'small' variations around the target are acceptable for most customers. The problem is in Shewhart's words:

> "*how much may the quality of a product vary and yet be controlled? In other words, how much variation should we leave to chance?*"

It is important to realize that this formulation of the problem explicitly states that, even if the quality of a certain product is 'controlled', some variability must be allowed for. The marksman may do his utmost to hit the bull's-eye with each shot, but he cannot hope to hit the target every time. A complex of causes of variation exerts its influence on the outcome of each shot. Both the manufacturer and the marksman are not always able to do what they want to do and cannot hope to precisely understand why.

The marksman's example is just one of many that can be given to illustrate that in all aspects of our lives, there is some variation that is considered to be 'normal' or 'acceptable', and does not call for action. The underlying sources that are responsible for this type of variation were originally called *chance causes of variation* by Shewhart. Nowadays, the term *common causes* is used, a reformulation that is due to Deming. An important aspect of the variation due to common causes is that it is to some extent *predictable*: it is possible, based on earlier experiences, to determine limits that bound the effect of the common causes. In the marksman's example, numerous common causes such as wind, trembling of the marksman's hands, and the precision of the rifle, affect the result of each shot. As long as the hits are within some range of the bull's-eye, the marksman leaves the variation to chance.

If the observed variability is such that it exceeds the boundaries of 'normal' variation we are inclined to undertake action. The presence of something out of the ordinary, not within the class of common causes, is suspected. Such causes of variation belong to the class of *special causes* (in Shewhart's terminology: *assignable causes*). If a single shot of the marksman deviates more than normally from the bull's-eye because he was startled by some noise of the bystanders, he might want to ask them to keep quiet or he might want to use earplugs.

In a manufacturing environment, it is, in most cases, not difficult to visualize a large number of common causes, the joint effect of which causes variation in the outcomes. This variation is inherently part of the process, and is always present, from day to day, from hour to hour. Usually, it is not within the power of an operator to remove common causes of variation; such variation is 'left to chance'. If it is necessary to remove common causes of variation, this requires in most cases a profound revision of the process,

which is the responsibility of the owner of the process, i.e. management.

Special causes of variation are *not* part of the process, and occur only accidentally. However, when a special cause of variation is present, it will have a large effect on the outcomes of the manufacturing process. If removal is possible, a special cause can usually be eliminated without revising the process. In many cases, an operator can be instructed to recognize and remove special causes of variation, thereby improving the quality of the outcomes of the process.

It is important to realize that the responsibility for reducing the effect of special causes of variation lies on a different management level than the reduction of common causes of variation. Counteracting the effect of special causes of variation can be delegated to operators, whereas reducing the effect of common causes of variation is the responsibility of the owner of the process. It is important for an operator to know whether or not special causes are present, so that he can undertake action to remove this cause of variation. But it is even more important to know when to leave the process alone when only common causes of variation are affecting the outcomes.

What happens in practice, is that operators try to counteract the effect of common causes of variation as if it was a special cause of variation. In many cases, this will result in larger variation in the outcomes. This phenomenon of intervening in a stable process when it would have been better to do nothing, is called '*tampering*'. It results in frustration, because of unsuccessful searches for special causes of variation, and in waste of time and money.

It is therefore of critical importance to be able to distinguish situations where only common causes of variation affect the outcomes of a process, from situations where also special causes are present. If only common causes of variation are present, the manufacturing process is said to be "*statistically in control*". This does not mean that there is no variation, or that there is small variation. It does mean that the outcomes are predictable, within statistical limits. In Shewhart's words:

> "*... a phenomenon will be said to be controlled when, through the use of past experience, we can predict, at least within limits, how the phenomenon may be expected to vary in the future. Here it is understood that prediction within limits means that we can state, at least approximately, the probability that the observed phenomenon will fall within the given limits.*"

The predictability of a process that is statistically in control is the basis for the control chart, a tool that can be utilized to distinguish between situations where only common causes of variation affect the outcomes of a process, and situations where also special causes are present. The control chart will be discussed in Section 2.3 for the case of independent observations.

In the case of independent observations, the term 'an in-control process' is associated with a sequence of independently and identically distributed observations. In most applications and in many textbooks, these requirements are adopted as necessary and sufficient conditions for an in-control process. However, it must be noted that Shewhart's original definition does not require independence of successive observations from an in-control process. It only demands that we can predict how the process may be expected to vary in the future. In Subsection 2.5, we will extend the definition of an in-control process to include cases where local deviations of the mean are allowed, due to stationary serial correlation. Such a definition is entirely in line with Shewhart's definition of an in-control process.

In the next subsection, which is based on an article by Snee (1990), it is shown that the philosophy behind SPC is not restricted to manufacturing processes alone.

2.2.2 Statistical thinking

Application of the idea of improving quality by reduction of variability is in most cases limited to manufacturing environments. However, the concept can be applied at many other levels of an organization, as well. In Snee (1990), the place of Statistical Process Control in the much broader context of *Total Quality Management* is discussed. Successful implementation of such a strategy requires a new way of thinking, which he calls *statistical thinking*. These ideas can provide a useful contribution to efforts aimed at quality improvement at all levels in an organization, from production to sales. Snee describes the essence of statistical thinking as follows:

> "*... all work is a series of interconnected processes and identifying, characterizing, quantifying, controlling, and reducing variation provide opportunities for improvement.*"

Even more insightful is the schematic presentation that Snee gave of statistical thinking in quality improvement, see Figure 2.1.

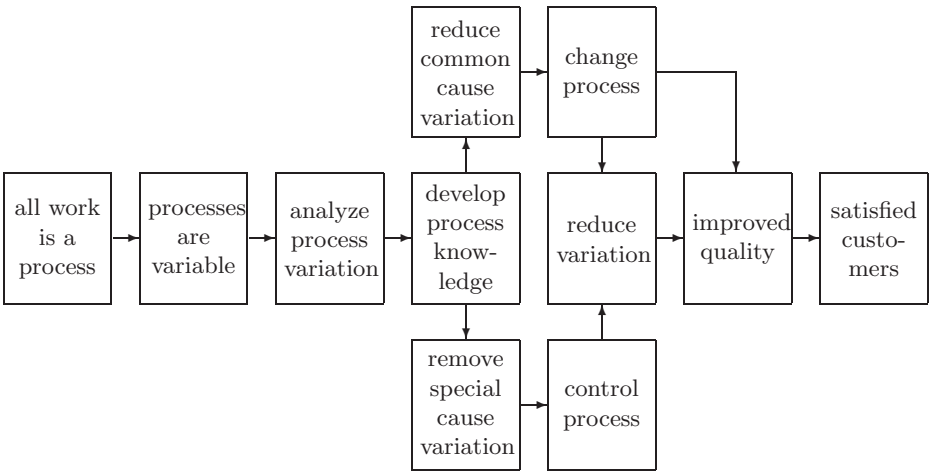


Figure 2.1: Statistical Thinking in Quality Improvement

In the first two boxes of Figure 2.1, it is indicated why statistical thinking is a logical approach to follow in all activities that are aimed at improving quality. All work that is done can be viewed as a process. A process can be defined as (see Nolan and Provost (1990)): "*a set of causes and conditions that repeatedly come together to transform inputs into outcomes*". The inputs might include people, materials, or information. The outcomes include products, services, behavior, or people.

In all such processes, variation is encountered. Careful analysis of this variation, combined with knowledge of the process may lead to reduction of variation. It is therefore important for a manager to realize that close cooperation with those who work with the process (e.g. operators), is an absolute necessity for successful quality improvement. The people that work with the process posses much of the knowledge that is needed to reduce variation.

Reduction of variation may be accomplished by one of two paths. Reduction of variation may be brought about by removing special causes of variation, which is the responsibility of the people working with the process. The other path is to reduce the effect of common causes of variation, which requires the management to undertake action. Removal of common causes requires a different approach than removal of special causes of variation.

However,

> *"Deming and his colleagues point out that managers typically treat all problem as due to special-cause variation, when, in fact, more than 85% of problems are due to defects in a system (common-cause variation), which only management can change. The result is that management spends too much time 'fire-fighting,' solving the same problem again and again because the system was not changed"*
>
> Snee (1990) page 120.

2.3 The Shewhart control chart

In the foregoing section, we argued that it is very important to detect the presence of special causes of variation. A tool is needed for this, since the effect of a possible special cause is hidden in the variation due to common causes. Shewhart developed the *control chart* for this purpose. It will be discussed in this section.

The control chart is based on the idea that if the process is in a state of statistical control, the outcomes are predictable. Based on previous observations, it is possible for a given set of limits to determine the probability that future observations fall within these limits.

Observations that are utilized for monitoring the process are usually grouped according to time, amount of production, or some other descriptive statistic. Sampling and subgrouping should be carried out with care. In Subsection 2.3.1, it will be discussed how to take and group the observations. From each subgroup sample, descriptive statistics such as the mean, the range or the sample standard deviation are computed.

In its original form, the control chart is a simple time plot of a sequence of subgroup statistics. The points in the plot are compared to limits, which indicate the bandwidth of the variation due to common causes. These limits are called *control limits*. The width of these limits is such that, as long as all points are within the control limits, it is reasonable to assume that the underlying process is statistically in control. A point outside the control limits is called an *out-of-control signal*, as it indicates that more variation is present than can be attributed to the effect of common causes of variation. However, due to the random nature of the observations, there is a small probability that an out-of-control signal is encountered while the process is

statistically in control. Such a signal is called a *false out-of-control signal*. In Figure 2.2, an example of a control chart is depicted.

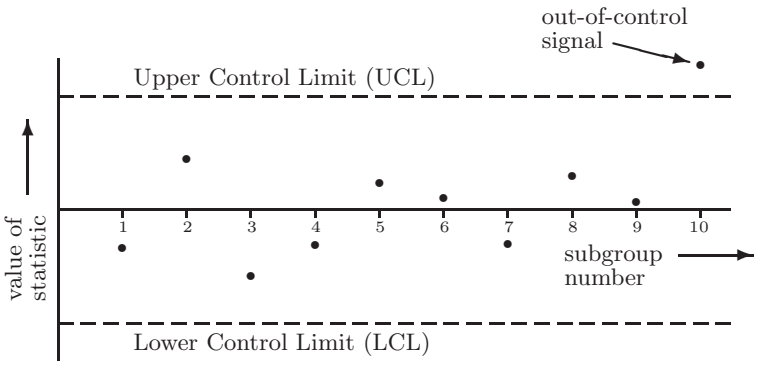


Figure 2.2: Illustration of a control chart.

Control limits are not to be confused with specifications or other targets for the process. They are simply a prediction of the variation that will occur due to common causes. If the variation due to common causes is relatively large, all points on the control chart may be within the control limits, but the process outcomes might fail to meet the specification limits. The presence of special causes does not necessarily mean that there is large variation, or that the specifications are not met. It does mean that there is some source of variation that causes the measurements to be more variable than can be attributed to common causes.

Specification limits are agreements between manufacturer and its customers concerning the tolerated deviation from target. They are in no way related to the actual performance of the process. Control limits on the other hand indicate the magnitude of the variability of the process when only common causes are present. They reveal what Nelson (1988) calls 'the heartbeat of the process'.

It is not advisable to plot specification limits in a control chart. Not only to avoid confusion: understanding the difference between the two concepts is hard enough. A more important reason is that specification limits relate to single products, whereas control limits are usually computed for a grouped variable such as a sample mean. A sample mean that is both within control limits and specification limits may give the impression that the process is performing as required, also in situations where some of the

individual observations exceed the specification limits.

2.3.1 Rational subgroups

Briefly, a control chart is a tool designed to judge whether a process is statistically in control or not. The control chart utilizes samples of observations, mostly drawn in subgroups, to obtain information about the behavior of the underlying process. Statistics that summarize the information in the subsamples are compared to limits which represent the variation due to common causes. If an out-of-control signal is observed, action is required to track down the special cause that is responsible. Sampling and subgrouping of the observations must be carried out in such a way that the search for a causal relation between the out-of-control signal and some underlying disturbing phenomenon is facilitated.

The way the observations are sampled and grouped may have a large effect on the behavior of a control chart. This phenomenon is illustrated in Section 6.1, where serially correlated measurements result in control limits that are too tight, leading to too many out-of-control alarms. Subgrouping of the data should be carried out with care, especially in situations where subsequent observations are serially correlated.

In addition to providing us with the control chart, Shewhart also gave guidelines for sampling subgroups of observations from a process. To this end, he introduced the concept of "*rational subgroups*". A rational subgroup is a sample in which all of the items are produced under conditions in which only common causes are responsible for the observed variation. Special causes do not occur *within* a rational subgroup, but only *between* subgroups. If the observations can be grouped according to these requirements, then appropriate control limits can be determined that discriminate between in-control situations and out-of-control situations.

Linking an out-of-control signal to a specific special cause of variation is facilitated by preserving the time order in the data points on the control chart. The time at which an out-of-control signal is given may give a hint as to when the special cause has occurred.

A control chart must be sensitive enough to detect the effect of special causes of variation, but must not generate too many false out-of-control signals. In practice, a balance between these two must be struck by determining the width of the control limits. How this is done for the classical Shewhart chart is the subject of the following two subsections. Following Does and Schriever (1992), we distinguish two situations.

In Subsection 2.3.2, the situation is discussed where we have k rational

subgroups of size n of past observations available. These observations are used to set up the control chart. This is sometimes called "Phase I", Does and Schriever (1992) call it "analysis of past data". In Phase I, the magnitude of the variation due to common causes is determined, so that the width of the control limits can be determined for what is sometimes called "Phase II". In Phase II, rational subsamples become available one by one, as they are drawn online from the process. Does and Schriever (1992) call this phase "*performance of current control*. Determining control limits for the Shewhart chart in Phase II will be discussed in Subsection 2.3.3.

2.3.2 Phase I: setting up of the control chart

In order to be able to compute the control limits for Phase I, some formalization of the foregoing is needed. Suppose that, from a certain manufacturing process, k rational subgroups of size n are sampled. Let us denote the j^{th} observation of the i^{th} subsample by X_{ij}, where $i = 1, \cdots, k$, and $j = 1, \cdots, n$. For the time being, we will assume that for all i, the individual observations $X_{i1}, \cdots, X_{in}$ are identically and independently distributed within sample i, with distribution function F_i. Furthermore, observations in different samples are also assumed to be independent. The independence assumption will be loosened in subsequent chapters.

Note that the foregoing includes a formalization of the concept of rational subgroups. Distribution functions $F_1, \cdots, F_k$ are used to model the joint effect of common and special causes of variation. Within each sample, the distribution function does not change. However, due to the presence of special causes, the distribution function may change over time.

If there are no special causes present, we assume that the distribution functions do not change over time. Monitoring the process for special causes of variation can then be formalized as testing the hypothesis

$$\mathrm{H}_0 : \ F_1 = F_2 =, \cdots, = F_k \equiv F_0$$

against the alternative

$$\mathrm{H}_1 : \ \text{there are } s, t \in \{1, 2, \cdots, k\} \text{ such that } F_s \neq F_t.$$

In many cases, it is assumed that the distribution functions $F_1, \cdots, F_k$ are known except for a few parameters (for example the expectation and the variance of a normal distribution). Control charts are set up for each of

these parameters. An out-of-control signal on one of these charts is an indication of the presence of a special cause of variation.

The control chart for a one-dimensional parameter η_i of F_i is set up in the following way. Let $T_i = T(X_{i1}, \cdots, X_{in})$ be an estimate of η_i based on the sample at time i. Furthermore, let M_k and V_k be statistics based on the k samples such that $M_k = M(X_{11}, \cdots, X_{1n}, \ldots, X_{k1}, \cdots, X_{kn})$ is a consistent estimator of the location of the distribution of T_i under H_0, and $V_k = V(X_{11}, \cdots, X_{1n}, \ldots, X_{k1}, \cdots, X_{kn})$ is a consistent estimator of the spread of the distribution of T_i under H_0.

In the control chart, realizations of T_i are plotted against i. These values are compared to control limits of the form

$$L = M_k + c(n, k, p)V_k, \tag{2.1}$$

where $c(n, k, p)$ is the p^{th} percentile of the null distribution of $(T_i - M_k)/V_k$. For the LCL and the UCL of a control chart, different constants $c(n, k, p_{\text{LCL}})$ and $c(n, k, p_{\text{UCL}})$ must be determined.

In situations where a location parameter or a spread parameter of the distribution function of T_i is known, these values are used in (2.1) instead of their estimates.

Note that the elements of the sequence $\{T_1, \cdots, T_k\}$ are mutually independent, but this does not, in general, hold for T_i, M_k, and V_k, since they are (partly) based on the same set of observations.

In most literature on SPC it is assumed that F_i is a normal distribution function with expectation μ_i and variance σ_i^2. The mean and/or variance of the observations may change over time due to the presence of special causes of variation. With these assumptions, the process is in control if and only if $\mu_i = \mu$ for some μ *and* if $\sigma_i = \sigma$ for some σ for all $i = 1, \cdots, k$. It is for this reason that in a lot of cases, a production process is monitored using two control charts, one for the standard deviation, and one for the mean of the process.

Shewhart did not consider statistical arguments to determine the constants $c(n, k, p_{\text{LCL}})$ and $c(n, k, p_{\text{UCL}})$. He decided to choose, "based on economic considerations"

$$c(n, k, p_{\text{LCL}}) = -3$$

and

$$c(n, k, p_{\text{UCL}}) = 3.$$

These values turn out to work well in a lot of practical cases. The underlying statistical arguments for a control chart for the mean of normal observations are the following. Suppose that we are testing the hypothesis $\mathrm{H}_0 : \mu_1 = \cdots = \mu_k \equiv \mu$ where μ is known, assuming a constant known variance σ^2. Furthermore, assume that the sample means $T_i = 1/n(X_{1i} + \cdots + X_{ni})$ are plotted in a control chart with limits $\mathrm{LCL} = \mu - 3\sigma/\sqrt{n}$, and $\mathrm{UCL} = \mu + 3\sigma/\sqrt{n}$. Then we have $p_{\mathrm{LCL}} = (1 - p_{\mathrm{UCL}}) = 0.00135$, so that under these assumptions a false out-of-control signal is quite unlikely.

If an out-of-control signal is generated in Phase I, a search is initiated for a responsible special cause of variation. If this can be found, action should be taken to prevent it from re-occurring. In cases where a special cause is found and removed, the corresponding sample does not provide information about the in-control state of the process. Therefore, in such cases, it should be removed from the data set, and the remaining $k - 1$ subsamples should be compared to re-estimated control limits.

This procedure should be repeated until no out-of-control signals are generated, or when underlying special causes either cannot be found or cannot be removed. At the end of Phase I, we have a data set at our disposal of, say, $m \leq k$ subsamples that provides information concerning the variability that can be attributed to common causes of variation. This information is needed for Phase II, when samples are drawn online.

2.3.3 Phase II: current control

In Phase II, we have an estimate of the in-control distribution available based on m samples from Phase I. This distribution function will be denoted by F_0. Each time a sample $X_{f1}, \cdots, X_{fn}$ becomes available at time $f > k$, we want to test the hypothesis

$$\mathrm{H}_0 : \ F_f = F_0$$

against the alternative

$$\mathrm{H}_1 : \ F_f \neq F_0.$$

The derivation of the control limits for Phase II is analogous to Phase I, with this difference: M_m and V_m are independent of $X_{f1}, \cdots, X_{fn}$. This will result in different constants $c(n, k, p_{\mathrm{LCL}})$ and $c(n, k, p_{\mathrm{UCL}})$ for the LCL and UCL, respectively.

The control charts discussed in this section are Shewhart-type control charts. Their performance under various levels of first-order autocorrelation (including the special case of independence) will be studied in Chapter 3. More efficient control charts are also developed, such as the EWMA control chart and the CUSUM control chart. Their performance for various levels of first-order serial correlation will be discussed in Chapter 4 and Chapter 5, respectively.

2.4 Serial correlation

In the previous sections, we made the assumption that the measurements that are used to monitor the process are independently distributed. This is standard practice in most literature on SPC. In this thesis, it is investigated how the performance of control charts is affected by serial correlation in the observations. This is motivated by many practical situations, where the assumption that the observations are (approximately) uncorrelated cannot be justified, see also Chapter 1. In this section, we will shortly review some of the types of serial correlation that will be considered in subsequent chapters. We will restrict ourselves to serial correlation that can be successfully modelled using ARIMA(p,d,q) models. This class of models is discussed in a large number of texts on time series, such as Box and Jenkins (1976), Pandit and Wu (1983), or Montgomery, Johnson and Gardiner (1990). We refer to such texts for a deeper discussion of time series analysis.

Throughout this thesis, a serially correlated sequence of variables will be denoted by $\{Y_1, \cdots, Y_T\}$, whereas a sequence of independent random variables will be denoted by $\{X_1, \cdots, X_T\}$.

Before we turn to the discussion of ARIMA models, we introduce some of the terminology that is used in time series analysis. This will facilitate the discussion in subsequent chapters. Much of it is taken from Anderson (1976).

A time series model that has proved to be useful in practice is the AR(1) model. This model deserves and will receive most of the attention in subsequent chapters. It is one of the simplest special cases of ARIMA(p,d,q) models. It will be discussed in Subsection 2.4.2. More general ARIMA(p,d,q) models will be discussed in Subsection 2.4.3

2.4.1 Some time series analysis terminology

We define a *time series* as a set of observations that are ordered in time. We will only consider discrete series, with observations drawn at equal

time intervals. A sequence of observations $\{y_1, y_2, \cdots, y_T\}$ can be viewed as a single realization of some underlying stochastic process. Each y_i is the realization of some random variable Y_i, which has an associated probability density function $f_{Y_i}(\cdot)$. We assume that for any set of Y_i's, say $Y_{j_1}, \cdots, Y_{j_r}$, a joint probability density function $f_{Y_{j_1},\cdots,Y_{j_r}}(\cdot)$ exists. If a statistical process is such that $f_{Y_{i+n_1},\cdots,Y_{i+n_m}}(\cdot)$ is independent of i for any positive integer m and for any choice of $n_1, \cdots, n_m$ then the probabilistic structure does not change over time and the process is said to be *strictly stationary*. Otherwise, it is *nonstationary*. If the definition of strict stationarity only holds for $m \leq p$ for some positive integer p, then the process is said to be *stationary of order p*.

A *Gaussian process* is defined by the property that the probability density function associated with any finite subset of $\{\cdots Y_{-2}, Y_{-1}, Y_0, Y_1, Y_2, \cdots\}$ is multivariate normal. For a Gaussian process, a sufficient condition for strict stationarity is stationarity of order 2, since all moments of higher order are then precisely fixed. Stationarity of order 2 is sometimes also called *weak stationarity*.

It is important to make the distinction between (weakly) stationary models and nonstationary models, since observations from the former wander around a fixed mean. Observations from a nonstationary model may wander away from a specified target value if no action is taken. That is, nonstationary processes require some form of *Automated Process Control* (APC) if the quality characteristic should fall between certain specification limits. Stationary processes are mean reverting and need not be controlled by APC techniques. However, the performance of such processes can be improved by applying APC, see Box and Luceño (1997b).

Weak stationarity implies that for all i

$$\mathrm{E}(Y_i) = \mu$$

and

$$\mathrm{Cov}(Y_i, Y_{i-k}) = \gamma_k,$$

where μ is the constant mean of the process, and γ_k, the *autocovariance* at lag k (integer), is also constant. In particular, Y_i has constant variance $\sigma_Y^2 = \gamma_0$.

Furthermore, we have for all integers k,

$$\gamma_{-k} = \gamma_k$$

since

$$\mathrm{Cov}(Y_i, Y_{i+k}) = \mathrm{Cov}(Y_{i+k}, Y_i) = \mathrm{Cov}(Y_i, Y_{i-k}).$$

Hence it is only necessary to consider γ_k for $k > 0$. The set $\{\gamma_0, \gamma_1, \cdots\}$ is sometimes called the *autocovariance function*. We define ρ_k, the *autocorrelation coefficient* at lag k by

$$\rho_k = \frac{\gamma_k}{\gamma_0}.$$

The set $\{\rho_0, \rho_1, \cdots\}$ is called the *AutoCorrelation Function* (ACF) of the process. An estimate of the ACF can be used to identify which models within the class of ARIMA(p,d,q) can be used to model the random behavior in a given set of observations.

2.4.2 The AR(1) process

A process $\{Y_t\}$ is said to be a first-order autoregressive process (AR(1) process) if it is generated by

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}, \tag{2.2}$$

where ϕ is some constant satisfying $\phi \in (-1, 1)$, and $\{\varepsilon_t\}$ is a sequence of i.i.d. disturbances, $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ for $t \in \mathbb{Z}$. Subsequent observations of model (2.2) are serially correlated, since

$$\mathrm{Cov}(Y_t, Y_{t-k}) = \phi^k \sigma_Y^2,$$

where

$$\sigma_Y^2 = \mathrm{Var}(Y_t) = \frac{\sigma_\varepsilon^2}{1 - \phi^2}.$$

Model (2.2) is strictly stationary since $|\phi| < 1$.

Obviously, $\mathrm{E}(Y_t) = \mu$ for all t in model (2.2). A model that includes the possibility of a shift in $\mathrm{E}(Y_t) \equiv \mu_t$ is

$$Y_t - \mu_t = \phi(Y_{t-1} - \mu_{t-1}) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}. \tag{2.3}$$

First-order autoregressive models are useful when disturbances affect not only the current outcome of the process, but also have an (exponentially declining effect) on future outcomes. Such situations occur for example when a tank containing raw material is refilled from time to time with raw material of varying quality, see also the example considered in Subsection 1.1.3.

2.4.3 ARIMA processes

More general AR processes are of the form

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}, \quad (2.4)$$

If we introduce the backward shift operator B, where $BY_t = Y_{t-1}$, then (2.4) can be rewritten as

$$\phi(B)(Y_t - \mu) = \varepsilon_t,$$

where $\phi(B)$ is a polynomial in B of degree p. Autoregressive models of order p are stationary if all roots of the polynomial $\phi(\cdot)$ are outside the unit circle (see Box and Jenkins (1976), Section 3.2).

A *Moving Average* (MA) process of order q is generated by

$$Y_t = \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q}.$$

Moving average models of any order are always stationary.

A useful class of models for time series is formed from a combination of MA and AR processes. A mixed autoregressive moving average process containing p AR terms and q MA terms is abbreviated to an ARMA(p,q) process, and is given by

$$\begin{aligned} Y_t - \mu \;=\; & \phi_1(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) \\ & + \varepsilon_t - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q} \qquad \text{for } t \in \mathbb{Z}, \end{aligned}$$

Or, rewritten using the backward shift operator B:

$$\phi(B)(Y_t - \mu) = \theta(B)\varepsilon_t, \quad (2.5)$$

where $\phi(B)$ and $\theta(B)$ are polynomials of degrees p and q, respectively. Such processes are stationary if the roots of the polynomial $\phi(\cdot)$ lie outside the unit circle. The class of ARMA(p,q) models can be used to model a wide range of stationary time series, with only a few parameters to estimate. In practice, values of p and q larger than 2 are rarely encountered.

However, many time series encountered in practice are nonstationary. In order to fit a stationary model it is necessary to remove the nonstationarity first. In many cases, this can be obtained by taking successive differences of the observations one or more times. That is, in case of first differences,

we consider $\{y_2 - y_1, y_3 - y_2, \cdots, y_T - y_{T-1}\}$ and try to fit a stationary ARMA(p,q) model to these new observations. If taking first differences is not enough to obtain stationarity, the observations are differenced once more, and so on, until stationarity is obtained. Taking differences can be expressed in terms of the backward shift operator as $(1 - B)y_t$. The notation $(1 - B)^d y_t$ is used to indicate that successive differences were taken d times.

If we replace $Y_t - \mu$ in Equation (2.5) by random variables that are differenced d times, we have

$$\phi(B)(1 - B)^d(Y_t - \mu) = \theta(B)\varepsilon_t.$$

Such models are called *AutoRegressive Integrated Moving Average* (ARIMA) models of order (p,d,q). The term 'integrated' is used because the stationary model which is fitted to the differenced data has to be summed or 'integrated' to provide a model for the nonstationary data. ARIMA(p,d,q) models are capable of describing certain types of nonstationary time series. An important special case is the IMA(1,1) model, which is often encountered in practice. In Chapter 8, an example of an IMA(1,1) process is discussed.

The ARIMA(p,d,q) models were popularized by Box and Jenkins (1976). Box and Jenkins (1963) themselves developed one of the first control charts to account for serial correlation. By assuming that the quality characteristic was drifting away from its target value according to an ARIMA(0,1,1) model, they derived a chart with action limits that are determined such that the total costs of running the process are minimized. The cost minimization procedure required trading off the cost of being off-target with the cost of resetting the machine.

2.5 Control charts for serially correlated data

In Section 2.3, it was discussed that the control chart is the tool to detect special causes of variation. The control limits that are drawn on the control chart bound the variation due to common causes of variation. For the proper placement of control limits in the case of serial correlated data, it is of crucial importance to decide which process behaviors are part of the process, and which are attributed to special causes. Gilbert, Kirby and Hild (1997) state the following.

> "*For example, if autoregressive behavior is a normal, unchangeable part of the process, then a chart that gives out-of-control signals because of the presence of autocorrelation is not very useful. On the other hand, if autocorrelation in a process is a symptom of a problem that should be addressed, then the control chart should detect the presence of the autocorrelation.*"

Crowder, Hawkins, Reynolds and Yashchin (1997) share this view. They argue that the *cause* of autocorrelation should be assessed before the data is analyzed and interpreted. As argued in Chapter 1, we will consider cases where autocorrelation in process data is unremovable and part of the process. Control charts should not signal because of autocorrelation, but give out-of-control signals because of the presence of special causes of variation.

Consequently, the definition of an in-control process that is most commonly used in practice and Shewhart's original definition do not necessary agree. In practice, the term 'an in-control process' is more often than not associated with a sequence of independently and identically distributed observations. As was discussed in Subsection 2.2.1, Shewhart's original definition of an in-control process only requires that we can predict (within statistically determined limits) how the process may be expected to vary in the future. A process that exhibits serial correlation *is* predictable. For this reason we extend the definition of an in-control process to include observations which may be serially correlated. Alwan (1988) refers to such processes as being 'in control in a broader sense'.

Chapter 3

Shewhart-type control charts for the mean

3.1 Introduction

In the previous chapter, we discussed the Shewhart control chart in the classical context: for independent univariate observations, and normality of the quality characteristic. In the remainder of this thesis, we will investigate how the performance of control charts is affected by various forms of serial correlation. In this chapter, which is based on Wieringa (1998), we will start with Shewhart-type control charts for the mean of a serially correlated sequence of observations. By Shewhart-type control charts we mean that the statistics that are plotted in the control chart are not smoothed, as in the EWMA control chart, which will be discussed in Chapter 4, or summed, as in the CUSUM control chart, which will be discussed in Chapter 5. We chose to discuss these charts in separate chapters, in order not to confound the efficiency differences between the control charts with the effect of serial correlation.

In this chapter and in the next two chapters, we will only consider charts for detecting a step change in the mean of the observations. In Chapter 7, charts for detecting a shift in the spread of serially correlated data are discussed.

For detecting a shift in the mean of a serially correlated process two approaches are suggested in the literature. The first approach is a modification of the classical Shewhart control chart. The control limits are modified to allow for serial correlation in the data. The modified Shewhart chart is discussed in Section 3.4. The second approach is based on residuals

from a time series model, fitted to the data. The residuals control chart will be discussed in Section 3.5. In Section 3.6 a third approach, the modified residuals chart, is discussed.

In Sections 3.4 through 3.6, we will study how the three charts are affected by first-order autocorrelation in the data. We believe that this is the most important case for practical purposes. In Subsection 2.4.2, the properties of the AR(1) model were quickly reviewed. The control charts that are discussed in this chapter are readily extended to other time series models. In the tables of Appendix A, the results of application of these control charts to various ARMA(p,q) models with $p + q \leq 2$ are presented. We will not discuss these results in detail in the text.

Throughout this thesis, the performances of different control charts are compared on the basis of *Average Run Length* (ARL) curves. Considering only the average of the run length has disadvantages, since the distribution function of the run length is typically skewed to the right. It is questionable whether the mean provides enough information to characterize the distribution of the run length. Therefore, it is sometimes recommended to study quantiles of the run length distribution, or the run length distribution itself. However, since the ARL curve is still most commonly used to evaluate the efficiency of a control chart, it will be the basis for the comparison of control charts throughout this thesis.

The ARL curve for the i.i.d. case is discussed in Section 3.2. This curve will serve as a reference point for subsequent sections, when the effect of serial correlation is studied.

3.2 ARL curve for the i.i.d. case

In the classical situation, it is assumed that subsequent observations of a quality characteristic are independently distributed. Throughout this thesis, a sequence of independent observations is denoted by $\{X_t\}$. An observation may be the mean of a sample, or an individual observation. In this chapter, it will assumed that the variance of the observations remains constant over time at level σ_X^2, and that a special cause of variation may cause a shift in the mean of the process. Assuming normality, we have the following model for X_t, an observation of a quality characteristic of interest at time t

$$X_t \sim \mathcal{N}(\mu_t, \sigma_X^2) \qquad \text{for } t \in \mathbb{Z}, \tag{3.1}$$

where successive X_t's are assumed to be independently distributed. The expectation of X_t is indexed by time, to indicate that the mean of the process may shift due to special causes of variation.

A special cause occurring at an unknown time point T is modelled as

$$\mu_t = \begin{cases} \mu & \text{for } t < T \\ \mu + \delta\sigma_X & \text{for } t \geq T. \end{cases} \tag{3.2}$$

The size of the shift is expressed in units of the standard deviation of the observations. This facilitates comparison of control charts for processes with different variances.

The classical Shewhart control chart has control limits $\mu \pm 3\sigma_X$, (see Subsection 2.3.2). After a shift of size $\delta\sigma_X$ has occurred, we can write for $\mathrm{P}(\delta)$, the probability that an observation falls within the control limits

$$\begin{aligned} \mathrm{P}(\delta) &= \mathrm{P}(\mu - 3\sigma_X \leq X_t \leq \mu + 3\sigma_X) \\ &= \Phi(\delta + 3) - \Phi(\delta - 3), \end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Since $\mathrm{P}(0) = 0.9973$, it is very unlikely to observe an out-of-control signal if $\delta = 0$. Therefore, if an observation outside the control limits is encountered, the presence of a special cause of variation is suspected. This interpretation of an out-of-control signal relies on the independence and normality assumptions made for model (3.1).

The *Average Run Length* (ARL) is defined as the average number of observations up to and including the first out-of-control observation. The ARL is a function of δ. In the present case of independent observations, $\mathrm{ARL}(\delta)$ can be computed as

$$\mathrm{ARL}(\delta) = \sum_{i=1}^{\infty} i\mathrm{P}(\delta)^{i-1}[1 - \mathrm{P}(\delta)] = \frac{1}{1 - \mathrm{P}(\delta)}. \tag{3.3}$$

In Figure 3.1, the ARL curve of a Shewhart chart with three-sigma limits is depicted.

Figure 3.1 shows that the ARL is high if $\delta = 0$, and that the ARL is low if δ is large. This is desirable behavior. A control chart having the same $\mathrm{ARL}(0)$, and lower $\mathrm{ARL}(\delta)$ for $\delta > 0$ is more efficient, since on average

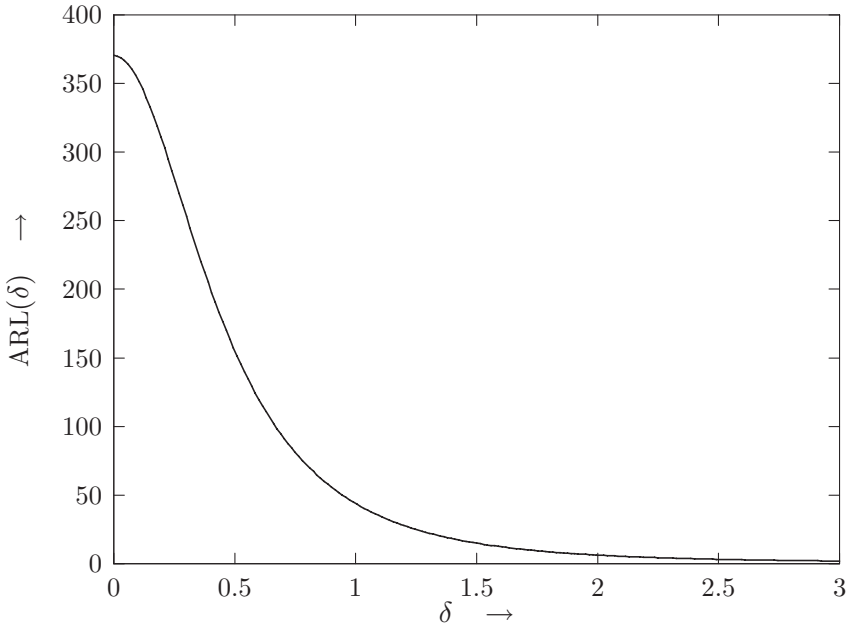


Figure 3.1: ARL curve of a Shewhart chart.

fewer observations are needed to detect a change in $\mathrm{E}(X_t)$. Analogously, a control chart having the same ARL(0) and higher ARL(δ) for $\delta > 0$ is less efficient. The ARL curves of control charts discussed in the remainder of this chapter will be compared to the curve in Figure 3.1.

3.3 Effect of ignoring serial correlation

In this section, we will investigate how the ARL curve of the standard Shewhart chart is affected by AR(1) dependence in the observations. In practice, when the presence of serial correlation in the data is not noticed or when it is ignored, an SPC practitioner will set up a control chart as if the data were independent. Since we are concerned with individual observations in this chapter, we will follow the most common design for a control chart for the mean of individual independently distributed observations.

Using the notation of Chapter 2, we assume that observations from a process of interest are generated by the AR(1) model (2.2):

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}, \tag{3.4}$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d. disturbances, $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ for $t \in \mathbb{Z}$. In the simulation study of this section that will be discussed next, we will use $\sigma_\varepsilon = 1$.

For σ_Y^2, the variance of the AR(1) observations, we have

$$\sigma_Y^2 = \mathrm{Var}(Y_t) = \frac{\sigma_\varepsilon^2}{1-\phi^2}.$$

Suppose that we have realizations of $\{Y_1, \cdots, Y_n\}$ at our disposal to set up a control chart for the mean. It is customary to estimate the center line of the control chart by the overall mean. This is an unbiased estimator. For the width of the control limits, the standard deviation of the process is usually estimated by $\overline{MR}/d_2(2)$, the mean of the moving ranges corrected by a factor $d_2(2)$. The mean of the moving ranges is defined as

$$\overline{MR} = \frac{1}{n-1}\sum_{i=2}^{n} MR_i,$$

where

$$MR_i = |Y_i - Y_{i-1}|.$$

The factor $d_2(2)$ is commonly used in the literature on SPC, and equals $2/\sqrt{\pi}$ (see Subsection 7.1.2 of this thesis for a derivation).

If the Y_t's would have been independent, $\overline{MR}/d_2(2)$ would have been an unbiased estimator for σ_Y. However, in Subsection 7.1.2, it is shown that $\mathrm{E}\left[\overline{MR}/d_2(2)\right] = \sqrt{1-\phi}\,\sigma_Y$, so that the usual estimator for the standard deviation is biased in case of AR(1) observations. Consequently, the control limits are not correctly computed. To observe how this affects the ARL curve of the Shewhart control chart with three-sigma limits, we simulated a sequence of AR(1) observations without a shift in the mean (the in-control situation), and judged these with control limits $\mu \pm 3\mathrm{E}[\overline{MR}/d_2(2)]$. We repeated this simulation 100,000 times for seven different values of the AR-parameter: $\phi = -0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9$. The mean of the 100,000 in-control run lengths can be found in the fourth column of Table 3.1, for each ϕ. The bracketed numbers below the simulated ARL values are the simulated standard errors of the averages. Subsequently, we repeated the simulations for the situation that a shift in the mean of size $1\,\sigma_Y$ has occurred. The average and the standard error of the average of 100,000 simulated run lengths are tabulated in the fifth column of Table 3.1. In

Table 3.1: ARL values of a three-sigma Shewhart chart, applied to various AR(1) processes.

ϕ	σ_Y	$\mathrm{E}\left[\overline{MR}/d_2(2)\right]$	ARL($0\,\sigma_Y$)	ARL($1\,\sigma_Y$)
-0.9	2.2942	3.1623	45179.41 (142.59)	1743.99 (5.53)
-0.6	1.2500	1.5811	7125.77 (22.57)	396.39 (1.25)
-0.3	1.0483	1.1952	1604.25 (5.06)	128.71 (0.41)
0.0	1.0000	1.0000	370.22 (1.16)	43.84 (0.14)
0.3	1.0483	0.8771	85.60 (0.27)	17.28 (0.05)
0.6	1.2500	0.7906	22.84 (0.07)	8.60 (0.03)
0.9	2.2942	0.7255	8.15 (0.02)	3.86 (0.02)

Table 3.1, the row corresponding to $\phi = 0$ represents the results for the case of independent observations. The simulated ARL values agree closely with the curve of Figure 3.1.

In the second column of Table 3.1, the standard deviation of the AR(1) observations is tabulated. The entries in the third column are the expectations of $\overline{MR}/d_2(2)$. The results illustrate that σ_Y is overestimated for $\phi < 0$, and underestimated for $\phi > 0$. This can be interpreted as follows.

In case of autocorrelated data, the short-term variation, which is represented by $\mathrm{E}\left[\overline{MR}/d_2(2)\right]$, is different from the long-term variation in the process, represented by σ_Y. This means that the width of control limits, which is based on (an estimate of) short-term variation, does not comply with the variation that is inherent in the process. In the case of positive autocorrelation, which is most commonly encountered in practice (see Faltin, Mastrangelo, Runger and Ryan (1997)), this will result in limits that are too tight. This explains the occurrence of a large number of false out-of-control signals in the case of positive autocorrelation.

In case of negative autocorrelation, the control limits are too wide, due to overestimation of σ_Y. Hence, the chart will be insensitive to changes in the mean of the process.

In this thesis, we acknowledge that biased estimation of the variation in serially correlated processes is a serious problem when constructing control charts. This will result in misinterpretation of out-of-control signals. Positive autocorrelation increases the number of false signals, compared to the case of monitoring uncorrelated observations. In case of negative autocorrelation, the control chart will not be sensitive enough to detect process upsets, so that the presence of special causes may not be detected.

However, as soon as the presence of serial correlation is established, it is relatively easy to obtain an unbiased estimate of the variance of the process. This can be done either by multiplying the estimate by a factor which eliminates the bias, or by estimating the variance of the residuals, and multiplying this by an appropriate factor to obtain an estimate of the process variance. See also Brockwell and Davis (1991). The second column of Table 3.1 illustrates that the variance of the autocorrelated observations is larger, compared to the case where autocorrelation is absent. Therefore, a control chart which accounts for serial correlation will generally have wider control limits than a control chart for uncorrelated data would have.

Nevertheless, the question remains whether a sequence of serially correlated observations with standard deviation σ_Y is properly monitored with a control chart with $3\sigma_Y$ limits. After all, such a control chart is intended to monitor an *independent* sequence of observations with standard deviation σ_Y. We will concern ourselves with the question what effect the dependence structure has on the behavior of the ARL curve.

This is the fundamental question when dealing with serially correlated data, since in our view, the use of a proper (estimate of) the process standard deviation to determine the width of the control limits a is minimal requirement for *any* control chart.

In the remaining sections of this chapter, we will consider control charts for the mean that account for serial correlation. We will assume that the underlying process model is correctly identified, and that its parameters σ_ε and ϕ (and hence σ_Y) are known. This is not a serious limitation in practice, since applications where serial correlation is an issue typically arise in situations where there is a high frequency of sampling. There is usually enough data available to identify and estimate the model with a high degree of accuracy.

3.4 The modified Shewhart control chart

In this section, we will study the ARL behavior of a control chart with control limits that are modified to account for serial correlation in the observations. This approach is also discussed in Vasilopoulos and Stamboulis (1978), Schmid (1995b), Kramer and Schmid (1996a), and Lu and Reynolds (1997). The word 'modified' refers to two adaptations compared to the control chart for i.i.d. observations.

3.4.1 Shewhart limits, modified for AR(1) data

In the i.i.d. case, the present observation is only affected by current causes of variation. There is no difference between short-term variation and long-term variation. When the data is dependent, past and present causes of variation have their effect on the current observation. Furthermore, as noted in the previous section, short-term variation is different from long-term variation in the process. Proper control limits for such processes are based on the long-term variation.

This is a first adaptation of the control chart, since it is usually recommended to use a short-term estimator of the standard deviation for independent individual data, to resemble the use of a within-sample estimator for the variation in the case where the sample size is larger than one. A second adaptation follows from the ARL properties of the first adaptation. Hence, it will be discussed after we have evaluated the ARL.

For evaluating the ARL of the modified Shewhart chart, the Markov-chain approach can be used (see Brook and Evans (1972)). However, we will use the Fredholm-integral approach presented by Crowder (1987). Suppose that the sequence $\{Y_t\}$ is generated by model (2.3):

$$Y_t - \mu_t = \phi(Y_{t-1} - \mu_{t-1}) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}.$$

where ϕ is some constant satisfying $\phi \in (-1, 1)$, and $\{\varepsilon_t\}$ is a sequence of i.i.d. disturbances, $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ for $t \in \mathbb{Z}$. Furthermore, assume that on time T, the mean of $\{Y_t\}$ shifts from μ to $\mu + \delta\sigma_Y$. Consequently, for $t \geq T + 1$, the observations of Y_t are generated by

$$(Y_t - \mu) - \delta\sigma_Y = \phi\,(Y_{t-1} - \mu) - \phi\delta\sigma_Y + \varepsilon_t$$

If the value of Y_{t-1} is s, we can write for v, the realization of Y_t

$$\begin{aligned} v &= \mu + \delta\sigma_Y + \phi\left(s - (\mu + \delta\sigma_Y)\right) + \varepsilon_t \\ &= \phi s + (1-\phi)\mu + (1-\phi)\delta\sigma_Y + \varepsilon_t, \end{aligned}$$

provided $t \geq T+1$. The observations are compared to control limits $\mu \pm 3\sigma_Y$. The run length is one if the next observation v falls outside the control limits. If v is within the control limits, the run length is one *plus* some additional observations, which can be regarded as a run length of the AR(1) process, starting in v. Let $L_\phi(\delta, s)$ denote the ARL of the modified Shewhart control chart as a function of δ and s, the value of the AR(1) process at the first observation. The latter is of importance since the observations are serially correlated. The function also depends on the AR parameter ϕ. We can write the following for $L_\phi(\delta, s)$:

$$\begin{aligned} L_\phi(\delta, s) &= 1 + \int_{\{\varepsilon|\ \mu-3\sigma_Y \leq v \leq \mu+3\sigma_Y\}} L_\phi\left(\delta, v\right) f(\varepsilon)\ \mathrm{d}\varepsilon \\ &= 1 + \int_{\mu-3\sigma_Y}^{\mu+3\sigma_Y} L_\phi\left(\delta, v\right) f\left[v - \phi s - (1-\phi)(\mu + \delta\sigma_Y)\right]\ \mathrm{d}v, \end{aligned}$$

where $f(\varepsilon)$ is the density function of ε. This integral equation is a so-called *Fredholm integral* of the second kind. The unknown function $L_\phi(\delta, s)$ can be numerically evaluated using for example Gaussian Quadrature, see Appendix B.

Note that for $\phi = 0$, $L_\phi(\delta, s)$ does not depend on s. This makes sense because the observations are independent if $\phi = 0$. From the integral equation above it follows that $L_0(\delta, s) = \mathrm{ARL}(\delta)$ for all $s \in (\mu - 3\sigma_Y, \mu + 3\sigma_Y)$, as expected.

The function $L_\phi(\delta, s)$ is the ARL function of the modified Shewhart chart if we start to take observations *after* the shift has occurred. In practice, we are often interested in how quickly on average a shift is detected starting from the moment that the shift occurs. The computation of this ARL function is slightly different. If it is assumed that the value of Y_{T-1} is s, we can write for v^*, the realization of Y_T

$$\begin{aligned} v^* &= \mu + \delta\sigma_Y + \phi\left(s - \mu\right) + \varepsilon_T \\ &= \phi s + (1-\phi)\mu + \delta\sigma_Y + \varepsilon_T. \end{aligned}$$

We will denote the ARL of a modified Shewhart chart when the first observation is taken at the time of the shift by $L^*_\phi(\delta, s)$. This function is related to $L_\phi(\delta, s)$ as follows

$$L_\phi^*(\delta, s) \;=\; 1 \;+\; \int_{\mu-3\sigma_Y}^{\mu+3\sigma_Y} L_\phi\,(\delta, v^*)\, f[v^* - \phi s - (1-\phi)\mu - \delta\sigma_Y]\;\mathrm{d}v^*.$$

Numerical evaluation of $L_\phi^*(\delta, s)$ and $L_\phi(\delta, s)$ shows that the difference between the functions is small. For negative values of ϕ, $L_\phi^*(\delta, s) \geq L_\phi(\delta, s)$, while for positive ϕ, $L_\phi^*(\delta, s) \leq L_\phi(\delta, s)$.

In Figure 3.2 $L_\phi^*(\delta, 0)$ curves are drawn for $\phi = -0.6, 0, 0.6$ as functions of δ. The curve for $\phi = 0$ is identical to the curve depicted in Figure 3.1.

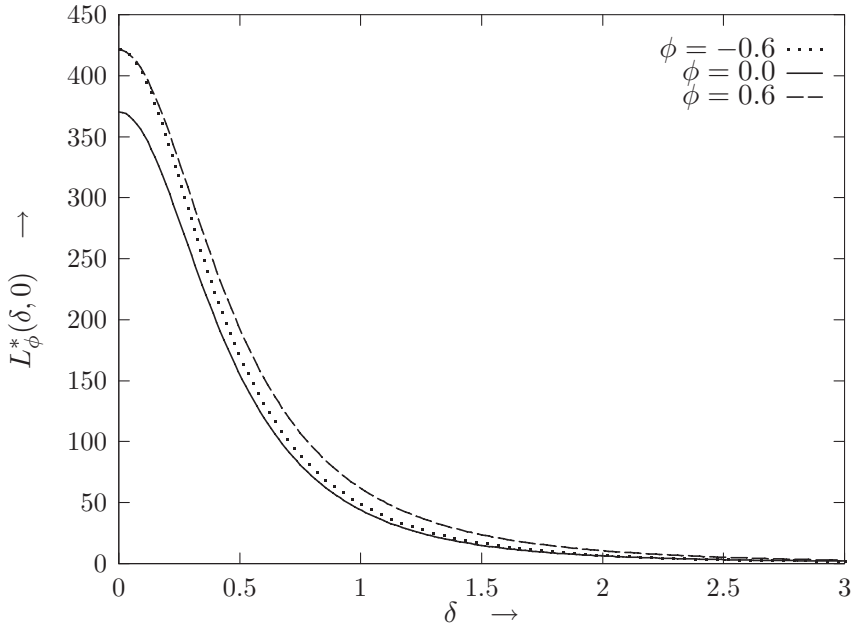


Figure 3.2: ARL curves of the modified Shewhart chart for various AR(1) processes.

The ARL curves in figure 3.2 show the result of a simple adaptation of the Shewhart chart. The correlation is basically ignored. The width of the control limits is based on σ_Y, which is the standard deviation of the observations, and not on σ_ε, the standard deviation of the disturbances. Figure 3.2 shows that, compared to the Shewhart chart for i.i.d. observations, the adapted control chart has a higher in-control ARL. This agrees with a result in Schmid (1995b), where it was proved that in-control ARL values for arbitrary Gaussian processes are always larger than the in-control ARL for i.i.d. observations. This effect is advantageous. For $\delta > 0$, the ARL curves are adversely affected by first-order autocorrelation. Compared to

the i.i.d. case, the adapted control chart is less sensitive for detecting shifts in the mean.

The curves drawn in Figure 3.2 show the typical behavior of $L^*_\phi(\delta, 0)$ for non-zero ϕ in the sense that a non-zero ϕ results in a larger $L^*_\phi(0, 0)$, but also in larger a $L^*_\phi(\delta, 0)$ for $\delta > 0$.

Whether the net effect of first-order autocorrelation is beneficial or not is not clear from Figure 3.2. To make a proper evaluation, a second adaptation of the control chart is needed. The adaptation consists of tightening the limits of the adapted Shewhart chart in such a way that $L^*_\phi(0, 0) \approx \text{ARL}(0)$.

The resulting ARL curves for $\phi = -0.9$ and $\phi = 0$ are depicted in Figure 3.3. ARL curves of intermediate values of ϕ were not drawn, since there is no visible difference with the curve of $\phi = 0$ (this can be checked in Figures 3.15 and 3.16). The curves for various $\phi \geq 0$ are drawn in Figure 3.4.

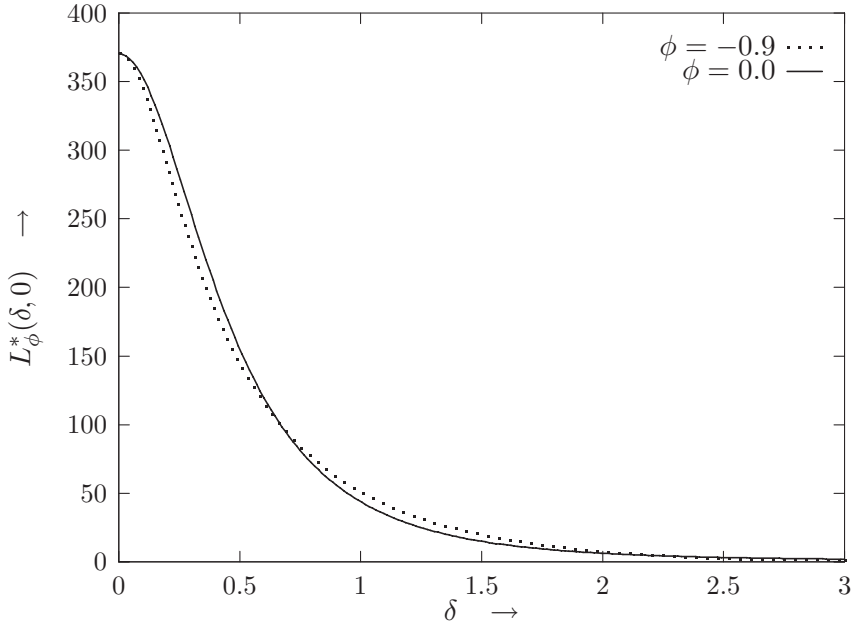


Figure 3.3: $L^*_\phi(\delta, 0)$ as a function of δ for $\phi = -0.9$ and $\phi = 0$.

All but one of the ARL curves drawn in Figures 3.3 and 3.4 are close together. We stated that, for negative ϕ, the ARL curves practically coincide with the curve corresponding to the i.i.d. case. For positive values of ϕ, the modified Shewhart chart is performing worse. However, for $\phi = 0.3$ and $\phi = 0.6$ the differences with the i.i.d. case are small. For the values

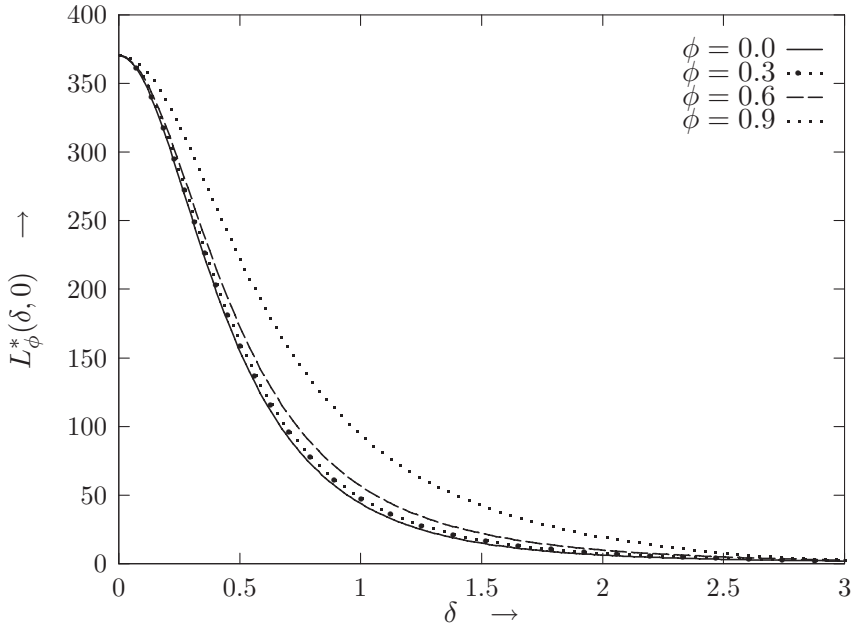


Figure 3.4: $L^*_\phi(\delta, 0)$ as a function of δ for various $\phi \geq 0$.

of ϕ considered, the modified Shewhart chart behaves considerably worse only for $\phi = 0.9$.

In conclusion, if one monitors AR(1) data for a shift in the mean with the aid of a modified Shewhart control chart, then the ARL performance is comparable to a Shewhart chart for independent observations, provided ϕ is not too large.

3.4.2 Discussion

In the previous subsections we argued that the control limits of the modified Shewhart chart for dependent data must be tightened to obtain a certain in-control ARL. We observed that negative first-order autoregressive serial correlation has virtually no impact on the ARL behavior, compared to the ARL curve of the corresponding control chart for independent observations. However, for positive correlation, the ARL behavior is worse, compared to the i.i.d.-case. In this subsection, we investigate why this is so. In addition, we will discuss how the control limits are determined if a certain in-control ARL is given.

In this section we considered $\{Y_t\}$, a sequence of AR(1) observations as

defined in Subsection 2.4.2:

$$Y_t - \mu_t = \phi(Y_{t-1} - \mu_{t-1}) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}.$$

where μ_t is the mean of the process at time t. It is assumed that $\{\mu_t\}$ behaves as

$$\mu_t = \begin{cases} \mu & \text{for } t < T \\ \mu + \delta\sigma_Y & \text{for } t \geq T, \end{cases}$$

where $\mu_t = \mu$ represents an in-control situation, and $\mu_t = \mu + \delta\sigma_Y$ represents an out-of-control situation, where a special cause of variation caused a step change in the mean from an unknown time point T and onwards. Without loss of generality, it is assumed that $\mu = 0$. For the explanation of the ARL behavior of the modified schemes it is helpful to consider $\mathrm{P}_\phi(\delta, s)$, which is defined as the probability of an out-of-control signal at the next observation, given that the current realization of the AR(1) process with parameter ϕ equals s, and that the shift in the mean is $\delta\sigma_Y$. For the modified Shewhart chart, this probability can be written as

$$\begin{aligned} \mathrm{P}_\phi(\delta, s) &= \mathrm{P}_\phi(|\phi s + (1-\phi)\delta\sigma_Y + \varepsilon_t| > h_\phi) \\ &= 1 - \Phi\left[h_\phi - \phi s - (1-\phi)\delta\sigma_Y\right] \\ &\qquad + \Phi\left[-h_\phi - \phi s - (1-\phi)\delta\sigma_Y\right], \end{aligned}$$

where $-h_\phi$ and h_ϕ denote the value of the control limits of the modified Shewhart chart, adjusted for AR(1) correlation with AR-parameter ϕ.

The ARL of the modified Shewhart control chart is closely related to $\mathrm{P}_\phi(\delta, s)$. For $\phi = 0$, the modified Shewhart chart is a standard Shewhart chart for the mean of independent observations. In this case, the ARL is directly related to $\mathrm{P}_\phi(\delta, s)$, see Equation (3.3). If $\phi \neq 0$, the relation between $L_\phi(\delta, s)$ and $\mathrm{P}_\phi(\delta, s)$ is more complicated. Previously, we derived the ARL of the modified Shewhart chart for $t \geq T + 1$ as

$$L_\phi(\delta, s) = 1 + \int_{-h_\phi}^{h_\phi} L_\phi(\delta, v)\, f\left[v - \phi s - (1-\phi)\delta\sigma_Y\right] \mathrm{d}v,$$

where v is the next AR(1) observation. In Appendix B, it is shown that the sequence of functions $\{L_0, L_1, \cdots\}$ converges uniformly to $L_\phi(\delta, s)$, where

$$L_i(\delta, s) = \quad 1 \quad + \quad \int_{-h_\phi}^{h_\phi} L_{i-1}\,(\delta, v)\, f\,[v - \phi s - (1-\phi)\delta\sigma_Y]\ \mathrm{d}v,$$

and L_0 is an arbitrary continuous function on $(-h_\phi, h_\phi)$. Suppose that we take $L_0(\delta, s) \equiv 1$, then we have

$$\begin{aligned} L_1(\delta, s) &= 1 + \int_{-h_\phi}^{h_\phi} f\,[v - \phi s - (1-\phi)\delta\sigma_Y]\ \mathrm{d}v \\ &= 2 - \mathrm{P}_\phi(\delta, s), \end{aligned}$$

and

$$\begin{aligned} L_2(\delta, s) &= 1 + 2\mathrm{P}_\phi(\delta, s) + \\ &\quad - \int_{-h_\phi}^{h_\phi} \mathrm{P}_\phi(\delta, v) f\,[v - \phi s - (1-\phi)\delta\sigma_Y]\ \mathrm{d}v. \end{aligned}$$

and so on. For every $n = 1, 2, \cdots$, the probability $\mathrm{P}_\phi(\delta, s)$ is part of $L_n(\delta, s)$. The behavior of $L_\phi(\delta, s)$ is therefore partly determined by $\mathrm{P}_\phi(\delta, s)$. Note that for $\phi = 0$, the ARL only depends on $\mathrm{P}_\phi(\delta, s)$.

Let us therefore consider $\mathrm{P}_\phi(\delta, s)$ for various values of ϕ. In Figures 3.5 through 3.11, $\mathrm{P}_\phi(\delta, s)$ is drawn as a function of the current realization $s \in (-h_\phi, h_\phi)$, and δ, the size of the shift in the mean expressed in units of σ_Y, for $\phi = -0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9$.

In Figure 3.8, the probability of an out-of-control signal at the next observation does not depend on s, the current realization of the process. As a result of ϕ being zero, subsequent observations are independent, and previous realizations do not affect the probability of future out-of-control signals. In case of negative ϕ and a positive shift in the mean (Figures 3.5, 3.6, and 3.7), values of s below $\mu = 0$ increase the probability of an out-of-control signal at the next observation. As a result of the negative correlation, a current realization below the mean is most likely to be followed by an observation larger than $\mu = 0$. Together with the positive shift in the mean this increases the probability of an out-of-control signal at the next observation. If the current realization is larger than $\mu = 0$, the positive shift and the high probability that the next observation is below the mean counteract each other, resulting in a small probability of an out-of-control signal.

The 'tips' around $(\delta, s) = (0, h_\phi)$ which get more pronounced for $\phi \downarrow -1$, show a surprising increase in $\mathrm{P}(\delta, s)$. For these values of (δ, s) we observe

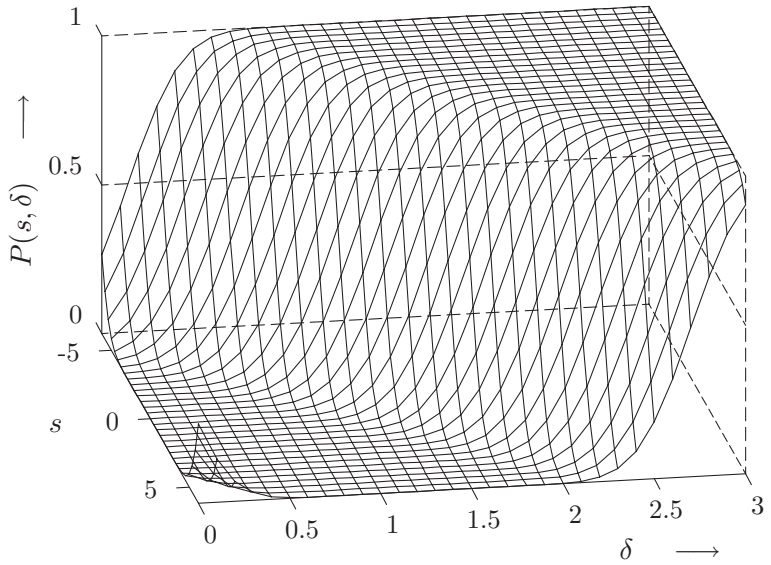


Figure 3.5: Probability of an out-of-control signal at the next AR(1) observation as a function of the shift in the mean δ and the current value $s \in (-h_\phi, h_\phi)$ for $\phi = -0.9$.

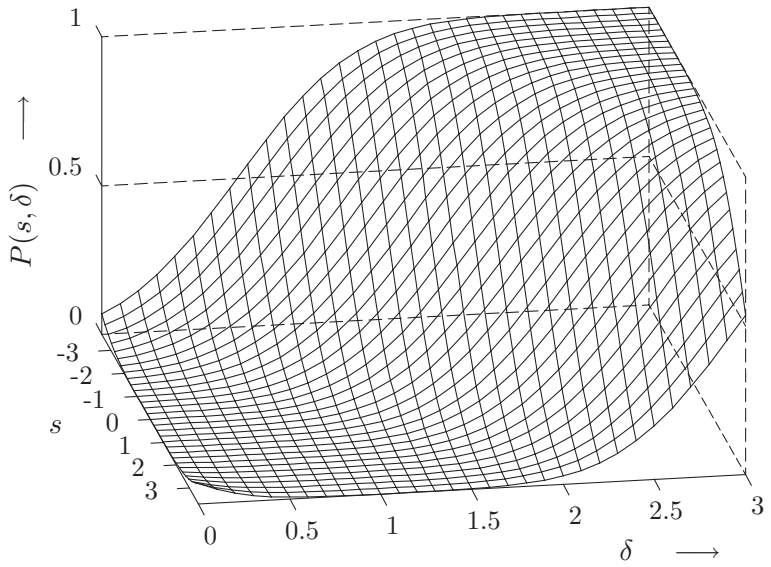


Figure 3.6: Probability of an out-of-control signal at the next AR(1) observation as a function of the shift in the mean δ and the current value $s \in (-h_\phi, h_\phi)$ for $\phi = -0.6$.

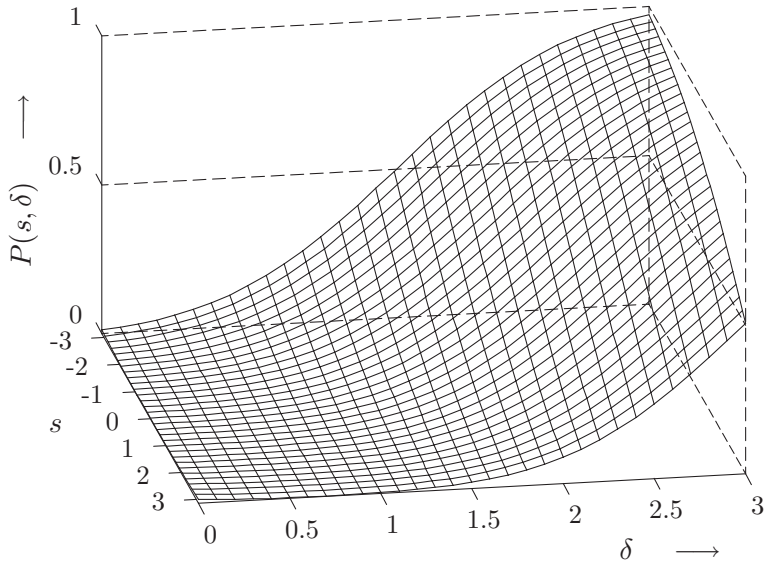


Figure 3.7: Probability of an out-of-control signal at the next AR(1) observation as a function of the shift in the mean δ and the current value $s \in (-h_\phi, h_\phi)$ for $\phi = -0.3$.

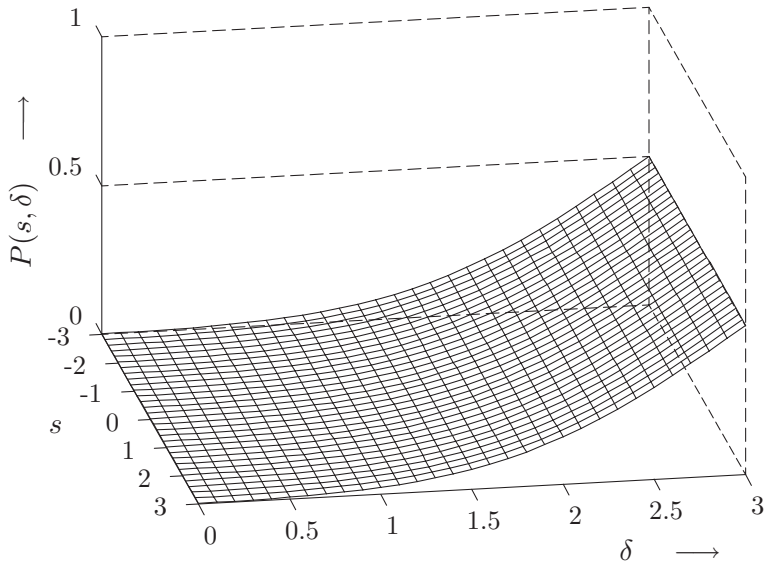


Figure 3.8: Probability of an out-of-control signal at the next AR(1) observation as a function of the shift in the mean δ and the current value $s \in (-h_\phi, h_\phi)$ for $\phi = 0.0$.

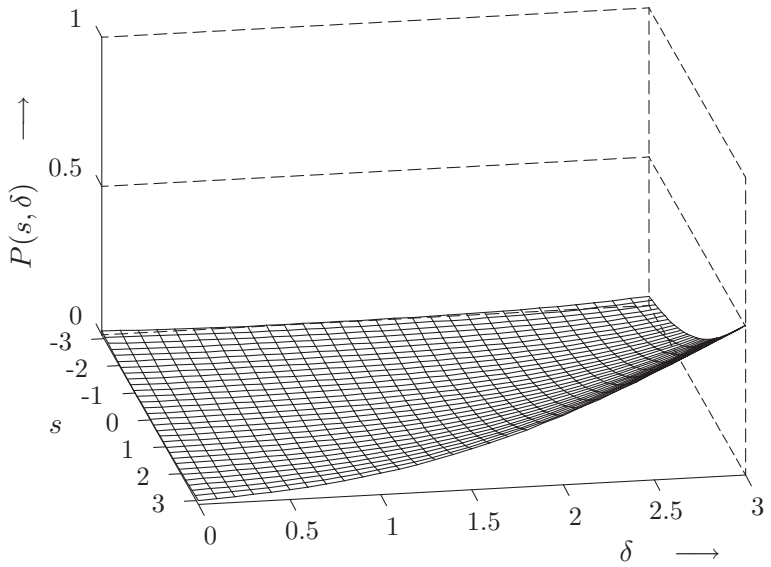


Figure 3.9: Probability of an out-of-control signal at the next AR(1) observation as a function of the shift in the mean δ and the current value $s \in (-h_\phi, h_\phi)$ for $\phi = 0.3$.

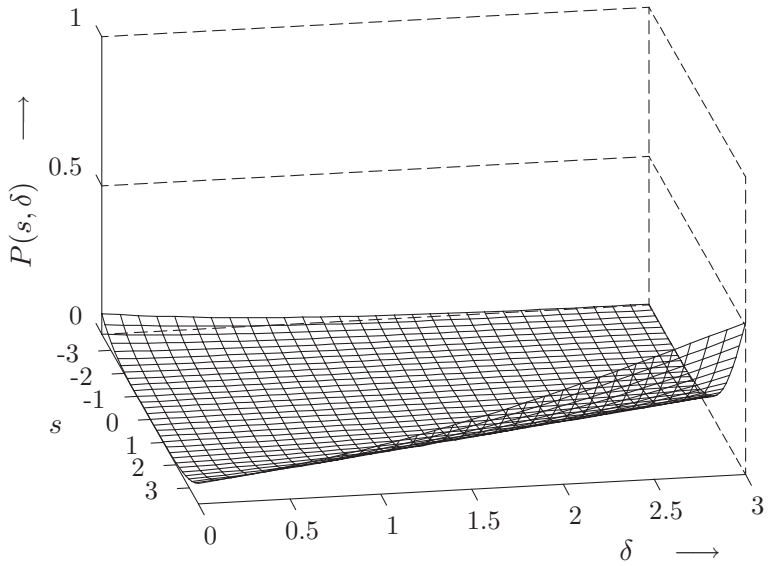


Figure 3.10: Probability of an out-of-control signal at the next AR(1) observation as a function of the shift in the mean δ and the current value $s \in (-h_\phi, h_\phi)$ for $\phi = 0.6$.

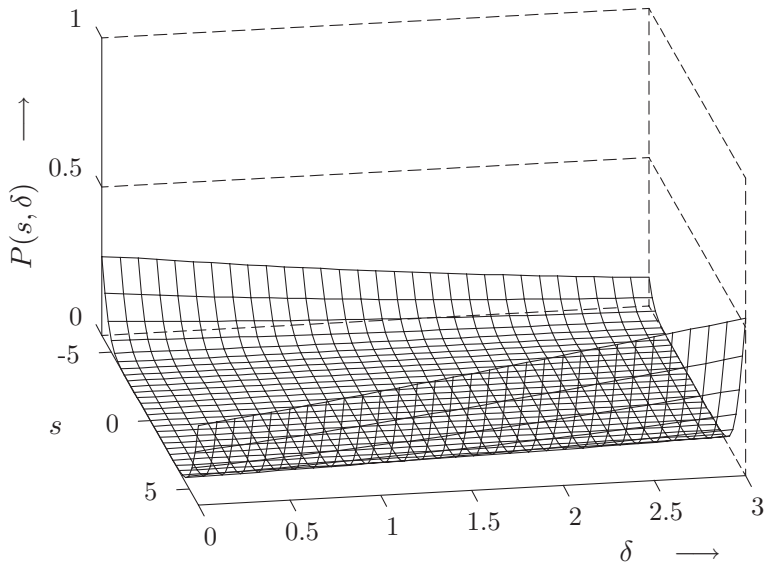


Figure 3.11: Probability of an out-of-control signal at the next AR(1) observation as a function of the shift in the mean δ and the current value $s \in (-h_\phi, h_\phi)$ for $\phi = 0.9$.

that the probability of an out-of-control signal at the next observation decreases if δ is increased! This can be explained as follows

$$\begin{aligned} \mathrm{P}_\phi(\delta, s) &= 1 - \Phi\left[h_\phi - \phi s - (1-\phi)\delta\sigma_Y\right] \\ &\quad + \Phi\left[-h_\phi - \phi s - (1-\phi)\delta\sigma_Y\right] \\ &\longrightarrow \Phi[0] = \frac{1}{2} \qquad (s \uparrow h_\phi,\ \delta \downarrow 0, \phi \downarrow -1). \end{aligned}$$

For a value of $s \approx h_\phi$, the next value is likely to be close to $-h_\phi$ if the AR-parameter is close to -1. Increasing δ then lowers the probability that the next observation is smaller than $-h_\phi$.

For positive ϕ, we observe that $\mathrm{P}_\phi(\delta, s)$ is increasing far less with δ compared to the case $\phi \leq 0$. The power of the control limits to discriminate between small and high values of δ diminishes. As a result, the decrease of the ARL curve from its in-control value to values for $\delta > 0$ is not as steep for $\phi > 0$, compared to the case $\phi \leq 0$.

The reason why $\mathrm{P}_\phi(\delta, s)$ is increasing far less with δ for $\phi > 0$ is the following. The shift in the mean enters the expression of $\mathrm{P}_\phi(\delta, s)$ through

the term

$$(1-\phi)\delta\sigma_Y = \sqrt{\frac{1-\phi}{1+\phi}}\sigma_\varepsilon\delta.$$

The factor $\sqrt{(1-\phi)/(1+\phi)} \to \infty$ for $\phi \downarrow -1$ and is strictly decreasing to zero for $\phi = 1$. For $\phi \to 0$, $\sqrt{(1-\phi)/(1+\phi)} \to 1$. Hence, a shift in the mean expressed in units of the standard deviation of the process under consideration is harder to detect for AR(1) processes with $\phi > 0$ compared to cases where $\phi \leq 0$. The effect of first-order autoregressive serial correlation on the ARL curve of the modified Shewhart chart is clearly not symmetric in the parameter ϕ.

Determining the control limits

In the previous subsections we argued that the control limits of the modified Shewhart chart for dependent data must be tightened to obtain a certain desired in-control ARL. The question now arises how much the in-control ARL is affected by first-order autocorrelation if the control limits are $\mu \pm 3\sigma_Y$. In Figure 3.12, $L^*_\phi(0,0)$ is drawn against values of $\phi \in (0,1)$. Only positive values of ϕ are considered since the curve is symmetric in ϕ.

From Figure 3.12, we conclude that $L^*_\phi(0,0)$ is not much affected if ϕ is not too large. For $|\phi| < 0.6$, the in-control ARL does not differ much from the in-control ARL for independent observations. For $\phi > 0.6$, the in-control ARL rises considerably. This conclusion agrees with the findings in Kramer and Schmid (1996a), where a similar graph was drawn.

In Figure 3.13, it is visualized how to modify the control limits in order to obtain an in-control ARL of $L^*_\phi(0,0) = 370.40$. For the i.i.d. case, the multiplication factor is 3. For $\phi \neq 0$, the multiplication factor needs to be lowered to compensate for the increase in the in-control ARL, induced by the presence of serial correlation. Again, the curve is symmetric in ϕ, so that only positive values of ϕ are considered.

From Figures 3.12 and 3.13, we conclude that using $3\sigma_Y$ limits for data with small to moderate first-order autocorrelation does not lead to severe misinterpretation. However, for values of $\phi > 0.6$, the control limits must be tightened.

In the light of the foregoing, the conclusion that small to moderate levels of first-order autocorrelation may be safely ignored, may force itself upon us. In Wheeler (1989) it was concluded:

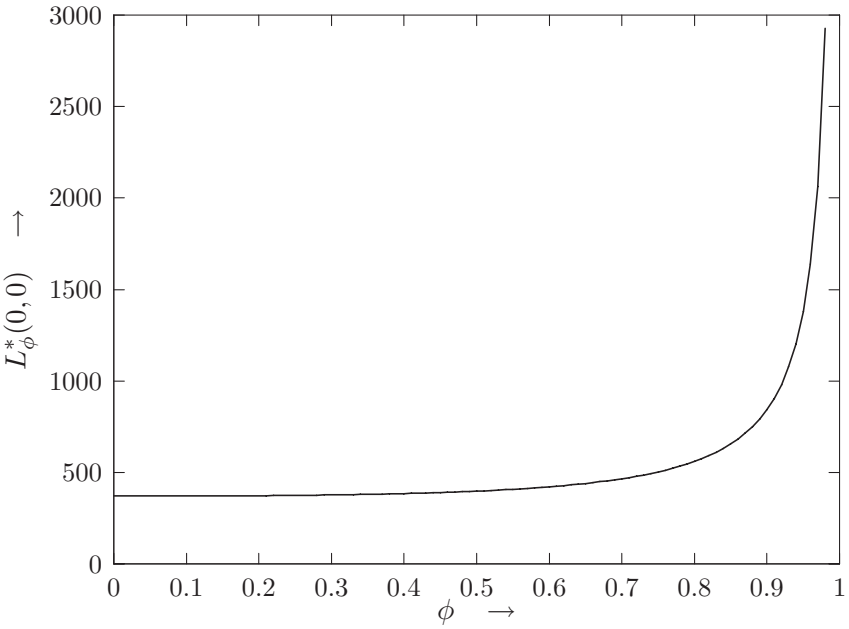


Figure 3.12: The behavior of $L^*_\phi(0,0)$ against ϕ.

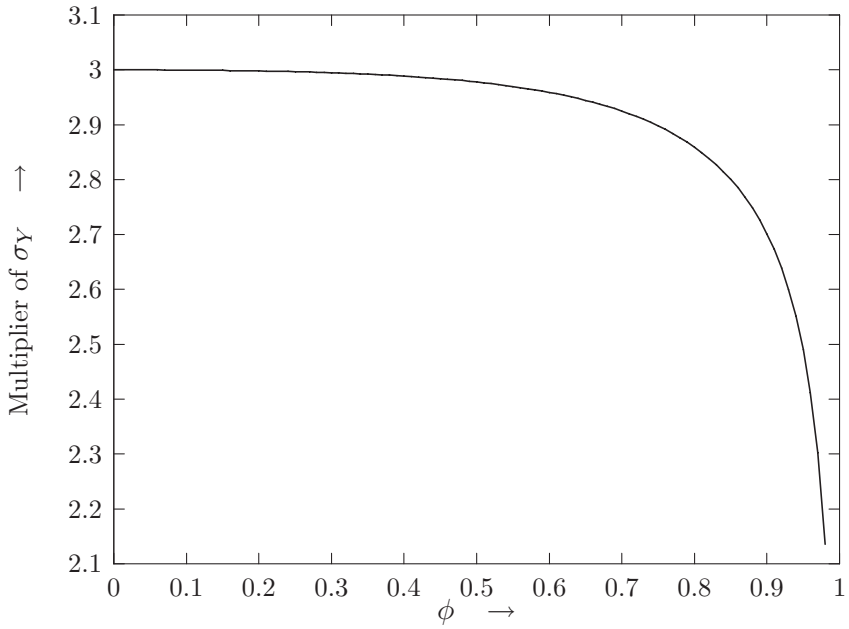


Figure 3.13: Control limit multiply factor of σ_Y against ϕ, when the in-control ARL is fixed at $L^*_\phi(0,0) = 370.40$.

> "*For most situations, 'significant' autocorrelations will have a minimal impact upon the control chart limits. ... The control limits will be contaminated by an appreciable amount only when the autocorrelation becomes excessive (say 0.80 or larger)*".

However, as illustrated by Table 3.1 of Section 3.3, ignoring even moderate levels of autocorrelation may have a large impact on the ARL behavior of Shewhart-type control charts. This is mainly due to the bias that is introduced if σ_Y is naively estimated as in the case of independent observations. A careful analysis of the correlation structure in the data remains necessary in all cases to obtain appropriate control limits.

Furthermore, even if σ_Y is known, the approach is not satisfactory for values of $\phi > 0.8$. In addition, there is extra information on the data structure available that is not explicitly used in the modified Shewhart chart. Taking the serial correlation explicitly into account might open up possibilities for improving ARL behavior. In the next two sections, Shewhart-type control charts will be discussed that explicitly utilize serial correlation in the data.

3.5 The residuals control chart

In the previous section, we discussed a first approach for monitoring serial correlated data. The limits of the modified Shewhart chart are adjusted to account for the effect of serial correlation. A second approach for monitoring AR(1) observations will be discussed in this section. It is based on the idea that the process can be monitored using the residuals from fitting a time series model to the data. If a shift in the mean of the process occurs, this will result in a shift in the mean of the residuals. Furthermore, if the time series model fits the data well, the residuals will be approximately uncorrelated. This provides a theoretically elegant way to monitor a serially correlated process using control charts that were designed for independent observations.

This approach has received much more attention in the literature on SPC than the modified Shewhart chart. References can be found in this section and in Section 3.7.

3.5.1 Residuals of an AR(1) process

The residual control chart is based on charting residuals

$$e_t \equiv Y_t - \hat{Y}_{t|t-1,t-2,\cdots}, \tag{3.5}$$

where $\hat{Y}_{t|t-1,t-2,\cdots}$ is a predictor of Y_t, based on observations up to and including time $t-1$. The linear predictor that minimizes the mean square error is $\mathrm{E}(Y_t|Y_{t-1}, Y_{t-2}, \cdots)$, see for example Harvey (1993). In the case of AR(1) data generated by the in-control model (2.2), we have

$$\hat{Y}_{t|t-1,t-2,\cdots} = \mathrm{E}\left(Y_t|Y_{t-1}, Y_{t-2}, \cdots\right) = \mu + \phi(Y_{t-1} - \mu). \tag{3.6}$$

In practice, μ and ϕ have to be estimated from a data set that was obtained in a period where only common causes of variation were affecting the process. Throughout this chapter we will assume that enough in-control observations are available so that μ and ϕ can be estimated accurately. In Section 3.7, references are cited that discuss the effect of estimating process parameters on the ARL curve.

As long as the process is in control, observations are assumed to be generated by model (2.2). Moreover, the quantities e_t that will be plotted in the residuals control chart satisfy

$$e_t = Y_t - \hat{Y}_{t|t-1,t-2,\cdots} \approx \varepsilon_t \qquad \text{for all } t \tag{3.7}$$

where the last relation is exact if μ and ϕ are known. Suppose that a special cause shifts $\mathrm{E}(Y_t)$ at time T by an amount of $\delta\sigma_Y$. Since we are not aware of this shift, we compute $e_T, e_{T+1}, \cdots$ as if the process were in control. Hence, also for $t > T$, computation of e_t is given by (3.5), with $\hat{Y}_{t|t-1,t-2,\cdots}$ computed as in (3.6).

The elements of the sequence of residuals $\{e_t\}$ satisfy

$$e_t \approx \begin{cases} \varepsilon_t & \text{for } t < T \\ \varepsilon_t + \delta\sigma_Y & \text{for } t = T \\ \varepsilon_t + (1-\phi)\delta\sigma_Y & \text{for } t = T+1, T+2, \cdots. \end{cases} \tag{3.8}$$

Note that the $\{e_t\}$ are independently distributed if the parameters μ and ϕ are known, since we assumed that $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$. Hence, we are back at the i.i.d. case. For $\phi > 0$, only a fraction of the shift in $\mathrm{E}(Y_t)$ is transferred to the residuals for $t > T$. For $\phi < 0$, the shift is magnified. Therefore, we expect the residuals chart to perform better for AR(1) data with negative ϕ than for AR(1) data with positive ϕ.

3.5.2 The ARL of the residuals chart for AR(1) data

In the case where the first observation is taken after the shift in $\mathrm{E}(Y_t)$ has occurred, the computation of the ARL of the residuals chart is analogous to the computation of the ARL curve of a Shewhart chart for independent observations, as described in Section 3.2.

However, if the first observation is taken at the time of the shift, or if we are interested in the expected number of observations that a shift goes undetected, the computation is slightly different. The reason for this is that the probability of observing a residual between the control limits at the time of the shift differs from the probability that a residual falls between the control limits after the shift. Let us denote the probability that a residual falls between the control limits for $t > T$ by $\mathrm{P}(\delta)$, and let $\mathrm{P}_1(\delta)$ denote the probability that a residual falls between the limits at time T. In Longnecker and Ryan (1992) it was shown that $\mathrm{ARL}_{\mathrm{rc}}(\delta)$, the ARL of the residuals chart, satisfies

$$\mathrm{ARL}_{\mathrm{rc}}(\delta) = 1 \; + \; \frac{1}{1 - \mathrm{P}(\delta)} \mathrm{P}_1(\delta) \tag{3.9}$$

if the first observation is taken at the time of the shift. Note that, if $\mathrm{P}_1(\delta) = \mathrm{P}(\delta)$, the right-hand side equals $1/(1 - \mathrm{P}(\delta))$, which is the ARL of the residuals chart if the first observation is taken after the shift in $\mathrm{E}(Y_t)$ has occurred.

The difference between $\mathrm{ARL}_{\mathrm{rc}}(\delta)$ and $1/(1-\mathrm{P}(\delta))$ is negligible for negative ϕ. However, for large positive ϕ, the difference is quite large. This can be explained by considering *signal-to-noise* ratios. A signal-to-noise ratio is a number that relates the size of the shift to the standard deviation of the process. This ratio allows us to compare the effect of a shift in the mean of processes with different variances. We define the signal-to-noise ratio as the size of the shift divided by the standard deviation of the process.

At time T, the size of the shift that is transferred to the residuals is approximately equal to $\delta\sigma_Y$. Dividing this quantity by σ, we have

$$\frac{\delta\sigma_Y}{\sigma} = \frac{\delta}{\sqrt{1 - \phi^2}}.$$

Hence, the signal-to-noise ratio converges to infinity as $\phi \to 1$. As a result, the probability that the shift will be detected at the first observation converges to one as $\phi \to 1$. Consequently, $\mathrm{P}_1(\delta) \to 0$ for $\delta \neq 0$. This affects

$\mathrm{ARL}_{\mathrm{rc}}(\delta)$ positively, since it follows from Equation (3.9) that $\mathrm{ARL}_{\mathrm{rc}}(\delta)$ converges to 1 as $\mathrm{P}_1(\delta) \to 0$.

For $t > T$, the signal-to-noise ratio is approximately equal to

$$\frac{(1-\phi)\delta\sigma_Y}{\sigma} = \delta\,\frac{\sqrt{1-\phi}}{\sqrt{1+\phi}}, \tag{3.10}$$

which converges to 0 as $\phi \to 1$. Hence, for values of ϕ very close to one it is very hard to detect a shift if it is not detected at the first observation.

In Figures 3.14 through 3.19, ARL curves of the residuals chart are compared to ARL curves of the modified Shewhart chart for various values of ϕ. For these curves it is assumed that the first observation is taken at the time of the shift. As a reference, also the ARL curve for the i.i.d. case is depicted. The ARL curves for the modified Shewhart chart were depicted earlier in Figures 3.3 and 3.4.

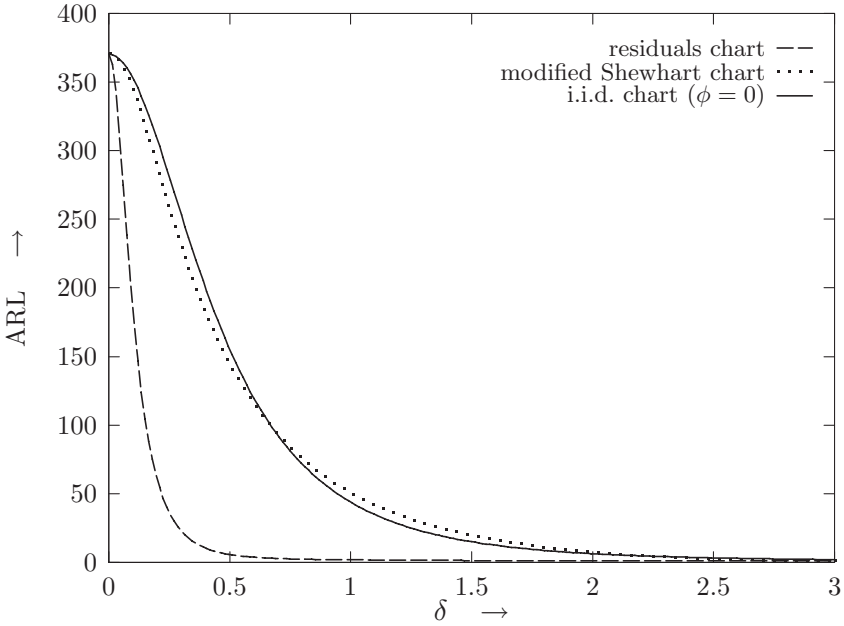


Figure 3.14: Various ARL curves for AR(1) process with $\phi = -0.9$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

3.5.3 Discussion

From Figures 3.14 through 3.19, we conclude that for negative first-order autocorrelation, the residuals chart is performing better than the Shewhart

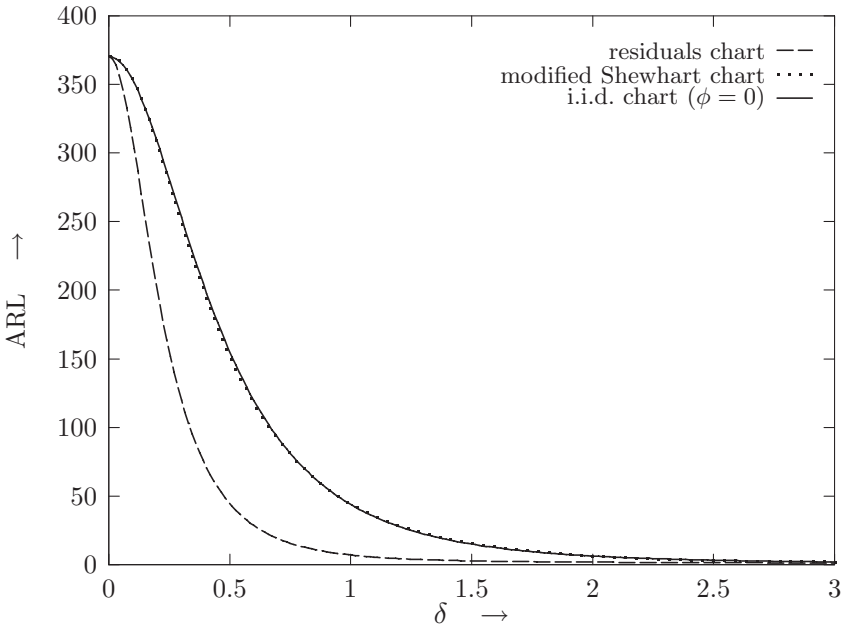


Figure 3.15: Various ARL curves for AR(1) process with $\phi = -0.6$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

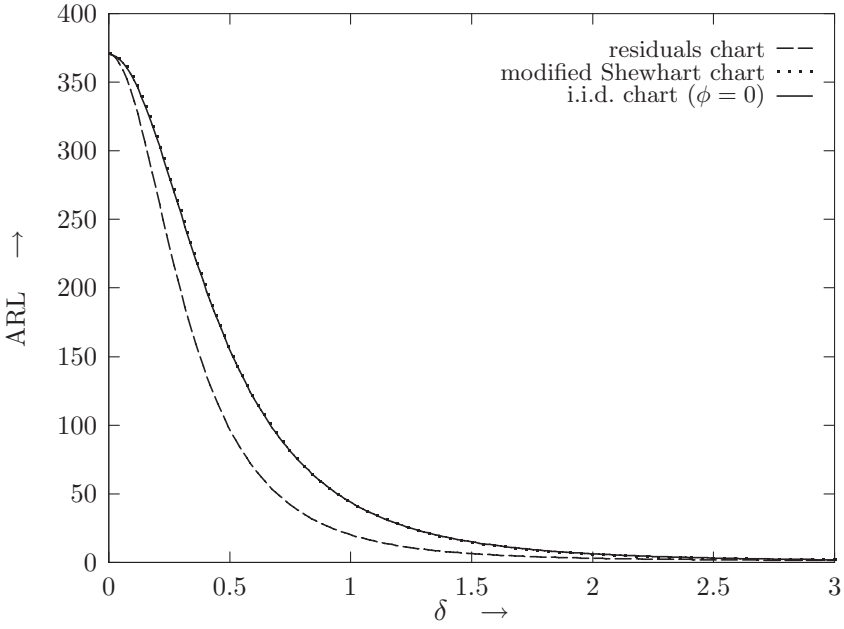


Figure 3.16: Various ARL curves for AR(1) process with $\phi = -0.3$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

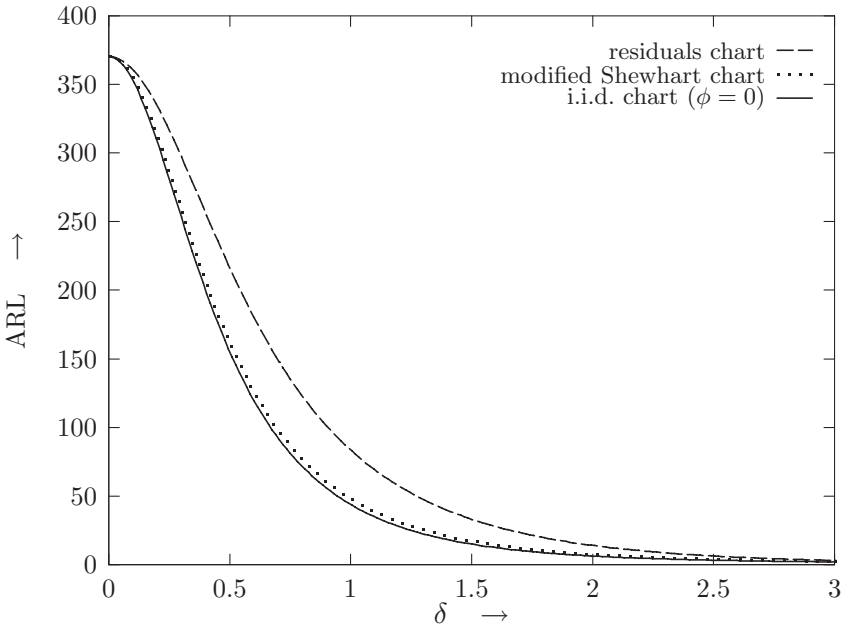


Figure 3.17: Various ARL curves for AR(1) process with $\phi = 0.3$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

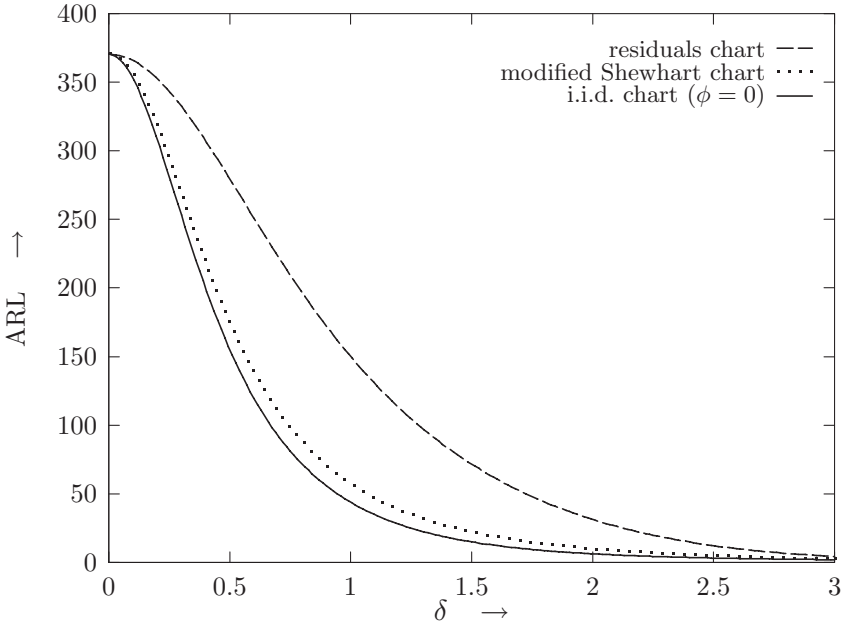


Figure 3.18: Various ARL curves for AR(1) process with $\phi = 0.6$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

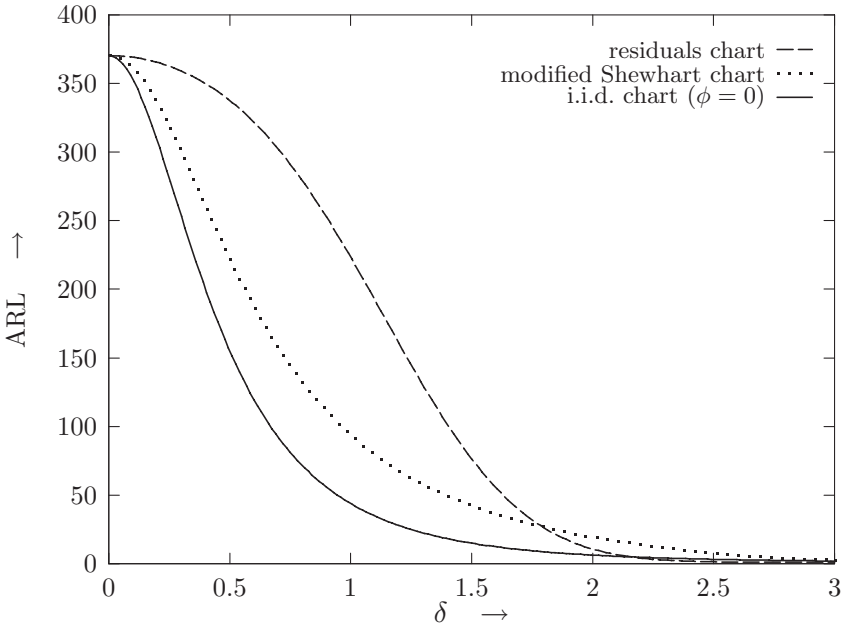


Figure 3.19: Various ARL curves for AR(1) process with $\phi = 0.9$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

chart for independent observations. This was to be expected, since in Subsection 3.5.1 it was shown that for negative autocorrelation, a shift in AR(1) observations is magnified in the residuals.

Compared to the Shewhart chart for independent observations, the performance of the residuals chart is worse for positive autocorrelation. This is caused by the fact that only a fraction of the shift is transferred to the residuals for positive ϕ.

For values of ϕ very close to one, we argued in the previous subsection that the corresponding ARL curve will converge to 1 for all $\delta > 0$. A value of $\phi = 0.9$ is not large enough to show this convincingly in Figure 3.19. Therefore, in Figure 3.20, the three ARL curves are drawn for $\phi = 0.99$. This graph shows that the ARL performance of the residuals chart is improved for values of ϕ very close to 1.

The conclusions drawn from Figures 3.14 through 3.19 and Figure 3.20 agree with a comment of Ryan (1991):

> "*A residuals chart for* AR(1) *data will perform poorly unless* ϕ *is negative or extremely close to* 1. *In most applications we would expect to have* $\hat{\phi} > 0$ *and not*

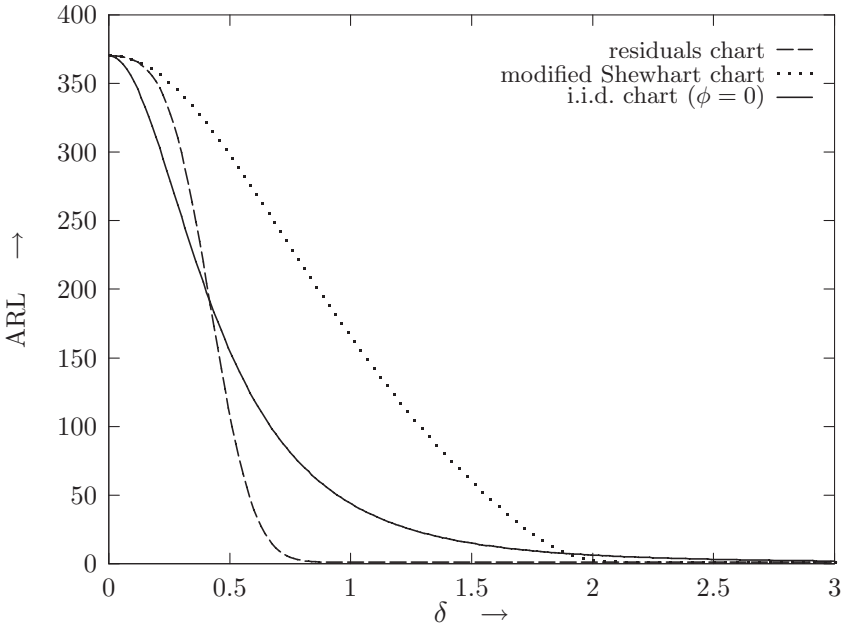


Figure 3.20: Various ARL curves for AR(1) process with $\phi = 0.99$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

particularly close to one".

This is an important disadvantage of the residuals chart.

At the start, the residuals chart seemed to be attractive. By removing the serial correlation, the problem is reduced to the well-known case of detecting a shift in the mean of independent observations. In contrast, the modified Shewhart chart basically ignores the serial correlation in the data. The control limits are adjusted in a rather ad-hoc manner to ensure a certain in-control ARL. However, for positive ϕ, the modified Shewhart chart is performing better than the residuals control chart.

The findings of the last two sections can be summarized in the advice to use the modified Shewhart chart for detecting a shift in the mean of AR(1) data with positive ϕ, and to use the residuals chart in case of AR(1) data with negative ϕ. For $\phi < 0$, a shift in the mean is then, on average, detected faster than in the i.i.d. case, while for $\phi > 0$ not much efficiency is lost compared to the i.i.d. case, provided that ϕ is not too large.

3.6 A modification of the residuals control chart

In the previous section it was concluded that the bad performance of the residuals chart for positive ϕ is caused by the fact that only a fraction of the shift in the mean is transferred to the residuals. As a result the signal-to-noise ratio is smaller than δ for $t > T$ and $\phi > 0$, see formula (3.10). This has a negative effect on the performance of the residuals control chart. In general, a higher signal-to-noise ratio will result in a more efficient control chart. For example, in Chapter 4 we will conclude that the efficiency of the well-known EWMA chart is mainly due to a good signal-to-noise ratio. The bad signal-to-noise ratio for positive ϕ is an important disadvantage of the residuals chart.

On the other hand, the residuals chart is theoretically very appealing because it takes the serial correlation explicitly into account, and reduces the problem to the well-known case of detecting a shift in the mean of independent observations.

3.6.1 The idea of the modified residuals chart

In this section, we try to combine the appeal of the residuals chart with a good signal-to-noise ratio. We suggest a modification of the residual chart that roughly maintains independence of the residuals, while the signal-to-noise ratio is larger than δ within a few observations after the shift. In this way, the main drawback of the residuals chart is overcome, and serial correlation is explicitly accounted for. The modified residuals can be monitored using control charts that were designed for detecting a change in the mean of independent observations.

Let us denote the sequence of modified residuals by $\{u_t\}$. Our suggestion is to set up a control chart for $\{u_t\}$, where u_t is defined for $t = 0, 1, 2, \cdots$ as

$$u_t \equiv Y_t - \phi Y_{t-1} + \phi \hat{\mu}_t, \tag{3.11}$$

where $\hat{\mu}_t$ is an estimator of μ_t that quickly responds to changes in the mean of the process. The rationale behind this suggestion is the following. Suppose that $\hat{\mu}_t$ is a very good estimator for μ_t, so that

$$\hat{\mu}_t \approx \begin{cases} \mu & \text{for } t = 0, 1, \cdots, T-1 \\ \mu + \delta\sigma_Y & \text{for } t = T, T+1, \cdots. \end{cases} \tag{3.12}$$

In that case

$$u_t \approx \begin{cases} \mu + \varepsilon_t & \text{for } t = 0, 1, \cdots, T-1 \\ \mu + \delta\sigma_Y + \varepsilon_t & \text{for } t = T, T+1, \cdots. \end{cases} \tag{3.13}$$

The right-hand side of (3.13) represents an ideal situation, where the successive u_t are independent, and the shift in μ_t is fully transferred to the modified residuals.

Of course, in practice, such a perfect estimator of $\{\mu_t\}$ is not available. In the research leading up to this thesis, simulation studies were performed, wherein we experimented with a regular moving average with a small window size of say, 5 to 10 observations. From these simulations we learned that a regular moving average is not the best option: hardly any efficiency is gained for positive ϕ compared to the modified Shewhart chart.

Thereafter, we simulated the ARL curve using a sequence of *Exponentially Weighted Moving Average* (EWMA) statistics to estimate $\{\mu_t\}$. This turned out to be a preferable alternative.

The EWMA statistic at time t will be denoted by W_t and is constructed as follows

$$W_t = (1-\lambda)W_{t-1} + \lambda Y_t \qquad \text{for } t = 1, 2, \cdots, \tag{3.14}$$

where $\lambda \in (0, 1)$. The sequence of EWMAs may be started by setting W_0 equal to a target value or (an estimator of) μ. If $W_0 = \mu$, it is easy to verify that $\mathrm{E}(W_t) = \mu$ for $t = 0, 1, 2, \cdots, T$. For $t \geq T$ we have that

$$\mathrm{E}(W_t) = \mu + \left[1 - (1-\lambda)^{t-T+1}\right] \delta\sigma_Y,$$

which approximately equals $\mu + \delta\sigma_Y$ for $t \gg T$. Hence, $\{\mathrm{E}(W_t)\}$, the sequence of expected values of the EWMA statistic, approximately mimics $\{\mathrm{E}(Y_t)\}$. In Subsection 3.6.3, some considerations are presented on the choice of the EWMA smoothing parameter λ. Chapter 4 is devoted to control charts based on the EWMA statistic.

3.6.2 Comparison to other procedures

To judge the effect of modifying the residuals, we compare the ARL curve of the modified residuals chart to ARL curves of the control charts we discussed earlier. In each of the Figures 3.21 through 3.26, ARL curves

corresponding to the four different control charts are drawn for a fixed value of ϕ ranging from $\phi = -0.9$ to $\phi = 0.9$. For the case $\phi = 0$ the four curves coincide, so that this graph is omitted.

For all of the curves it is assumed that the first observation is taken at the time of the shift. The ARL curve for the modified Shewhart chart is $L^*_\phi(\delta, s)$, and the ARL curve for the residuals chart is computed using formula (3.9). The ARL curve for the modified residuals chart is derived by simulation. The curve consists of 101 points, which are means of 10,000 replications. The random number generator we used is described in Knypstra (1997). A value of $\lambda = 0.1$ was chosen for computation of the EWMA.

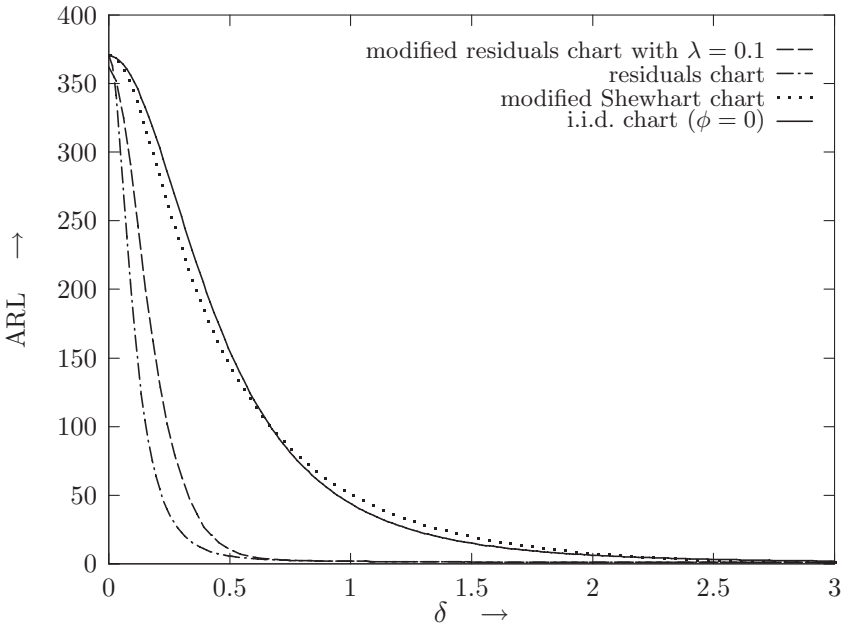


Figure 3.21: Various ARL curves for AR(1) process with $\phi = -0.9$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

From Figures 3.21 through 3.26 we conclude that for negative ϕ, the ARL performance of the modified residuals is better than the ARL performance of the modified Shewhart chart and the ARL performance of the the Shewhart chart for independent observations. However, the excellent ARL performance of the residuals chart for negative ϕ is not equalled by the modified residuals chart. Hence, for $\phi < 0$, the residuals chart remains the best choice, and the modified residuals chart is a well-performing second-

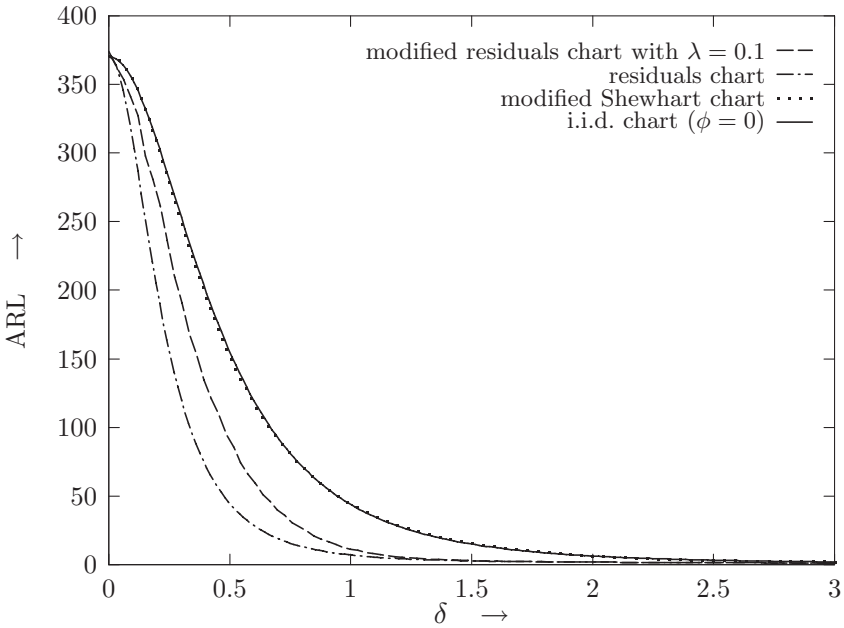


Figure 3.22: Various ARL curves for AR(1) process with $\phi = -0.6$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

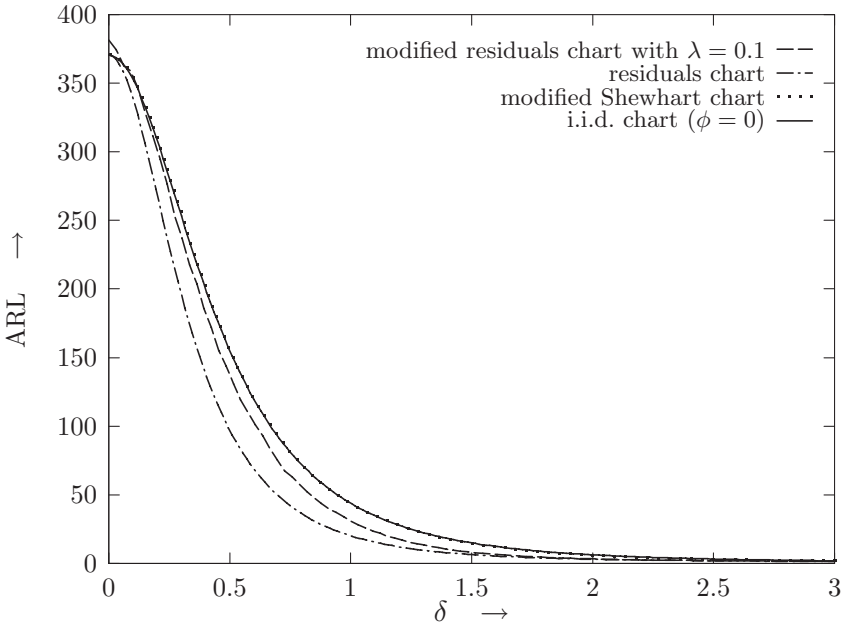


Figure 3.23: Various ARL curves for AR(1) process with $\phi = -0.3$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

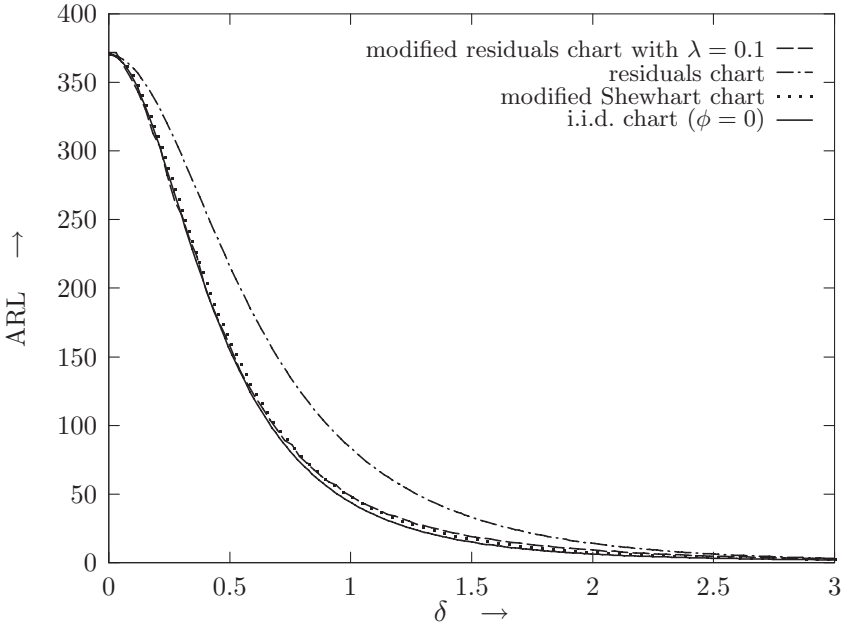


Figure 3.24: Various ARL curves for AR(1) process with $\phi = 0.3$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

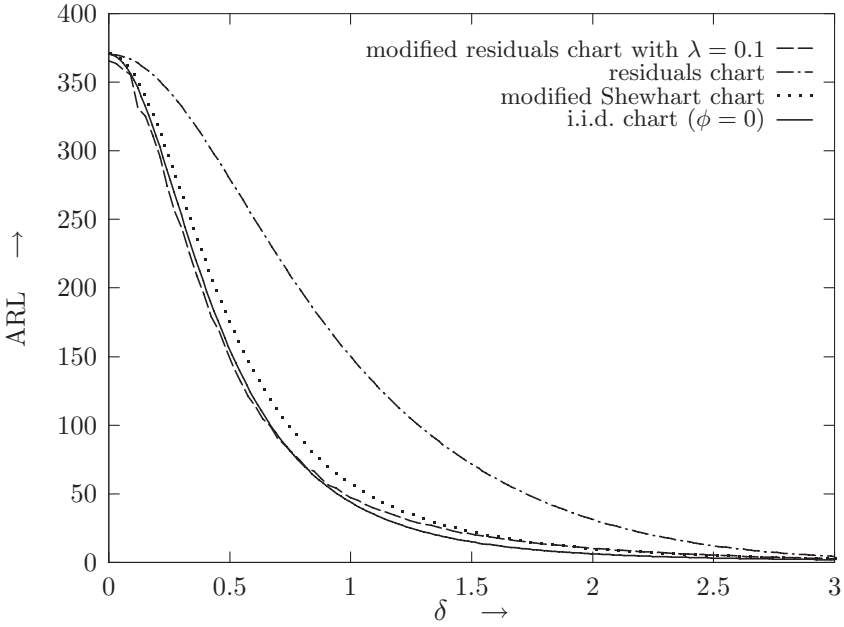


Figure 3.25: Various ARL curves for AR(1) process with $\phi = 0.6$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

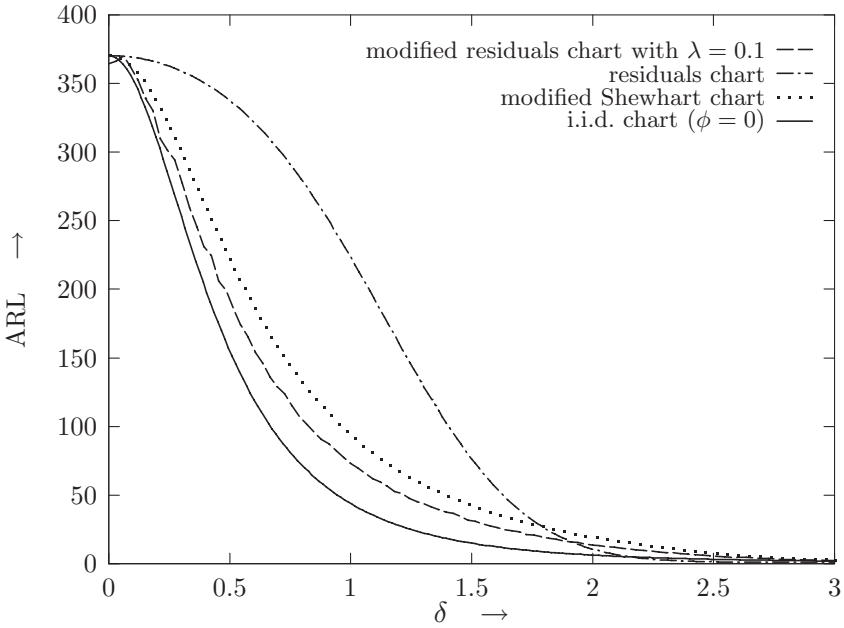


Figure 3.26: Various ARL curves for AR(1) process with $\phi = 0.9$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

best option.

For positive ϕ, the modified residuals chart outperforms both the modified Shewhart chart and the residuals chart. The loss of efficiency due to serial correlation compared to the i.i.d. case is negligible for $\phi \leq 0.6$. However, for larger positive values of ϕ, this procedure becomes less efficient, too.

3.6.3 Choice of the EWMA smoothing parameter

For a good performance of the suggested modification, it is necessary to have a good estimator for $\mathrm{E}(Y_t)$ available, one that quickly adapts to persisting changes in the mean which may occur due to the presence of special causes of variation. As mentioned before, in various simulation studies, we experimented with a regular moving average and an EWMA of the observations. It turned out that use of an EWMA leads to better ARL performance of the modified residuals chart compared to a modified residuals chart which uses a regular moving average. Use of the EWMA requires choosing a value for the smoothing parameter λ. In Table 3.2, it is tabulated how the ARL of the modified residuals chart for an AR(1) process with $\phi = 0.9$ is affected

by the choice of λ. This value of ϕ was chosen since for the values of ϕ considered, this ARL curve could use some improvement. The entries in Table 3.2 are obtained by simulation. The values in the column labeled 'ARL(0)' are the simulated ARL values of in the in-control situation, i.e. there has not been a shift in the mean. The simulated ARL values in the following columns correspond to shifts in the mean of size $1\sigma_Y$, $2\sigma_Y$ and $3\sigma_Y$, respectively. The bracketed numbers below the ARL values indicate the corresponding simulated standard deviations.

From Table 3.2 we conclude that to a limited extent, the choice of λ can be used as a design parameter for the modified residuals chart. If large shifts of size $2\sigma_Y$ or $3\sigma_Y$ are to be detected quickly, a small value of $\lambda = 0.01$ is recommended. A value of $\lambda = 0.05$ is a good choice for detecting a shift of size σ_Y. This value of λ is used in the simulation studies of Appendix A.

Table 3.2: The effect of λ on the ARL of the modified residuals chart for $\phi = 0.9$.

	ARL(0)	ARL(1)	ARL(2)	ARL(3)
$\lambda = 0.01$	377.4 (3.94)	77.3 (0.67)	9.7 (0.22)	1.1 (0.02)
$\lambda = 0.025$	372.1 (3.75)	69.3 (0.62)	10.6 (0.19)	1.8 (0.03)
$\lambda = 0.05$	366.9 (3.76)	67.4 (0.64)	12.3 (0.18)	1.8 (0.04)
$\lambda = 0.075$	368.6 (3.75)	70.6 (0.71)	13.4 (0.18)	2.1 (0.05)
$\lambda = 0.1$	364.4 (3.68)	73.1 (0.74)	13.7 (0.18)	2.3 (0.05)
$\lambda = 0.125$	365.0 (3.63)	75.0 (0.77)	14.2 (0.18)	2.4 (0.05)
$\lambda = 0.15$	364.7 (3.65)	78.0 (0.80)	14.3 (0.19)	2.5 (0.05)

3.6.4 Discussion

Summarizing, it may be said that the modified residuals chart has an overall good performance: if ϕ is negative, it is more efficient than the Shewhart chart for independent observations, and it also performs better than the

modified Shewhart chart. The efficiency gain of the residuals chart for negative ϕ is only partly attained by the modified residuals chart. For positive ϕ, the modified residuals chart outperforms both the residuals chart and the modified Shewhart chart. For small to moderate ϕ there is virtually no loss of efficiency compared to the benchmark-curve of the Shewhart chart for independent observations.

The crux of the modification is the addition or subtraction of a portion of $\hat{\mu}_t$. The estimator that is used needs to adapt quickly to persisting changes in the level of the observations, but must be insensitive to the effect of short term random disturbances. As it turns out, an EWMA is a better choice than a regular moving average. The smoothing parameter λ should be chosen somewhere within the range $[0.01, 0.15]$. Within this range, it is to a limited extent possible to choose λ in such a way that the modified residuals chart is most sensitive to shifts of a given size. Table 3.2 can provide some guidance when choosing a value of λ.

In practice it is, of course, not possible to estimate the sequence $\{\mu_t\}$ perfectly. If this were possible, formula (3.13) would become an exact relationship, and the corresponding ARL curve would equal the curve of the Shewhart chart for independent observations. Hence, the latter may be viewed as a kind of limiting ARL curve for the modified residuals chart with a very good smoothing procedure. In the simulation studies leading up to this chapter, we only considered regular moving averages with a small window size and the EWMA. It is possible that another smoothing procedure performs even better for large positive values of ϕ, in the sense that the corresponding ARL curve is closer to the ARL of the Shewhart chart for independent observations. This remains to be investigated.

For practical purposes, the modified residuals chart is an improvement on existing procedures since the chart outperforms both other Shewhart-type control charts for the case of $\phi > 0$. This case is more likely to occur in practice than the case of negative ϕ. But also from a theoretical point of view this approach is appealing. The extra information on the data structure that is provided by the presence of serial correlation is used explicitly, and the problem is approximately transformed into the more familiar case of monitoring a sequence of independent observations. However, due to the imperfect estimate of $\{\mu_t\}$, some serial correlation remains, and some ad-hoc adjustments to the control limits are needed to attain a desired in-control ARL.

3.7 Related work of other authors

In a recent article by Faltin, Mastrangelo, Runger and Ryan (1997), a nice overview is presented of the developments concerning control charts for autocorrelated data. They categorize the articles that appeared on this subject into two groups. In the first group of articles, the modified Shewhart chart is discussed. In the second group of articles, the residuals chart is considered.

Vasilopoulos and Stamboulis (1978) introduced the modified Shewhart chart, and used it for monitoring subgrouped AR(2) data. The authors calculated the variance of the mean in the in-control state and gave curves for the modified quality control factors so that the width of the control limits is appropriate for the variance of the sample mean and sample standard deviations. It was shown that the classical quality control factors are substantially affected by dependence in the data.

The first and main contribution of the paper by Vasilopoulos and Stamboulis is the introduction of the modified Shewhart chart. The second contribution is that attention is drawn to the fact that subgrouping of serially correlated data has a substantial effect on the variance of the sample mean and on the variance of the sample standard deviation. However, modifying the limits so that the width of the control chart is appropriate for the variance of the statistic that is plotted in the chart is only half of the story in the case of serially correlated data.

Dependence in the data does not only affect the ARL curve through the variance of the observations. Even if a right normalizing variance is used for determining the width of the control limits, the ARL curve is affected by the 'memory' in the data. This was encountered in Figure 3.2, where the right normalizing variance was used for computing the control limits for AR(1) data. We observed that the dependence structure itself was responsible for a higher in-control ARL, compared to the in-control ARL for independent observations. This observation is typical for the much more general class of Gaussian processes. In an excellent article by Schmid (1995b), several theorems were proved about the behavior of the run length of the modified Shewhart chart for Gaussian processes. Based on these results, Kramer and Schmid (1996a) proved that the in-control ARL for a modified Shewhart chart for Gaussian process is larger than or equal to the in-control ARL of the Shewhart chart for independent observations.

In most references on control charts for serially correlated data, only the residuals chart is discussed. The first reference we encountered on Shewhart-type control charts of residuals is Berthouex, Hunter, and Palle-

sen (1978). In this article, the residuals of an autoregressive model containing a daily and a weekly component were used to monitor the level of effluent BOD_5 of two sewage treatment plants.

In Ermer, Chow, and Wu (1979) and in Ermer (1980), the residuals of an ARMA(2,1) model were used to monitor the level of neutron flux data of a nuclear reactor. In Notohardjono and Ermer (1986) fourth-order ARMA models were used to model two data sets of a blast furnace shell. The corresponding residuals were used to monitor the processes for upsets.

In Alwan and Bissell (1988), the concept of the residuals chart is explained for use in clinical laboratories. A data set is used to illustrate its use in clinical chemistry. In Alwan and Roberts (1988), an IMA(1,1) data set from Box and Jenkins (1976) is used to illustrate the "*common cause chart*", a chart of fitted values, and the "*special cause chart*", the regular residuals chart. The common cause chart is used to monitor the local level of the process. Values exceeding action limits on this chart are not signals notifying that it is time to look for special causes; rather, they indicate that a specific corrective action is needed. Furthermore, Alwan and Roberts argue that the time series model underlying the EWMA is a flexible time series model that, for many but not all processes, may serve as a satisfactory approximation to ARIMA modelling. These EWMA residuals of any kind of serially correlated data are monitored in the special cause chart. In Tseng and Adams (1994), the ARL behavior of this procedure is evaluated. Their advise is not to use EWMA residuals, but to use residuals of an appropriate time series model instead.

Montgomery and Mastrangelo (1991) also use the EWMA as an approximation to ARIMA modelling. However, they propose to combine information about the state of statistical control and process dynamics on one control chart: the so-called 'EWMA Center Line Control Chart'. This is based on the idea that the conditional mean of a stationary time series model is constantly changing. The sequence of fitted EWMA values estimate these changes. The centerline of the control chart is therefore not taken to be constant, but equal to the sequence of fitted values. The band of control limits moves up and down with the centerline.

Whether to use fixed or variable control limits to monitor a serially correlated process has been the subject of some controversy in the literature on SPC. Alwan (1992) argues that

> "*... control limits with any width which are horizontally positioned at a fixed amount above and below the estimated process average are intrinsi-*

> *cally inappropriate in the sense that they misdirect the detection of special causes reflected as extreme points*"

Also in Alwan and Radson (1992b) the 'inappropriateness' of fixed limits for monitoring the mean of correlated data is stressed. It is the opinion of Montgomery and Mastrangelo (1991) that an appropriate monitoring control chart should have variable limits to allow for movements in the conditional mean.

However, using variable control limits is, in essence, applying a residuals chart with control limits moving up and down with a center line which is not equal to zero, but an estimate of the local level of the process. The advantage of a moving centerline chart over a residuals chart is that the original observations are plotted in the control chart. However, its ARL behavior is equal to that of a residuals chart.

In several articles in which the ARL of residuals charts is evaluated, notably Schmid (1995b) and Kramer and Schmid (1996a), it is concluded that the ARL performance of modified control charts is much better than that of residuals charts for positive autocorrelation. We encountered the earliest reference explaining this phenomenon in the discussion of Montgomery and Mastrangelo (1991), where Thomas P. Ryan remarks that a shift in the mean of AR(1) observations is only partly transformed to the residuals for positive ϕ (see also the comments following Equation (3.8)).

Although we are a little uncomfortable with the fact that the extra information on the data structure is not used, we prefer modified Shewhart charts over the residuals or moving centerline charts. The bad ARL behavior of residuals and moving centerline control charts for positive autocorrelation is a strong argument against their use. However, we recommend using the modified residuals chart discussed in Section 3.6, which explicitly accounts for serial correlation and has a good ARL behavior for negative autocorrelation, and the best ARL behavior for positive autocorrelation.

In Wardell, Moskowitz, and Plante (1992) an ARL-based comparison is carried out for ARMA(1,1) data. In this simulation study, the combination of common cause and special cause control charts is compared to the modified Shewhart, and the EWMA control chart. In Maragah and Woodall (1992), realizations of AR(1) and MA(1) models are simulated. It is investigated how the Shewhart chart for the mean with control limits based on moving ranges is affected by these types of serial correlation. It turns out that for positive first-order autocorrelation, the control limits are too tight, so that many false alarms are generated. For negative first-order

autocorrelation, the control limits are unnecessarily wide, so that significant shifts in the process mean may go undetected. In Wardell, Moskowitz, and Plante (1994a), run length distributions of the special cause control chart proposed by Alwan and Roberts are derived, and the ARL and the standard deviation of the run length is derived for any AR(p) process. Numerical results are given for the ARMA(1,1) model, given that there was a step change in the mean. In Wardell, Moskowitz, and Plante (1994b), an accompanying computer program listing is presented.

In Tseng and Adams (1994), the performance of Shewhart, EWMA and CUSUM charts is investigated, when applied to EWMA forecast errors. It is shown that the EWMA forecasts do not adequately account for the dependence structure that is present in AR(1) models. Longnecker and Ryan (1992) discuss the properties of the residuals chart in case of AR(1), AR(2) and ARMA(1,1) data. They conclude that the residuals chart will perform poorly for most parameter values likely to occur with process control data. In Lu and Reynolds (1996), several types of control charts and combinations of control charts are evaluated in terms of their ability to detect changes in the mean and variance of an AR(1) process. They recommend using an EWMA control chart for the observations, in combination with a Shewhart chart of the residuals for practical applications.

Lin and Adams (1996) advocate the use of a combination of a Shewhart chart of residuals with an EWMA chart of residuals which will be discussed in the next chapter. In this article, it is shown that the Shewhart chart of residuals has a higher probability of detecting a shift in the mean at the first observation following the shift, compared to the EWMA chart of residuals. The EWMA chart of residuals is shown to have a good ARL performance. The combination of these two charts takes advantage of the possible fast detection of the Shewhart chart of residuals and the desirable ARL behavior of the EWMA chart of residuals.

3.8 Conclusions

In this chapter we discussed three Shewhart-type control charts that are able to take first-order autocorrelation in the observations into account. The performance of these control charts was compared by means of ARL curves.

The first control chart that was evaluated is a modification of the classical Shewhart control chart. It basically ignores the autocorrelation; the control limits are adjusted in a rather ad-hoc manner in order to attain a

certain in-control ARL. For negative and moderate positive first-order autocorrelation, the ARL performance of this control chart is comparable to that of the classical Shewhart control chart for independent observations. This is a comforting observation: out-of-control signals of the modified Shewhart chart can be interpreted just as in the i.i.d. case in a large number of practical situations.

Secondly, the residuals chart was discussed. This chart utilizes the residuals of a fitted time series model to monitor the process for shifts in the mean. If the time series model is appropriate for the data, the residuals are approximately uncorrelated. In this way, the serial correlation is explicitly taken into account, and the problem is transformed into the well-known case of detecting a shift in the mean of independent observations. For negative first-order autocorrelation and for values of ϕ that are extremely close to one, the ARL performance of the residuals chart is excellent. Compared to the case of independent observations, a shift in the mean is, on average, detected faster. However, for values of $\phi > 0$ and not close to one, the residuals chart performs poorly. This is an important drawback of the residuals chart.

To overcome this drawback, a third control chart is introduced in this chapter. It is a modification of the residuals chart. This chart was shown to have a good overall ARL performance. For negative values of ϕ, it is not as efficient as the regular residuals chart, but it is more efficient than the modified Shewhart chart and the classical Shewhart chart for the i.i.d. case. For positive ϕ, it outperforms both the modified Shewhart chart and the residuals chart. The difference with the classical Shewhart chart for independent observations is negligible for a large region of ϕ-values. For large ϕ (say $\phi > 0.8$) however, improvement of ARL performance remains desirable. Further research into this is needed.

Finally, we would like to make two general remarks. First, in this chapter, only Shewhart-type control charts were discussed. For each of the procedures considered, improvement of the ARL performance is possible by utilizing control schemes that are based on the EWMA statistic, or on the *CUMulative SUM* (CUSUM) of the observations. Such control charts are the subject of Chapter 4 and Chapter 5, respectively. Secondly, throughout this chapter we assumed that the process parameters were known. In practice, these have to be estimated. In Kramer and Schmid (1996b) it is shown that both the modified Shewhart and the residuals chart react sensitively to parameter estimation. The robustness of the modified residuals chart for parameter estimation needs to be investigated.

Chapter 4

EWMA-type control charts for the mean

In the previous chapter, we discussed Shewhart-type control charts. These control charts only use the current observation or sample to monitor the process. In the next two chapters we consider control charts that also utilize previous observations. In Chapter 5, the *CUmulative SUM* (CUSUM) chart is discussed. In its basic form, an unweighted cumulative sum of the (standardized) observations is plotted against time in the CUSUM chart. This chart has a long 'memory'.

In the present chapter, we consider the *Exponentially Weighted Moving Average* (EWMA) control chart. Like the CUSUM, the EWMA utilizes all previous observations, but the weight attached to data is exponentially declining as the observations get older and older. By varying the parameter of the EWMA statistic the 'memory' of the EWMA control chart can be influenced. A control chart based on the EWMA was introduced by Roberts (1959). More recent references include Hunter (1986), Crowder (1987), and Lucas and Saccucci (1990).

In the previous chapter, the EWMA statistic was used as a local estimator for the level of the data. The EWMA 'smoothes out' the effect of single disturbances, and shows the behavior of the level of the data. This suggests using the EWMA as a statistic to monitor the mean of a process. Originally, the EWMA was developed by time series analysts to distinguish short term variation from long term variation such as trends and cyclic behavior.

Another application of the EWMA was mentioned in Section 3.7 of the previous chapter. The EWMA of previous observations provides an

approximate one-step-ahead predictor. It is easily shown that the EWMA is an optimal (in the sense of minimal *Mean Squared Error* (MSE)) one-step-ahead predictor if the underlying time series model is IMA(1,1). This property will be illustrated in Chapter 8.

The setup of this chapter is similar to that of the previous chapter. In Section 4.1, we discuss the EWMA control chart for independent observations. The ARL curve that is derived in this section will serve as a benchmark in Sections 4.3, 4.4 and 4.5, where the modified EWMA chart, the EWMA chart of residuals, and the EWMA of modified residuals will be discussed, respectively. In Section 4.6, the ARL behavior of the three types of EWMA control charts are compared, and conclusions are presented.

4.1 The EWMA control chart for i.i.d. observations

As in the previous chapter, the sequence of independent observations of a quality characteristic of interest is denoted by $\{X_t\}$. Assume that X_t, an observation at time t with $t \in \mathbb{Z}$, are independently distributed as

$$X_t \sim \mathcal{N}(\mu_t, \sigma_X^2) \qquad \text{for } t \in \mathbb{Z},$$

where the index t of μ_t indicates that the mean of the observations may shift over time. The value of the EWMA statistic at time t, which we will denote by $W_{X,t}$, is computed as follows

$$W_{X,t} = \lambda X_t + (1-\lambda) W_{X,t-1}, \tag{4.1}$$

where the parameter λ is a constant satisfying $\lambda \in (0,1)$. Usually $W_{X,0}$ is set equal to a target value, or (an estimation of) the mean μ. All information on previous observations that is needed for computing $W_{X,t}$ is stored in $W_{X,t-1}$.

If $W_{X,t}$ is interpreted as a one-step-ahead predictor of X_{t+1}, formula (4.1) can be rewritten as

$$\hat{X}_{t+1} = \hat{X}_t + \lambda(X_t - \hat{X}_t).$$

This formula shows that the predictor for X_{t+1} equals the predictor for X_t, corrected with a fraction of the error made in the forecast of X_t. As

Hunter (1986) remarks, this makes plotting of the EWMA almost as easy as plotting the successive observations.

Equation (4.1) can also be rewritten as

$$W_{X,t} = \begin{cases} W_{X,0} & \text{for } t = 0 \\ \lambda \sum_{i=0}^{t-1} (1-\lambda)^i X_{t-i} + (1-\lambda)^t W_{X,0} & \text{for } t = 1, 2, \cdots. \end{cases} \tag{4.2}$$

From expression (4.2) it is clear why the EWMA is sometimes called a *geometric moving average*, since the weights of past observations are declining as in a geometric series. The use of the term *average* can be justified by observing that for any t the weights sum to one, since

$$\lambda \sum_{i=0}^{t-1} (1-\lambda)^i + (1-\lambda)^t = 1.$$

The choice of λ determines the decline of the weights, and thereby the effect of past observations in the computation of $W_{X,t}$. However, for all possible choices of λ, more recent observations always receive more weight in the computation of $W_{X,t}$ than older observations. If $\lambda \to 1$ then $W_{X,t} \to X_t$, and the EWMA places all of its weight on the most recent observation. The EWMA control chart will then behave as the Shewhart control chart. If $\lambda \to 0$, then the most recent observation receives a small weight, whereas the weight attached to previous observations only slightly declines with the age of the observations. The EWMA then takes on the appearance of the CUSUM. How to determine λ is discussed later in this section.

If we take $W_{X,0} = \mu$ it is easily seen from Equation (4.2) that the expectation of $W_{X,t}$ is equal to μ whereas for the variance of $W_{X,t}$ we have

$$\sigma^2_{W_{X,t}} = \sigma^2_X \left(\frac{\lambda}{2-\lambda} \right) \left[1 - (1-\lambda)^{2t} \right]. \tag{4.3}$$

Using this expression, the control limits of an EWMA chart for the mean of $\{X_t\}$ can be computed. The lower control limit (UCL) at time t is constructed as follows

$$\text{LCL}_t = \mu - c\sigma_X \sqrt{\left(\frac{\lambda}{2-\lambda} \right) [1 - (1-\lambda)^{2t}]}, \tag{4.4}$$

while the upper control limit (UCL) of the EWMA is computed as

$$\mathrm{UCL}_t = \mu + c\sigma_X \sqrt{\left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2t}]}. \tag{4.5}$$

The EWMA chart generates an out-of-control signal at time t if the realization of $W_{X,t}$ at time t (which we will denote by w_t) is larger than UCL_t or smaller than LCL_t. In Equations (4.4) and (4.5) $c > 0$ is a constant that needs to be chosen by the designer of the control chart.

From Equation (4.3) we can see that $\sigma^2_{W_{X,t}}$ is increasing over time. Hence, the control limits in Equations (4.4) and (4.5) will become wider. However, unless λ is very small, $\sigma^2_{W_{X,t}}$ converges very quickly to $\lambda/(2-\lambda)\sigma^2_X$. When constant limits are preferred, the computations will be based on the asymptotic standard deviation instead of the exact $\sigma_{W_{X,t}}$. Control limits are then obtained by

$$\mathrm{LCL} = \mu - c\sigma_X \sqrt{\left(\frac{\lambda}{2-\lambda}\right)} \tag{4.6}$$

and

$$\mathrm{UCL} = \mu + c\sigma_X \sqrt{\left(\frac{\lambda}{2-\lambda}\right)}. \tag{4.7}$$

Usually, μ and σ_X will not be known in practice. Therefore, μ and σ_X will be replaced by estimators $\hat{\mu}$ and $\hat{\sigma}_X$ in Equations (4.4) through (4.7).

The EWMA control chart has two parameters, c and λ. Their value will be set based on requirements on the ARL curve. The first to study the ARL of the EWMA was Roberts (1959). Using simulation results, he derived nomograms for the ARL of normally distributed variables. Numerical results were obtained by Robinson and Ho (1978), they evaluated the ARL of the EWMA control chart using Edgeworth expansion. Crowder (1987) showed that the ARL of the EWMA chart can be written as a so-called Fredholm integral of the second kind. The ARL curve can then be evaluated by approximating the integral numerically, see Appendix B. This Fredholm-integral approach of Crowder will be discussed in this chapter. Lucas and Saccucci (1990) used a Markov-chain approach to evaluate the ARL of the EWMA control chart. This approach is discussed in Appendix C.

Assume that an out-of-control signal is given if $|W_{X,t} - \mu| > h$ for some constant h, and define $L_{W_X}(\delta, u)$ as the ARL of the EWMA chart for the mean of $\{X_t\}$, given that the shift in the mean is equal to $\delta\sigma_X$, and that the EWMA starts in $W_{X,0} = u$. The run length is 1 if x_1, a realization of X_1, is such that $|(1-\lambda)u + \lambda x_1 - \mu| > h$. Otherwise the run continues from $(1-\lambda)u + \lambda x_1 - \mu$. From this point on, we expect an additional run length of $L_{W_X}(\delta, (1-\lambda)u + \lambda x_1 - \mu)$. This leads to the following integral equation for $L_{W_X}(\delta, \cdot)$

$$\begin{aligned} L_{W_X}(\delta, u) &= 1 \cdot \Pr(|(1-\lambda)u + \lambda X_1 - \mu| > h) \; + \\ &\quad + \int_{\{|(1-\lambda)u+\lambda x_1-\mu| \le h\}} \left[1 + L_{W_X}(\delta, (1-\lambda)u + \lambda x_1 - \mu)\right] f(x_1) \mathrm{d}x_1 \\ &= 1 + \frac{1}{\lambda} \int_{-h}^{h} L_{W_X}(\delta, x) \, f\!\left(\frac{x + \mu - (1-\lambda)u}{\lambda}\right) dx, \qquad (4.8) \end{aligned}$$

where $f(\cdot)$ denotes the probability density function of X_t $(t = 1, 2, \cdots)$. Equation (4.8) is a Fredholm integral of the second kind. Some considerations on Fredholm-integral equations of the second kind are discussed in Appendix B.

The ARL of the EWMA chart depends on the smoothing parameter λ, and on the width of the control limits, which is determined by c in formulas (4.6) and (4.7). These parameters need to be chosen with care. In the following, we will present some considerations on how to choose λ and c. We will make use of the asymptotic LCL and UCL which appeared in formulas (4.6) and (4.7). We determine the value of the two parameters by fixing two points on the ARL curve. The first point that will be fixed is the desired in-control ARL. Fixing the in-control ARL is related to *producer's risk*. If the process is in control, an out-of-control signal is unwanted. The producer will suffer a certain loss if it accidentally happens that the process is stopped as the result of a false out-of-control signal. This demand on the ARL curve is usually formulated in terms of a high minimal in-control ARL. The second point that is fixed on the ARL curve, relates to *consumer's risk*. It is desirable for the customer to have a low ARL if the products are of unacceptable quality. Hence, the second point that will be fixed is usually formulated in terms of a maximal ARL when there is a certain large shift in the mean. By fixing this second point, the chart will be designed for detecting a shift in the mean of a certain size as fast as possible.

In Figure 4.1, iso-$L_{W_X}(0, 0)$ curves are drawn for various values of the in-control ARL. All combinations of λ and c on one curve yield the same

producer's risk.

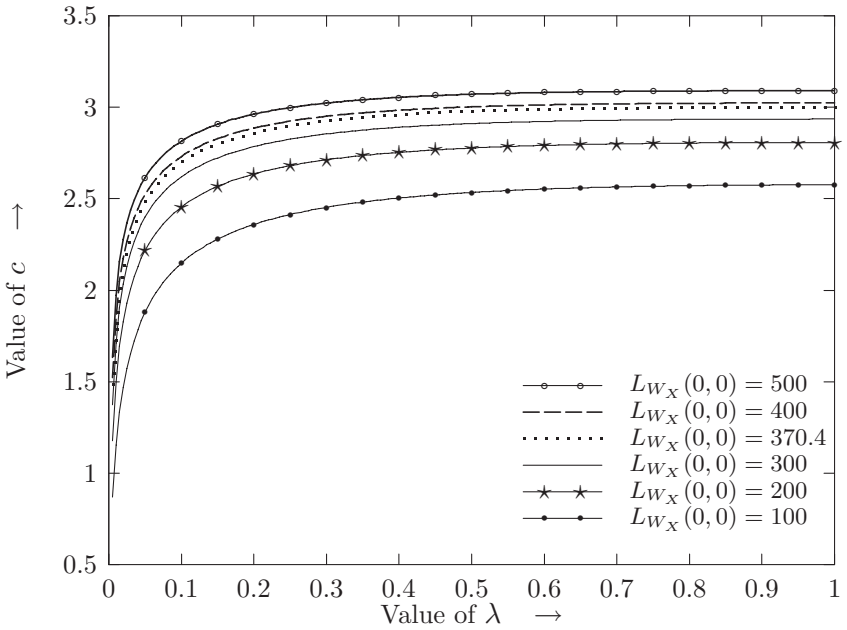


Figure 4.1: Iso-$L_{W_X}(0,0)$ curves for various in-control ARLs

From Figure 4.1, it is not clear which point on an iso-$L_{W_X}(0,0)$ curve to choose. Therefore, for shifts in the mean ranging from 0 to 3, the point of the iso-$L_{W_X}(0,0)$ curve is determined that yields the minimal out-of-control ARL, thereby minimizing the consumer's risk. This was done for each of the six choices of iso in-control ARL curves. This results in six curves in the c, λ, δ space. The projection of these curves on the λ, δ plane is shown in Figure 4.2. The projections in the c, δ plane is depicted in Figure 4.3.

With the aid of Figures 4.2 and 4.3, the EWMA chart can be designed to have minimal out-of-control ARL for a predetermined shift in the mean, given a certain in-control ARL.

In Figure 4.4, the ARL curve of an EWMA chart that is most sensitive to a shift of size $1\sigma_X$ in the mean of $\mathrm{E}(X_t)$, given an in-control ARL of 370.4, is depicted. The parameters of $c = 2.7878$ and $\lambda = 0.1417$ were determined using Figures 4.2 and 4.3 as being the optimal choice of parameters for detecting a shift of $1\sigma_X$ as soon as possible with the EWMA control chart. In addition, the ARL curve of the Shewhart chart for independent observations with the same in-control ARL is depicted. This curve was drawn earlier in Figure 3.1 of the previous chapter.

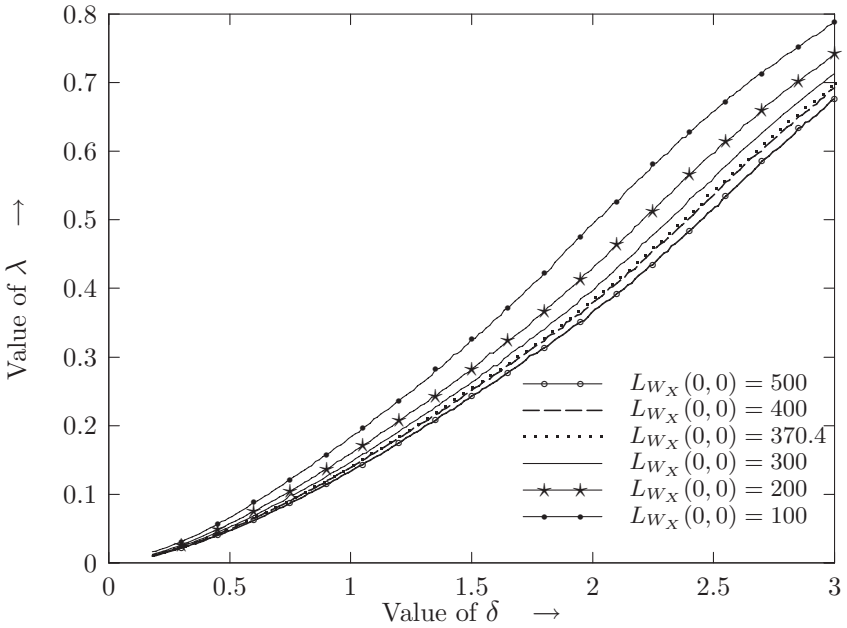


Figure 4.2: Optimal choice of λ for each δ.

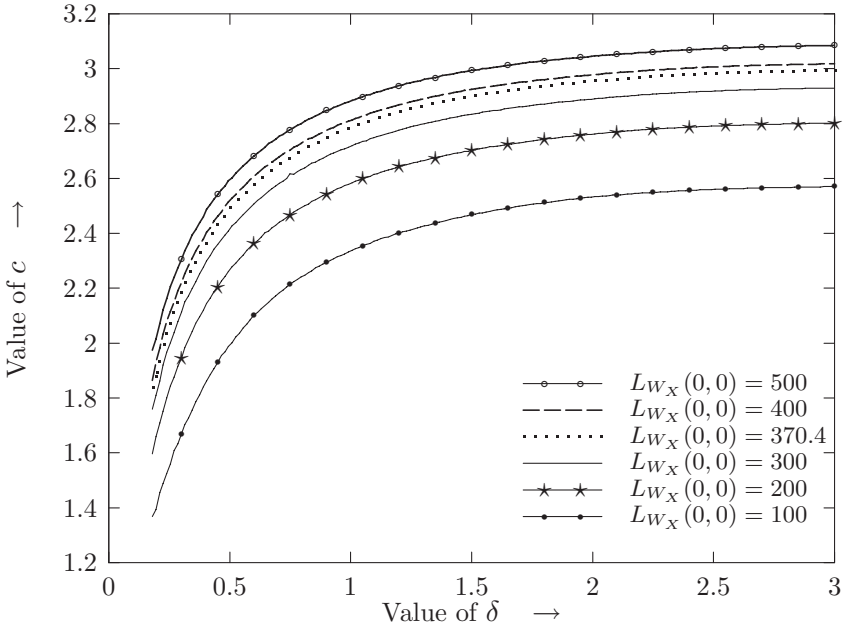


Figure 4.3: Optimal choice of c for each δ.

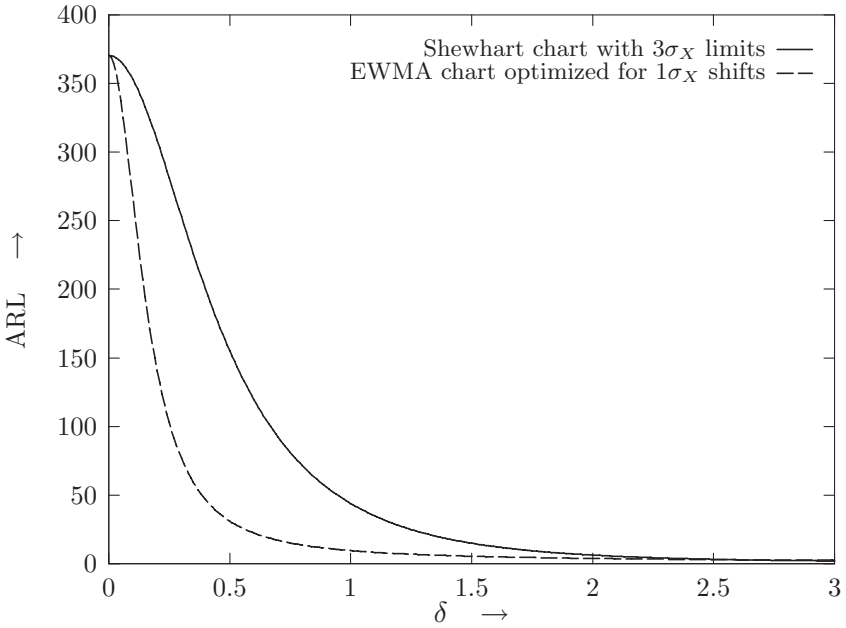


Figure 4.4: ARL curves of the Shewhart chart and an EWMA chart for independent observations.

From Figure 4.4, we conclude that a shift of size $1\sigma_X$ in the mean of the observations is, on average, detected much earlier on the EWMA chart. The out-of-control ARL of the EWMA chart at $\delta = 1$ is 9.58, whereas the out-of-control ARL of the Shewhart chart for independent observations is 43.89. However, for values of $\delta > 2.6$, the Shewhart chart is slightly more efficient. The difference is so small that this is not visible in Figure 4.4.

In conclusion, the EWMA is more sensitive to small shifts in the mean of a series of observations, whereas the Shewhart chart is a little more sensitive in detecting large shifts in the mean. In Section 4.6, the sensitivity of the EWMA control chart to small changes in the mean is explained.

Lucas and Saccucci (1990) describe three enhancements of the EWMA chart that were presented earlier by Lucas and Crosier (1982) for use with the CUSUM chart. These include a so-called FIR feature, which makes the control chart more sensitive at startup, a combined Shewhart EWMA chart, which is sensitive in detecting both large and small shifts in μ and robust EWMA, that provides protection against outliers.

In the next section, we will investigate how the ARL of the EWMA control chart is affected by AR(1) dependence, when the chart is designed as if the data were independent.

4.2 Effect of ignoring serial correlation

In this section, we will simulate the effect of ignoring AR(1) dependence in the data on the behavior of the EWMA control chart. To this end, we will assume that an EWMA control chart has been designed for independent data. We will use the same values for the parameters c and λ as in the previous section. An SPC practitioner who uses such a chart (un)knowingly with AR(1) data, will interpret an out-of-control signal as in Figure 4.4. For example: an out-of-control signal will, on average, occur once every 370 observations if the process is actually in control, and an out-of-control signal will occur, on average, every 10 observations if there is a shift of one standard deviation in the mean. However, if the data is correlated, these interpretations are not valid anymore. This is illustrated by Table 4.1, where we tabulated for selected values of ϕ the in-control ARL and the out-of-control ARL when there has been a shift of size $1\sigma_Y$ in the mean of AR(1) data. The ARLs are simulated using 100,000 replications, except for the ARL(0) corresponding to $\phi = -0.3$; there we used 1,000 replications. The bracketed numbers are the corresponding standard errors.

Table 4.1: ARL values of an i.i.d. EWMA chart, applied to various AR(1) processes.

ϕ	σ_Y	$\mathrm{E}\left[\overline{MR}/d_2(2)\right]$	ARL($0\,\sigma_Y$)	ARL($1\,\sigma_Y$)
-0.9	2.2942	3.1623	—.—	—.—
-0.6	1.2500	1.5811	—.—	—.—
-0.3	1.0483	1.1952	33119.69 (1069.60)	11.96 (0.02)
0.0	1.0000	1.0000	365.56 (1.16)	9.34 (0.02)
0.3	1.0483	0.8771	42.55 (0.13)	7.42 (0.02)
0.6	1.2500	0.7906	12.64 (0.04)	5.84 (0.02)
0.9	2.2942	0.7255	5.86 (0.02)	3.59 (0.01)

The entries in Table 4.1 have been arrived at assuming that the standard

deviation of the observations is estimated using $\overline{MR}/d_2(2)$. Comparing the true value of σ_Y (column 2) with the expectation of its estimator (column 3), we noted earlier in Section 3.3 that $\overline{MR}/d_2(2)$ is positively biased in case of negative ϕ, and negatively biased in case of positive ϕ. Hence, the control limits for the EWMA chart will be too wide in case of negative autocorrelation. In fact, the width of the control limits for $\phi = -0.9$ and $\phi = -0.6$ turned out to be so large that it was not feasible to simulate the ARL in these cases. In case of $\phi = -0.3$, simulating the run length was so time consuming that it was necessary to cut down the number of replications to 1,000 in the in-control situation.

Despite the missing results for $\phi - 0.9$ and $\phi - 0.6$, the warning from Table 4.1 is compelling. Serial correlation seems to have an even stronger effect on EWMA control charts than on Shewhart control charts. This agrees with the findings of Harris and Ross (1991). Negative autocorrelation makes the control chart very insensitive, positive autocorrelation will result in many false out-of-control signals. An out-of-control signal on an EWMA control chart that was designed for independent observations, but used with AR(1) data is thus seriously misinterpreted.

One important reason for the misinterpretation is the bias in the estimation of σ_Y, a problem that is relatively easily resolved. The question remains how an out-of-control signal on an EWMA control chart that accounts for serial correlation must be interpreted.

In the Sections 4.3 through 4.5, we will discuss EWMA control charts that account for AR(1) dependence in the data. In Section 4.3, it is discussed how to modify the limits of the EWMA chart to allow for first-order autoregressive correlation in the data. In Section 4.4, the EWMA residuals chart is discussed, and in Section 4.5 the EWMA control chart for modified residuals is discussed. In Section 4.6, the ARL behavior of these charts will be compared, assuming that all model parameters are known.

4.3 The modified EWMA chart

In the following sections, we return to the case in which successive observations are correlated. As in the previous chapter, the AR(1) case will be discussed. In Appendix A, ARL tables of control charts for other time series models are presented.

In Section 3.4, we discussed how the Shewhart chart can be modified to account for AR(1) data. Schmid (1997b) employs the same reasoning for the EWMA control chart. In this section, we will consider this approach.

Compared to the EWMA chart for i.i.d. observations, the limits of the EWMA chart for correlated observations are widened to account for the increase in the variance. A second adjustment is required due to the fact that the ARL behavior of an EWMA chart of independent observations is not the same as that of an EWMA chart for correlated observations. For a proper evaluation of the effect of serial correlation on the EWMA chart, the limits will be adjusted so that the in-control ARL of the EWMA chart for dependent observations equals a predefined in-control ARL of the EWMA chart for independent observations. The resulting control chart is called the *modified EWMA chart*.

The statistic that will be plotted in the control chart at time t is an EWMA of the correlated data, and will be denoted by $W_{Y,t}$. In Zhang (1998) it was shown that if the sequence $\{Y_t\}$ is stationary, then $\{W_{Y,t}\}$ is asymptotically stationary. Successive realizations of $\{W_{Y,t}\}$ are generated by

$$W_{Y,t} = \lambda Y_t + (1-\lambda) W_{Y,t-1},$$

where the sequence $\{Y_t\}$ consists of AR(1) observations, generated by model (2.3)

$$Y_t - \mu_t = \phi(Y_{t-1} - \mu_{t-1}) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z},$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d. disturbances, $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ for $t \in \mathbb{Z}$. Suppose that a special cause of variation may occur at some unknown time T, resulting in a persistent shift in $\mathrm{E}(Y_t) = \mu_t$ of size $\delta\sigma_Y$:

$$\mu_t = \begin{cases} \mu & \text{for } t < T \\ \mu + \delta\sigma_Y & \text{for } t \geq T. \end{cases} \tag{4.9}$$

We will investigate how quickly, on average, such a change is detected by monitoring the sequence $\{W_{Y,t}\}$.

For determining the control limits for the EWMA chart for AR(1) data,

we derive the variance of $W_{Y,t}$.

$$
\begin{aligned}
&\mathrm{Var}(W_{Y,t}) \\
&= \mathrm{Var}\left(\lambda\sum_{i=0}^{t-1}(1-\lambda)^i Y_{t-i} + (1-\lambda)^t W_{Y,0}\right) \\
&= \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2t}\right] + \\
&\quad + 2\lambda^2\frac{\sigma_\varepsilon^2}{1-\phi^2}\sum_{i=0}^{t-2}\sum_{j=i+1}^{t-1}(1-\lambda)^{i+j}\phi^{j-i} \\
&= \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2t}\right] + \\
&\quad + 2\lambda^2\frac{\sigma_\varepsilon^2}{1-\phi^2}\sum_{i=0}^{t-2}\left(\frac{1-\lambda}{\phi}\right)^i\sum_{j=i+1}^{t-1}((1-\lambda)\phi)^j \\
&= \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2t}\right] + \\
&\quad + 2\frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda^2}{1-\phi(1-\lambda)}\right)\left\{\phi(1-\lambda)\sum_{i=0}^{t-2}(1-\lambda)^{2i} + \right. \\
&\qquad\qquad \left. - (\phi(1-\lambda))^t\sum_{i=0}^{t-2}\left(\frac{1-\lambda}{\phi}\right)^i\right\} \\
&= \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2t}\right] + \\
&\quad + 2\frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda}{2-\lambda}\right)\left(\frac{\phi(1-\lambda)}{1-\phi(1-\lambda)}\right)\left[1-(1-\lambda)^{2t-2}\right] \\
&\quad - 2\frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda^2}{1-\phi(1-\lambda)}\right)\left(\frac{\phi^{t+1}(1-\lambda)^t}{\phi+\lambda-1}\right)\left[1-\left(\frac{1-\lambda}{\phi}\right)^{t-1}\right]
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda}{2-\lambda}\right)\left(\frac{1+\phi(1-\lambda)}{1-\phi(1-\lambda)}\right)\left[1-(1-\lambda)^{2t}\right] + \\
& + 2\frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda}{2-\lambda}\right)\left(\frac{\phi(1-\lambda)}{1-\phi(1-\lambda)}\right)\left[(1-\lambda)^{2t}-(1-\lambda)^{2t-2}\right] \\
& - 2\frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda^2}{1-\phi(1-\lambda)}\right)\left(\frac{\phi^2(1-\lambda)^t}{\phi+\lambda-1}\right)\left(\phi^{t-1}-(1-\lambda)^{t-1}\right).
\end{aligned}
$$

The variance of $W_{Y,t}$ is not constant over time. However,

$$
\mathrm{Var}(W_{Y,t}) \approx \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\frac{\lambda}{2-\lambda}\right)\left(\frac{1+\phi(1-\lambda)}{1-\phi(1-\lambda)}\right) \qquad \text{for large } t.
$$

Control limits for monitoring the sequence $\{W_{Y,t}\}$ are of the form

$$
\begin{aligned}
\mathrm{LCL}_t &= \mu - c\sigma_{W_{Y,t}} \\
\mathrm{UCL}_t &= \mu + c\sigma_{W_{Y,t}},
\end{aligned}
$$

where either the exact standard deviation of $\{W_{Y,t}\}$ can be used, which will result in control limits that vary over time, or the asymptotic standard deviation, which results in constant control limits. Surprisingly, simulation studies performed by Schmid (1997b) indicate that utilizing the exact variance for the control limits does not always lead to better ARL behavior. His results also show that there is essentially no difference in the ARL behavior. Therefore, in the remainder of this section, we will work with the asymptotic variance.

In order to implement the EWMA chart the two parameters λ and c have to be chosen. Schmid (1997b) presents combinations for λ and c that yield an in-control ARL of 500 for various values of ϕ. For other in-control ARLs, combinations of c and λ can be determined by fixing two points on the ARL curve of the control chart. For the ARL of the modified EWMA chart in case of AR(1) observations we present the following considerations, which are inspired by an excellent article by VanBrackle and Reynolds (1997), where the ARL curve of the modified EWMA chart for ARMA(1,1) data was presented. In Schmid and Schöne (1997) some theoretical results on the EWMA control chart in the presence of autocorrelation are presented.

Let $L_{W_Y}(\delta, u, v)$ denote the ARL of the modified EWMA chart if the first observation is taken after the mean has shifted by an amount of $\delta\sigma_Y$, and that the value of EWMA statistic is u at the first observation, while the corresponding AR(1) observation is v. Then the following holds

$$\begin{aligned} L_{W_Y}(\delta, u, v) \; &= \\ &= \; 1 + \int\limits_{\{|\lambda y + (1-\lambda)u - \mu| \leq c\sigma_{W_Y}\}} L_{W_Y}(\delta, (1-\lambda)u + \lambda y, y) \; \times \\ &\qquad \times \; f\left(y - \phi v - (1-\phi)(\mu + \delta\sigma_Y)\right) \mathrm{d}y \\ &= \; 1 + \frac{1}{\lambda} \int\limits_{\mu - c\sigma_{W_Y}}^{\mu + c\sigma_{W_Y}} L_{W_Y}\left(\delta, w, \frac{w - (1-\lambda)u}{\lambda}\right) \; \times \\ &\qquad \times \; f\left(\frac{w - (1-\lambda)u}{\lambda} - \phi v - (1-\phi)(\mu + \delta\sigma_Y)\right) \mathrm{d}w, \end{aligned}$$

where

$$w = (1-\lambda)u + \lambda y,$$

and $f(\cdot)$ is the probability density function of the disturbances. Compared to the derivation of the ARL curve of the EWMA chart for independent observations in (4.8), the expression above is more complicated, since the information that is needed to go from one time point to another is no longer one-dimensional. In addition to the value of the EWMA statistic, it is also necessary to know the value of the AR(1) observation.

If the first observation is taken at the time of the shift, the corresponding ARL function is obtained as follows

$$\begin{aligned} L^*_{W_Y}(\delta, u^*, v^*) \; &= \; 1 + \frac{1}{\lambda} \int\limits_{\mu - c\sigma_{W_Y}}^{\mu + c\sigma_{W_Y}} L_{W_Y}\left(\delta, w^*, y^*\right) \; \times \\ &\qquad \times \; f\left(y^* - \phi v^* - (1-\phi)\mu - \delta\sigma_Y\right) \mathrm{d}w^*, \end{aligned}$$

where u^* is the last observed value of the EWMA statistic before the shift in the mean occurred, and v^* is the corresponding AR(1) observation. Analogously, w^* is the value of the EWMA statistic at the time of the shift. The corresponding AR(1) observation is denoted by y^* and equals

$$y^* = \frac{w^* - (1-\lambda)u^*}{\lambda}.$$

The dependence of $L^*_{W_Y}(\delta, u^*, v^*)$ on u^* and v^* can be 'averaged out' to obtain

$$L^{**}_{W_Y}(\delta) = \int_{\mu - c\sigma_{W_Y}}^{\mu + c\sigma_{W_Y}} \int_{-\infty}^{\infty} L^*_{W_Y}(\delta, u^*, v^*)\; k(v^*)\, g(u^*)\; \mathrm{d}v^* \mathrm{d}u^*,$$

where $k(\cdot)$ is the probability density function of the last AR(1) observation just before the occurrence of the shift, and $g(\cdot)$ is the truncated probability density function of the last observation of the EWMA just before the shift.

The presentation and the discussion of the ARL curves of the modified EWMA chart are deferred to Section 4.6. In that section, the ARL curves of the modified EWMA chart are compared to the ARL curves of the EWMA chart of residuals and the EWMA chart of modified residuals for various values of the AR-parameter ϕ. But first we discuss the EWMA chart of residuals in Section 4.4, and the EWMA chart of modified residuals in Section 4.5.

4.4 The EWMA chart of residuals

In this section, the EWMA chart for the mean of residuals of a fitted AR(1) process is discussed. As in Section 3.5, residuals $\{e_t\}$ are computed as follows, assuming that the mean of the AR(1) observations equals μ:

$$e_t = Y_t - \mu - \phi(Y_{t-1} - \mu).$$

If μ and ϕ are not known, they have to be replaced by appropriate estimates. However, we will assume that μ and ϕ are known. Concerning the effect of parameter estimation on the performance of control charts, as yet little work has been done. This is also outside the scope of this thesis.

If μ and ϕ are known, $e_t = \mathrm{E}(e_t) + \varepsilon_t$. If $\mathrm{E}(Y_t)$ shifts from μ to $\mu + \delta\sigma_Y$ at time T, we have for the expectation of the residuals

$$\mathrm{E}(e_t) = \begin{cases} 0 & \text{for } t < T \\ \delta\sigma_Y & \text{for } t = T \\ (1-\phi)\delta\sigma_Y & \text{for } t > T, \end{cases}$$

where we recall from Chapter 3 that

$$\sigma_Y = \frac{\sigma_\varepsilon}{\sqrt{1-\phi^2}}$$

when $\{Y_t\}$ is generated by an AR(1) model with 'white noise' disturbances $\{\varepsilon_t\}$, with $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$. The EWMA of the residuals at time t is denoted by $W_{e,t}$ and is computed as follows

$$W_{e,t} = \lambda e_t + (1-\lambda) W_{e,t-1}.$$

Assuming that μ and ϕ are known, the variance of $W_{e,t}$ is, analogously to Equation (4.3), derived as

$$\sigma^2_{W_{e,t}} = \sigma_\varepsilon^2 \left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2t}\right].$$

As an asymptotic approximation,

$$\sigma^2_{W_{e,t}} \approx \sigma_\varepsilon^2 \left(\frac{\lambda}{2-\lambda}\right)$$

may be used. Control limits for the EWMA residuals chart are of the form

$$\begin{aligned} \text{LCL}_t &= \mu - c\sigma_{W_{e,t}} \\ \text{UCL}_t &= \mu + c\sigma_{W_{e,t}}, \end{aligned}$$

where either the exact or the asymptotic expression for the standard deviation of $W_{e,t}$ can be used. The two design parameters λ and c are again chosen by fixing two points on the ARL curve.

In Chapter 3, it was argued that monitoring a serially correlated process by monitoring the residuals of a fitted time series model is appealing. If the time series model is appropriate for the data, then the residuals will be approximately uncorrelated.

In Section 4.1, the ARL curve of a sequence of independent observations was derived. Since under the assumptions made above, the sequence $\{e_t\}$ consists of independently distributed random variables, the ARL curve of the EWMA residuals chart is derived analogously. That is, the ARL of the EWMA residuals chart that started in u, which is denoted by $L_{W_e}(\delta, u)$, satisfies the following integral equation if the first observation is taken after the shift in the mean has occurred

$$L_{W_e}(\delta, u) = 1 + \frac{1}{\lambda} \int_{-c\sigma_{W_e}}^{c\sigma_{W_e}} L_{W_e}(\delta, v)\, f\Big(\frac{v - (1-\lambda)u}{\lambda} - (1-\phi)\delta\sigma_Y \Big)\, \mathrm{d}v,$$

where v is the value of the EWMA statistic that succeeds u. If the first observation is taken at the time of the shift, $L^*_{W_e}(\delta, u^*)$, the ARL of the EWMA residuals chart that started in u^*, satisfies

$$L^*_{W_e}(\delta, u^*) = 1 + \frac{1}{\lambda} \int_{-c\sigma_{W_e}}^{c\sigma_{W_e}} L_{W_e}(\delta, v^*)\, f\Big(\frac{v^* - (1-\lambda)u^*}{\lambda} - \delta\sigma_Y \Big)\, \mathrm{d}v^*,$$

where v^* is the value of the EWMA of residuals at the time of the shift. Note that u^* is the last value of the EWMA statistic just before the shift. The dependence of $L^*_{W_e}(u^*)$ on u^* can be 'integrated out' to obtain

$$L^{**}_{W_e}(\delta) = \int_{-c\sigma_{W_e}}^{c\sigma_{W_e}} L^*_{W_e}(\delta, u^*)\, g(u^*) \mathrm{d}u^*,$$

where $g(\cdot)$ is the density function of the EWMA statistic that is observed just before the shift. The resulting ARL curves are drawn, together with the ARL curves of the modified EWMA chart and the EWMA modified residuals chart in Section 4.6. In the next section, the EWMA modified residuals chart is discussed.

4.5 The EWMA chart of modified residuals

In Chapter 3, a modified residual was proposed that can be used to monitor serially correlated data. In this section, we discuss monitoring the mean of an AR(1) process with an EWMA control chart of the modified residuals.

Recall from Section 3.6 that the modified residuals $\{u_t\}$ of an AR(1) process are computed as

$$u_t \equiv Y_t - \phi Y_{t-1} + \phi\hat{\mu}_t, \tag{4.10}$$

where $\hat{\mu}_t$ is an estimator of the level of the process that quickly responds to changes in the 'local mean' μ_t. Simulation studies (which are not included in this thesis) have shown that an EWMA of the observations is a better choice than an unweighted moving average, in terms of ARL behavior of the resulting control chart.

Assuming that μ and ϕ are known, and that $\hat{\mu}_t$ is an EWMA of previous observations with parameter λ', the expectation of u_t equals

$$\mathrm{E}(u_t) = \begin{cases} (1-\phi)\mu \;+\; \phi\mu\,\lambda' \sum_{i=0}^{\infty}(1-\lambda')^i \;=\; \mu & \text{for } t < T \\ \mu \;+\; (1+\phi\lambda')\delta\sigma_Y & \text{for } t = T \\ \mu \;+\; (1-\phi)\delta\sigma_Y \;+\; \phi\delta\sigma_Y\,\lambda' \sum_{i=0}^{t-T}(1-\lambda')^i & \text{for } t > T, \end{cases}$$

where T is again the unknown time point when a shift in $\mathrm{E}(Y_t)$ occurs. Note that the mean of the sequence $\{u_t\}$ is μ before a possible shift in the mean of the process. After a shift in the mean of the process, the mean of $\{u_t\}$ converges to $\mu + \delta\sigma_Y$ for $t \gg T$.

The EWMA of $\{u_t\}$ at time t is denoted by $W_{u,t}$ and is computed as

$$W_{u,t} = \lambda u_t + (1-\lambda)W_{u,t-1} \qquad \text{for } t = 0, 1, 2, \cdots.$$

The ARL curves for the EWMA chart of modified residuals of AR(1) data are obtained by simulation. In Section 4.6, these are compared to the ARL curves of the modified EWMA chart and the EWMA chart of residuals for various values of ϕ.

4.6 Discussion

In the previous sections, various EWMA control charts for monitoring the mean of AR(1) observations were discussed. In the next subsection, the ARL curves corresponding to these control charts are presented. We will see that the ARL behavior of EWMA-type control charts is in general much better than the ARL behavior of Shewhart-type control charts for small shifts in the mean. This was also observed in Section 4.1 of this chapter. In Subsection 4.6.2, we will explain why this is the case.

4.6.1 ARL comparison

In the previous three sections, respectively the modified EWMA, the EWMA chart of residuals and the EWMA chart of modified residuals were discussed. For each of the charts, considerations on the computation of ARL curves were presented. However, compared to the ARL curves in Chapter 3, the computation time needed to compute the ARL curves of the EWMA-type control charts rose dramatically. Therefore, we decided to

evaluate the ARL curves of the EWMA charts by means of simulation. In Figures 4.5 through 4.10, the ARL curves of the three charts are compared to the ARL curve of the EWMA control chart for the mean of independent observations, for values of $\phi = -0.9, -0.6, -0.3, 0.3, 0.6, 0.9$. The value of the EWMA parameter λ was chosen equal to 0.2. For the EWMA chart of modified residuals, we took $\lambda' = 0.1$, as in Section 3.6. Each of the curves in Figures 4.5 through 4.10 consists of 101 points, and each point is the mean of 100,000 run lengths.

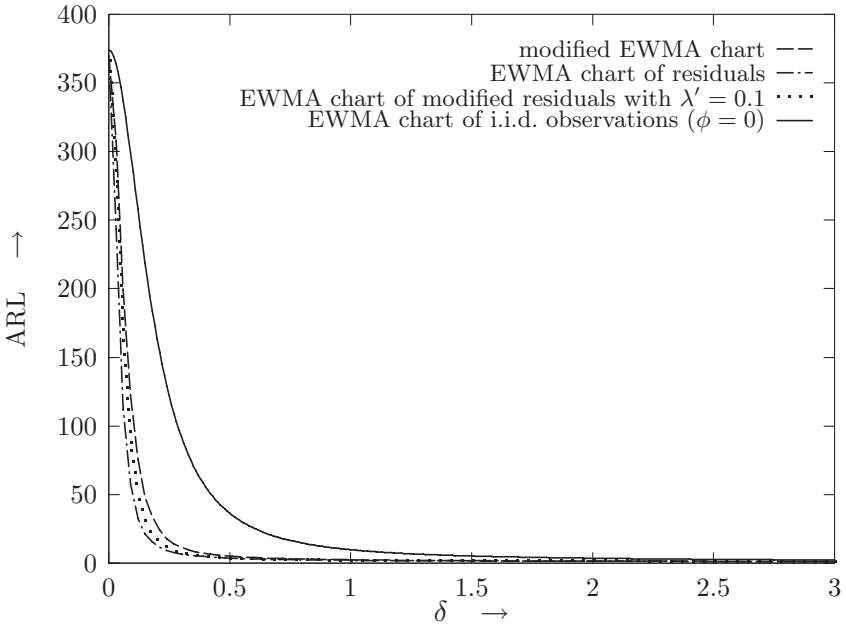


Figure 4.5: Various ARL curves for AR(1) process with $\phi = -0.9$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

From Figures 4.5 through 4.10, we conclude that, generally speaking, the effect of first-order autocorrelation on EWMA charts is the same as on Shewhart-type charts. Compared to the i.i.d case, the ARL behavior is better for negative ϕ, whereas ARL behavior is worse for positive ϕ.

The difference in ARL behavior of the three charts is small for negative ϕ. Also, for positive ϕ, the ARL behavior of the modified EWMA chart and the ARL behavior of the EWMA of modified residuals does not differ much. For large ϕ, the EWMA chart of residuals is performing worse than the other two. These conclusions agree with the findings of Schmid (1997a) and Schmid (1997b).

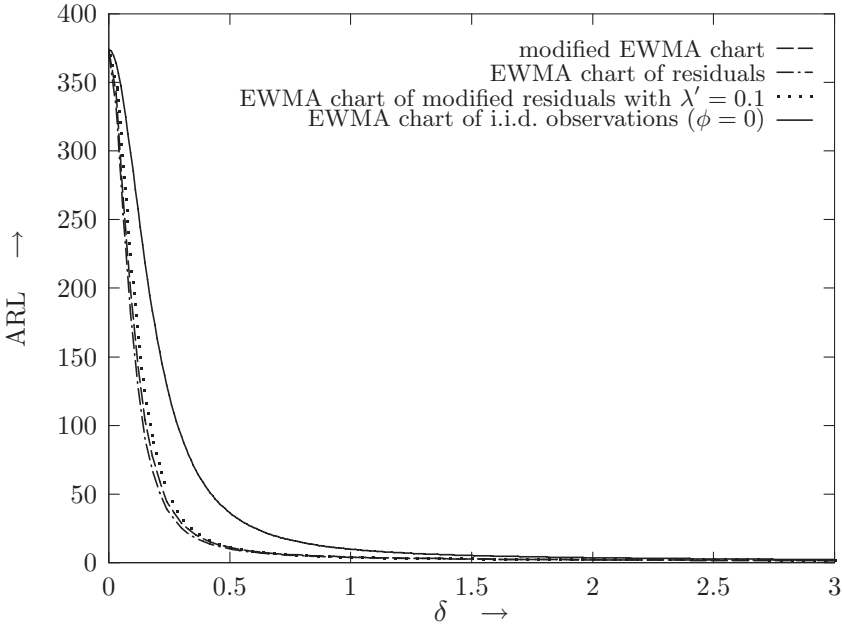


Figure 4.6: Various ARL curves for AR(1) process with $\phi = -0.6$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

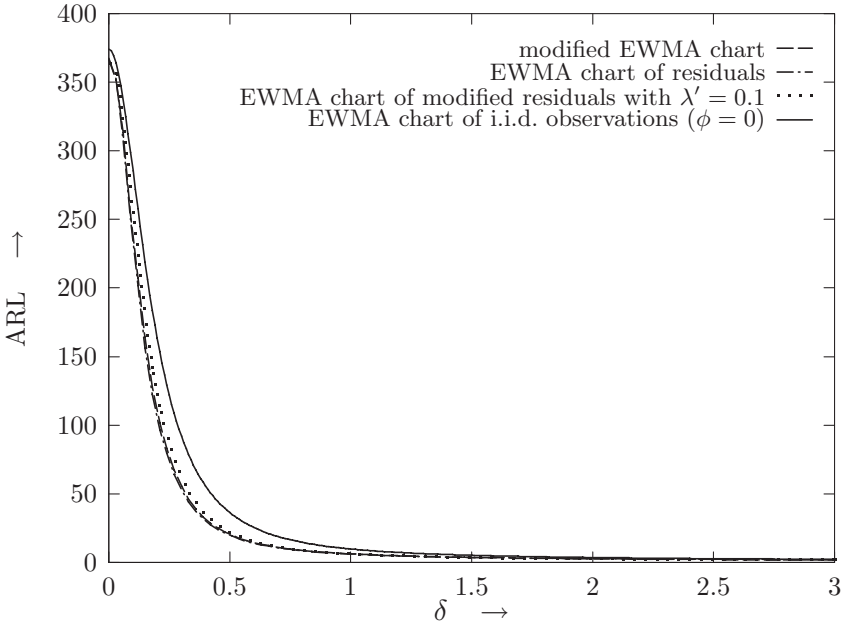


Figure 4.7: Various ARL curves for AR(1) process with $\phi = -0.3$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

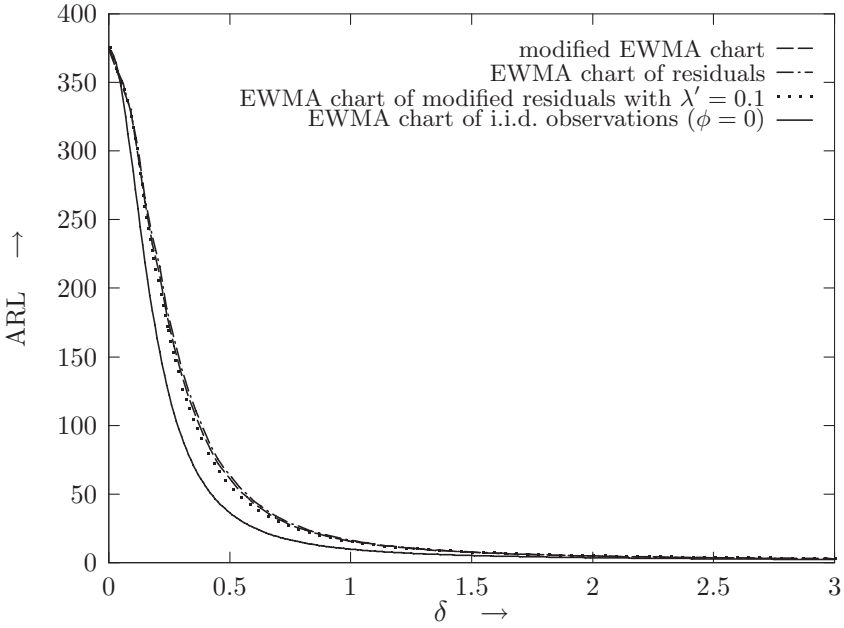


Figure 4.8: Various ARL curves for AR(1) process with $\phi = 0.3$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

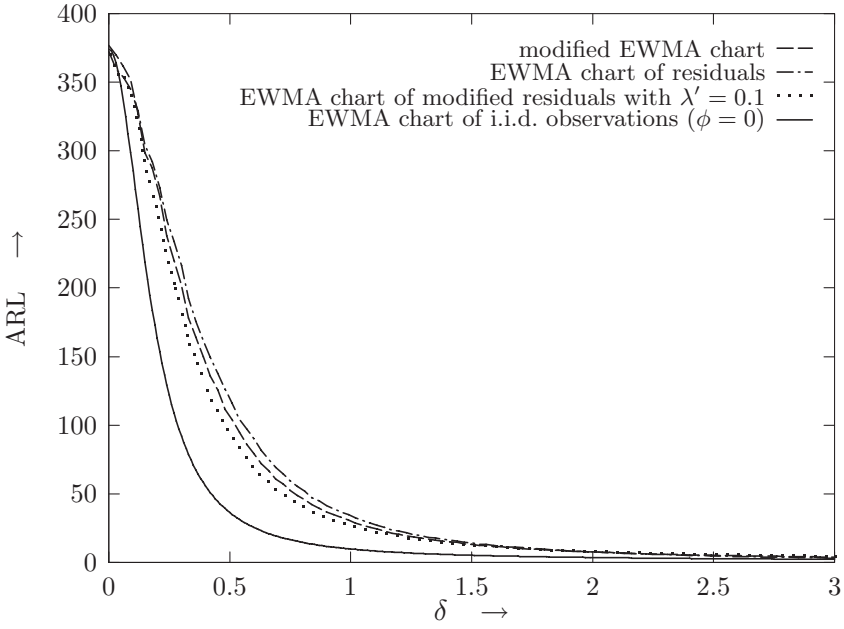


Figure 4.9: Various ARL curves for AR(1) process with $\phi = 0.6$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

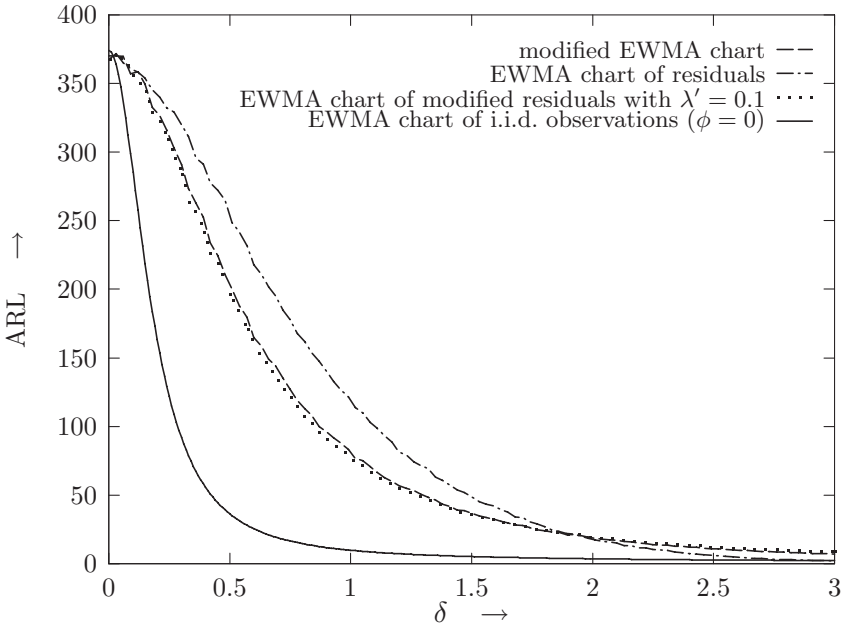


Figure 4.10: Various ARL curves for AR(1) process with $\phi = 0.9$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

Note that for small ϕ the ARL behavior of EWMA-type control charts appears to be better than the ARL behavior of Shewhart-type charts (compare to Figures 3.21 through 3.26 of the previous chapter). However, for large ϕ, the ARL performance of EWMA-type control charts is comparable to Shewhart-type control charts.

Hence, for small to moderate levels of first-order autoregressive serial correlation, it is advisable to use an EWMA-type control chart to monitor the process for changes in the mean. However, for larger values of ϕ, a Shewhart-type control chart might be preferred because it is easier to compute and to interpret.

4.6.2 Relationship between the EWMA and the modified Shewhart chart

In this subsection the modified Shewhart chart for AR(1) data is compared to the EWMA control chart for independent observations. It turns out that they are very similar. In Figure 4.4, it is shown that the ARL behavior of the EWMA control chart is better than the ARL performance of the Shewhart chart for independent observations. Therefore, one might expect

that the ARL performance of the modified Shewhart chart for AR(1) data is also better than that of the Shewhart chart for independent observations. However, in Chapter 3, it was argued that the ARL performance of the modified Shewhart chart for AR(1) data is comparable to that of the Shewhart chart for independent observations. In this subsection, it is explained why the ARL performance of the modified Shewhart chart for AR(1) data is not as good as the ARL performance of the EWMA chart for independent observations.

Suppose that we have a sequence of i.i.d. observations $\{X_t\}$, which satisfy (3.1) and (3.2). The EWMA statistic at time t is denoted by $W_{X,t}$ and is constructed as follows

$$W_{X,t} = (1-\lambda)W_{X,t-1} + \lambda X_t \qquad \text{for } t = 1, 2, \cdots, \tag{4.11}$$

where $\lambda \in (0, 1)$. The EWMA chart may be started by setting $W_{X,0}$ equal to a target value or (an estimator of) μ. If $W_{X,0} = \mu$, it is easy to verify that $\mathrm{E}(W_{X,t}) = \mu$ for $t = 0, 1, 2, \cdots$,T. For $t \geq T$ we have

$$\mathrm{E}(W_{X,t}) = \mu + \left\{1 - (1-\lambda)^{t-T+1}\right\} \delta\sigma_X,$$

which approximately equals $\mu + \delta\sigma_X$ for $t \gg T$. Hence, $\{\mathrm{E}(W_{X,t})\}$, the sequence of expected values of the EWMA statistic, approximately mimics $\{\mathrm{E}(X_t)\}$.

That there is a relation between (4.11) and the AR(1) model (2.3) becomes clear if we subtract μ_t on both sides of Equation (4.11):

$$W_{X,t} - \mu_t = (1-\lambda)(W_{X,t-1} - \mu_t) + \lambda(X_t - \mu_t). \tag{4.12}$$

Note that $\{\lambda(X_t - \mu_t)\}$ may be considered a white noise process with variance $\lambda^2\sigma_X^2$. As long as $\mu_t = \mu$ for $t = 0, 1, 2, \cdots$, we conclude that computing the EWMA statistic is equivalent to converting a sequence of i.i.d. observations into an AR(1) sequence with AR parameter $\phi = 1 - \lambda$. Moreover, using an EWMA control chart is in a certain sense equivalent to monitoring AR(1) observations using modified Shewhart chart. In Lucas and Saccucci (1990) it was shown that the properties of the EWMA chart are very close to those of CUSUM schemes. Small shifts in the mean are, on average, more quickly detected on an EWMA control chart than on a standard Shewhart control scheme.

The argument above seems somewhat in contradiction with the conclusions of Section 3.4. There we observed that for $\phi = 0.9$ the modified

Shewhart control chart for AR(1) data is not very sensitive in detecting small changes in the mean. A value of $\phi = 0.9$ corresponds to a value of $\lambda = 0.1$. For this value of λ, an EWMA chart is much more sensitive than a Shewhart control chart for detecting small shifts in the mean of a sequence of independent observations. The discrepancy between these results can be explained by studying a signal-to-noise ratio: a number that relates the size of the shift to the standard deviation of the process. This ratio allows us to compare shifts in means in processes with different variances. In Section 3.5.2, the signal-to-noise ratio was defined as the size of the shift divided by the standard deviation of the process.

For the modified Shewhart chart, the size of the shift is $\delta\sigma_Y$ after time T. The standard deviation of the AR(1) process is σ_Y. Hence, the signal-to-noise ratio is δ.

In the case of the EWMA control chart, the size of the shift in $\mathrm{E}(W_{X,t})$ approximately equals $\delta\sigma_X$ for $t \gg T$. The variance of the EWMA statistic is

$$\mathrm{Var}(W_{X,t}) = \sigma_X^2 \left(\frac{\lambda}{2-\lambda}\right) \left\{1 - (1-\lambda)^{2t}\right\}.$$

Hence, for large $t \gg T$, the signal-to-noise ratio equals approximately $\delta\sqrt{(2-\lambda)/\lambda}$, which is larger than δ for $0 < \lambda < 1$. A popular choice of λ for the EWMA chart is $\lambda = 0.1$. In this case, the signal-to-noise ratio is approximately $\delta\sqrt{19}$ for large $t \gg T$.

Hence, computing the EWMA of a sequence of i.i.d. observations leaves the pattern of expectations approximately unaltered, but improves the signal-to-noise ratio. This is combined with the introduction of first-order positive autocorrelation. From the ARL curves in the previous subsection, we learned that first-order positive autocorrelation has a (small) negative effect on the ARL performance. Apparently, this effect is offset by the positive effect of the improved signal-to-noise ratio. We conclude that the efficiency of the EWMA control chart is not the result of the autocorrelation that is introduced. It is the improvement of the signal-to-noise ratio that makes the EWMA control chart efficient.

Chapter 5

CUSUM-type control charts for the mean

In the previous chapters, we discussed control charts for the mean of AR(1) data. In Chapter 3, Shewhart-type control charts were discussed. In Chapter 4, the effect of first-order autoregressive correlation on the performance of EWMA-type control charts was studied. In this chapter, we will investigate how CUSUM-type control charts are affected by serial correlation in the data. As in the Chapters 3 and 4, we will assume that the measurements of a quality characteristic of interest are generated by an AR(1) model.

Shewhart-type control charts only use the last sample to monitor the process. These charts have no memory: previous observations do not influence the probability of future out-of-control signals. Large shifts in the mean are quickly detected by Shewhart-type control charts. However, Shewhart-type control charts are not sensitive in detecting small shifts in the mean. Additional so-called runs rules are sometimes applied to improve the performance of the Shewhart chart (see for example Champ and Woodall (1997) or Does and Schriever (1992)). These runs rules introduce some 'memory' which results in faster detection of small shifts in the mean.

In EWMA-type control charts the process is monitored using a weighted mean of all previous observations. The weight attached to recent observations is high compared to the weights of older observations. The weights decline exponentially as the observations get older and older. The EWMA parameter λ determines the memory of the EWMA control chart.

The CUSUM chart, which was originally introduced by Page (1954), uses an unweighted sum of all previous observations. This chart has a

rather long memory. In Section 5.1, the CUSUM chart for the mean of independent observations is discussed. As in the previous two chapters, the behavior of this control chart will serve as a benchmark for CUSUM charts for dependent observations, which will be discussed in Section 5.3 through 5.5. In Section 5.3, the behavior of a CUSUM chart with limits which are modified to account for AR(1) data, is studied. In Section 5.4, a CUSUM chart for the residuals of a fitted AR(1) model is discussed. A CUSUM of modified residuals for monitoring the mean of AR(1) data is presented in Section 5.5. In Section 5.6, the ARL curves of these charts are compared to the ARL curve of the CUSUM for independent observations.

5.1 The CUSUM chart for i.i.d. observations

In this section, we will investigate the properties of the CUSUM chart for the mean of $\{X_t\}$, a sequence of independently distributed observations, generated by model (3.1). Page (1954) proposed plotting a cumulative sum

$$S_t = \sum_{i=0}^{t} (X_i - \mu_0)$$

on a chart, where μ_0 indicates some target mean level of the process. If the mean of the process μ_t is larger than μ_0, the mean path of $\{S_t\}$ is upward sloping. On the other hand, if $\mu_t < \mu_0$, the sequence $\{S_t\}$ will show a downward sloping trend. Hence, when the cumulative sum is plotted against time, changes in the mean of the process are visible as changes of the slope. In the case of one-sided testing for a positive change in the mean of $\{X_t\}$, Page (1954) proposed to undertake action at time t if

$$S_t - \min_{0 \le i \le t} S_i \ge h$$

for a certain fixed value of h. This approach is equivalent to monitoring S'_t, which is defined as

$$S'_t = \max(S'_{t-1} + x_t, 0),$$

since $S'_t = 0$ if $S_t \le \min_{0 \le i \le t} S_i$. An out-of-control signal is issued at time t if $S'_t \ge h$.

In the case of two-sided testing, an out-of-control signal is given if either

$$S_t - \min_{0\le i\le t} S_i \ge h \quad \text{or} \quad \max_{0\le i\le t} S_i - S_t \ge h'$$

for constants h and h'. A method of judging the CUSUM that became popular after Barnard's (1959) article, is the so-called 'V-mask' scheme. In Figure 5.1, an example of a V-mask is drawn.

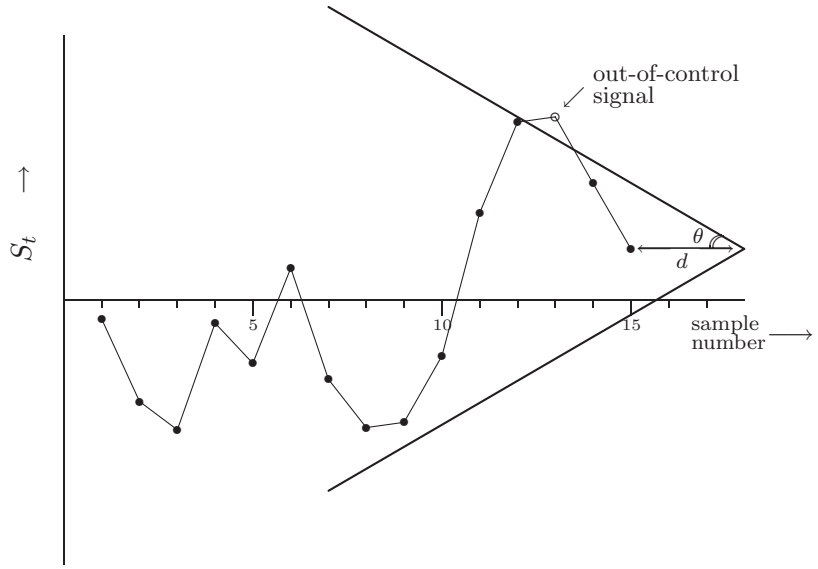
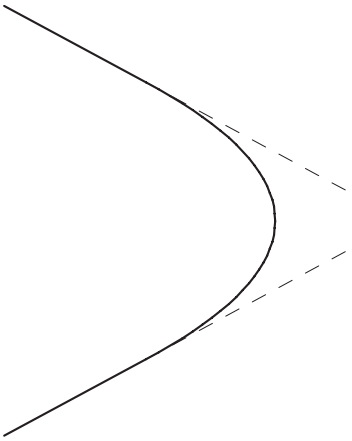


Figure 5.1: An example of a V-mask CUSUM monitoring scheme.

With each new observation that becomes available, the turned-down V is moved to a distance d of the current value of the sum S_t. The angle between the arms of the V is 2θ. As long as all previous values of the sum are within the arms of the V-mask, no out-of-control signal is given. If one or more points cross the lower arm of the V-mask, an out-of-control signal indicating a positive shift in the mean is given. One or more points crossing the upper arm of the V-mask indicate a negative shift in the mean. In Goldsmith and Whitfield (1961) and Lucas (1973), it is recommended to scale one unit of the horizontal axis of the graph of the V-mask CUSUM chart equivalent to $2\sigma_X$ on the vertical axis. With this scaling, a $2\sigma_X$ shift in the mean gives a 45° trend on the CUSUM plot. As we will see later in this section, for small shifts in the mean, the ARL behavior of the CUSUM is much better than the ARL of the Shewhart chart. However, the Shewhart chart is faster



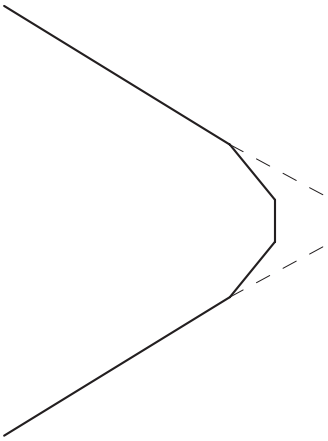


Figure 5.2: Semi-parabolic V-mask. **Figure 5.3:** Snub-nosed V-mask.

in detecting large shifts in the mean. Modifications of the V-mask scheme that improve the performance of the CUSUM for large shifts include the semi-parabolic CUSUM mask, which replaces the last end of the vertex of the V-mask by a parabolic, and a snub-nosed V-mask, which obtained its name for obvious reasons, see Figure 5.2 and 5.3.

In Kemp (1961), it is shown that the use of a V-mask on a CUSUM chart is equivalent to monitoring two sums

$$\begin{aligned} S_{L_t} &= \max(S_{L_{t-1}} - Z_t - k, 0) \\ S_{H_t} &= \max(S_{H_{t-1}} + Z_t - k, 0) \end{aligned} \tag{5.1}$$

with a horizontal limit, placed at h. The variable Z_t denotes the standardized variable

$$Z_t = \frac{X_t - \mu_0}{\sigma_X}.$$

As soon as $S_L \geq h$ or $S_H \geq h$, an out-of-control signal occurs. The constant k was called the *reference value* by Ewan and Kemp (1960), and h is called the *decision interval*. This CUSUM monitoring scheme is called a *Decision Interval Scheme* (DIS). The relationship between the parameters d and θ of the V-mask monitoring scheme and k and h of an equivalent DIS will

be discussed later in this section. An example of a DIS is depicted in Figure 5.4.

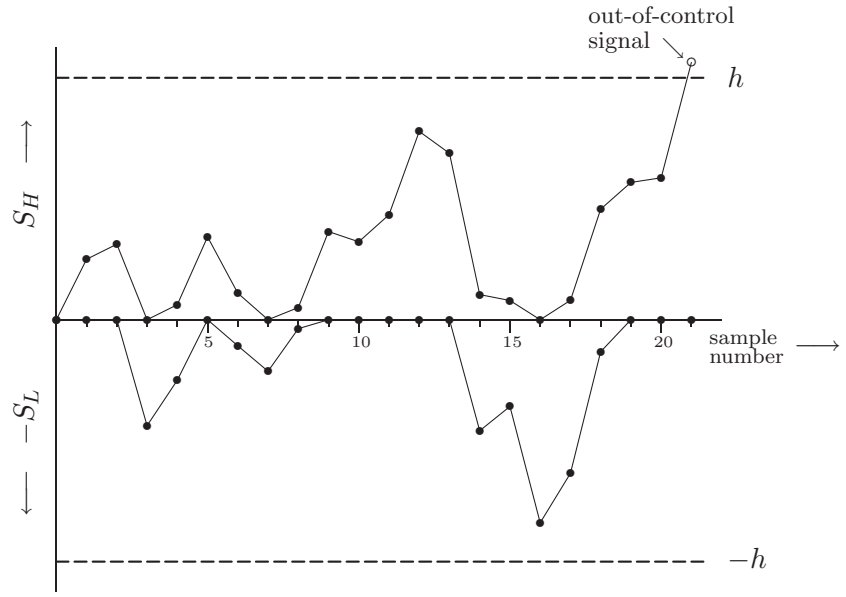


Figure 5.4: An example of a DIS CUSUM monitoring scheme.

When plotting a two-sided DIS it is convenient to plot both sums in one graph. Therefore, as in Figure 5.4, it is customary to plot S_H (which takes on only nonnegative values), and $-S_L$ (which takes on only nonpositive values).

Compared to the V-mask CUSUM scheme, a practical advantage of a DIS is that the cumulative sums do not run off the paper. More convincingly, Montgomery (1996) strongly advises not to use the V-mask procedure because of the following reasons:

1. being a two-sided scheme, the V-mask is not very useful in one-sided process monitoring problems;

2. the useful FIR feature (see Lucas and Crosier (1982)) cannot be applied to the V-mask;

3. it is not clear how far backwards the arms of the V-mask should extend, this complicates interpretation of the V-mask;

4. designing the V-mask using Johnson's (1961) approach, which will be discussed in the following subsection, produces inaccurate results.

In the remainder of this thesis, we will only utilize the DIS to judge the CUSUM. Therefore, when we speak of a 'CUSUM chart' we are referring to the combination of the sums S_L, S_H and the corresponding DIS with reference value k and decision interval h, unless explicitly stated otherwise.

5.1.1 Relation between the CUSUM and the SPRT

Monitoring a sequence of independent (normally) distributed random variables with a CUSUM procedure is closely related to Wald's *Sequential Probability Ratio Test* (SPRT). Johnson (1961) obtained simple formulas for the design parameters d and θ of the V-mask scheme by noting that application of the V-mask monitoring scheme is approximately equivalent to the application of an SPRT 'in reverse'. But, as was demonstrated by Woodall and Adams (1993), designing the V-mask scheme by these methods does not produce the intended values for the ARL.

Nevertheless, the analogy between the CUSUM control chart and the SPRT provides a useful insight in the theoretical foundations of the CUSUM chart. Let us therefore consider this approach. Suppose that we have a sequence of m observations $X_1, X_2, \cdots, X_m$, which are independently and identically normally distributed with expectation μ and variance σ_X^2. The SPRT can be used to discriminate between simple hypotheses. Suppose that we want to distinguish between three hypotheses

$$\mathrm{H}_{-1} : \mu = \mu_0 - \delta \qquad \mathrm{H}_0 : \mu = \mu_0 \qquad \mathrm{H}_1 : \mu = \mu_0 + \delta$$

Define reversed standardized variables $Z'_1, X'_2, \cdots, Z'_m$ as

$$Z'_1 = \frac{X_m - \mu_0}{\sigma_X}, \quad Z'_2 = \frac{X_{m-1} - \mu_0}{\sigma_X}, \ldots$$

The likelihood-ratio test statistic for testing H_0 against H_{-1}, based on $Z'_1, \cdots, Z'_r$ is

$$\begin{aligned}\lambda_r &= \frac{\left(\frac{1}{2\pi}\right)^{\frac{r}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{r}(Z_i')^2\right)}{\left(\frac{1}{2\pi}\right)^{\frac{r}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{r}(Z_i'+\delta)^2\right)} \\ &= \exp\left(\delta\sum_{i=1}^{r} Z_i' + \frac{r\delta^2}{2}\right).\end{aligned}$$

For large λ_r we find it very unlikely that H_{-1} is true. On the other hand, the smaller λ_r, the more likely it becomes that H_{-1} is true. This suggests that it is possible to define two constants A and B with $A < B$, for which we decide to accept H_{-1} if $\lambda_r < A$, whereas we decide to accept H_0 if $\lambda_r > B$. If $A \leq \lambda_r \leq B$, it is decided to take observation Z'_{r+1} and compare the likelihood ratio λ_{r+1} to these limits. The choice of values for A and B is closely related to α, the probability to accept H_{-1} while H_0 is true, and β, the probability of falsely accepting H_0. However, in general, it is difficult to determine A and B by fixing α and β. Fortunately, a good approximation can be found. It can be shown (see for example Mood, Graybill and Boes (1974)) that if the following values are used

$$A' = \frac{\alpha}{1-\beta} \quad \text{and} \quad B' = \frac{1-\alpha}{\beta},$$

that α' and β', the corresponding probabilities of errors of the first and the second kind respectively, satisfy

$$\alpha' + \beta' \leq \alpha + \beta.$$

When A' and B' are used as limits, we decide to accept H_{-1} at observation r if

$$\sum_{i=1}^{r} Z_i' < -\frac{r\delta}{2} + \frac{\log(\alpha)}{\delta} - \frac{\log(1-\beta)}{\delta},$$

whereas H_0 is accepted if

$$\sum_{i=1}^{r} Z_i' > -\frac{r\delta}{2} + \frac{\log(1-\alpha)}{\delta} - \frac{\log(\beta)}{\delta}.$$

These formulas can be simplified even further if we realize that in the case of monitoring for a shift in the mean, β must be chosen equal to zero. We never decide to accept H_0; we always keep on sampling if it is not decided to accept the hypotheses that there has been a shift in the mean. Hence, we cannot falsely accept H_0. This effectively leads to the following decision rule: keep on sampling until

$$\sum_{i=1}^{r} Z_i' < -\frac{r\delta}{2} + \frac{\log(\alpha)}{\delta}. \tag{5.2}$$

The hypothesis that there has been a negative shift in the mean is accepted as soon as inequality (5.2) holds.

Analogously, for discriminating between H_0 and H_1 with the same approximate error probabilities, the following rule can be derived: keep on sampling until

$$\sum_{i=1}^{r} Z_i' > \frac{r\delta}{2} - \frac{\log(\alpha)}{\delta}, \tag{5.3}$$

and accept the hypothesis that there has been a positive shift in the mean if inequality (5.3) holds. Joint monitoring for a positive and a negative shift in the mean is done by combining the rules above. The probability of falsely rejecting H_0 is then approximately 2α. Let us illustrate the combined test graphically, see Figure 5.5.

In Figure 5.5, it is shown that the boundary of the continuation area forms a reversed V-mask. The lines $-r\delta/2 + \log(\alpha)/\delta$ and $r\delta/2 - \log(\alpha)/\delta$ intersect at $r = 2\log(\alpha)/\delta^2$, so that for the parameters d and θ we have

$$\begin{aligned} d &= -2\log(\alpha)/\delta^2 \\ \tan\theta &= \tfrac{1}{2}\delta. \end{aligned}$$

The V-mask monitoring scheme can be designed using the above formulas. By specifying δ, the chart is designed for shifts of size δ. To complete the design, 2α, the probability of falsely rejecting the hypothesis that the mean is on target, needs to be specified. In Adams, Lowry and Woodall (1992), it was pointed out that defining 2α as the probability of a false alarm is incorrect. Essentially, 2α cannot be the probability of a false alarm on any single sample, because this probability changes over time on the CUSUM.

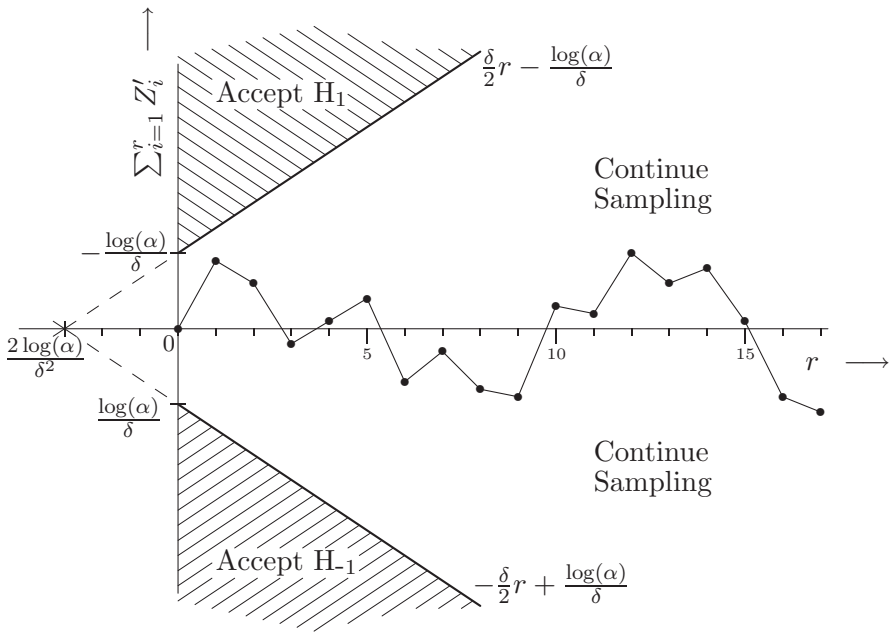


Figure 5.5: SPRT for joint testing for a negative and a positive shift in the mean.

Nor can 2α be the probability of eventually obtaining a false alarm (this probability is, of course, 1). In fact, 2α must be the long-run proportion of observations resulting in false alarms. If this is the case, then the in-control ARL should be $1/(2\alpha)$. However, Johnson's design method produces in-control ARLs that are substantially larger than $1/(2\alpha)$.

Johnson's analogy between the CUSUM and the SPRT can also be employed to design the DIS. Rewrite the decision rules in formulas (5.2) and (5.3) as

$$\begin{aligned} &\text{accept } \mathrm{H}_{-1} \text{ if} \quad \sum_{i=1}^{r}\left(-Z_i' - \frac{\delta}{2}\right) > -\frac{\log(\alpha)}{\delta} \\ &\text{accept } \mathrm{H}_{1} \text{ if} \quad \sum_{i=1}^{r}\left(Z_i' - \frac{\delta}{2}\right) > -\frac{\log(\alpha)}{\delta} \qquad (5.4) \\ &\text{otherwise, continue sampling.} \end{aligned}$$

This suggests choosing

$$k = \frac{\delta}{2} \quad \text{and} \quad h = \frac{\log(\alpha)}{\delta}.$$

Note that this value of k agrees with the general recommendation to set k equal to half the size of the shift one wants to detect.

The equivalence of the decision rules in (5.4) and the decision to generate an out-of-control signal the first time one of the sums S_{L_i} or S_{H_i} of Equation (5.1) is larger than h, is easily established by noting that in the first two rows of (5.4), two cumulative sums of the last r unreversed observations of a sample of size m are compared to an upper limit. If it is not decided to accept one of the alternative hypotheses, the two cumulative sums of the last $r+1$ unreversed observations are considered, and so on, until the two cumulative sums of all m previous observations have been compared to the upper limit. If neither H_{-1} or H_1 is accepted, a new observation is taken, and for each of the first two rows in (5.4), $m+1$ cumulative sums are compared to the upper limit. If it is decided to accept H_{-1} at time $m+1$ for the first time, the largest value of the $m+1$ sums of the first row of (5.4) is exactly $S_{L_{m+1}}$. Analogously, if it is decided to accept H_1 at time $m+1$ for the first time, the largest value of the $m+1$ sums of the second row in (5.4) is exactly $S_{H_{m+1}}$.

When applying the SPRT to distinguish between two hypotheses H_0 and H_1, the number of observations that need to be taken before a decision is reached, is a random variable, denoted by N. It is a well-known remarkable result (Wald's theorem) that for fixed α and β, the SPRT minimizes *both* $E(N|H_0)$ *and* $E(N|H_1)$.

5.1.2 The ARL of the CUSUM for i.i.d observations

For determining the ARL of the CUSUM for the mean of independent observations, two approaches can be used. The Markov-chain approach is discussed in Appendix C. Here, we will utilize the Fredholm-integral approach to derive the ARL of the CUSUM control chart for the mean of independently distributed observations, when a DIS is used to judge the CUSUM.

The derivation of this integral equation is a little more complicated than the integral equation of the EWMA control chart that was discussed in Section 4.1. The reason for this is that the sums S_L and S_H may consist of several so-called *runs* before an out-of-control signal occurs. A *run* of S_L is defined as follows. A *run* starts when S_L has been zero, and ends as soon as S_L becomes zero again or if an out-of-control signal occurs ($S_L \geq h$ or if $S_H \geq h$). Conversely, a run of S_H starts the next observation after the sum has been zero, and ends as soon as $S_H = 0$ or if $S_L \geq h$ or if $S_H \geq h$.

As an illustration, we have 11 runs for S_L in Figure 5.4, they are observations $\{1\}$, $\{2\}$, $\{3,4,5\}$, $\{6,\cdots,9\}$, $\{10\}$, $\{11\}$, $\{12\}$, $\{13\}$, $\{14,\cdots,19\}$, $\{20\}$, and $\{21\}$. For S_H we have four runs in Figure 5.4. They are observations $\{1,2,3\}$, $\{4,\cdots,7\}$, $\{8,\cdots,16\}$, and $\{17,\cdots,21\}$.

In Kemp (1961), it was proved that if an out-of-control signal occurs for the first time at time T indicating a negative shift in the mean, i.e. $S_{L_T} > h$ while $0 \leq S_{L_i} \leq h$ and $0 \leq S_{H_i} < h$ for $i = 1,\cdots,T-1$, then $S_{H_T} = 0$. Conversely, if $S_{H_T} > h$ for some T while $0 \leq S_{L_i} \leq h$ and $0 \leq S_{H_i} \leq h$ for $i = 1,\cdots,T-1$, then $S_{L_T} = 0$. This means that a run which ends at or above h cannot cut short a run of the other sum. Moreover, if we let L^- denote the ARL of a one-sided DIS monitoring S_L, then m/L^-, the expected number of out-of-control signals occurring in S_L after m observations, is not affected by simultaneous monitoring of S_H. Vice versa, if L^+ denotes the ARL of a one-sided DIS monitoring S_H, then m/L^+, the expected number of out-of-control signals after m observations, is not affected by simultaneous monitoring of S_L. As a result, the expected number of out-of-control signals of a two-sided DIS is then

$$\frac{m}{L^-} + \frac{m}{L^+},$$

so that

$$\frac{1}{L} = \frac{1}{L^-} + \frac{1}{L^+}, \tag{5.5}$$

where L denotes the ARL of the two-sided DIS.

In some cases, it may be desirable to set k^- and h^-, the reference value and the decision interval of S_L, respectively, different from k^+ and h^+, the reference value and the decision interval of S_H, respectively. Such a DIS is called an *asymmetrical two-sided DIS*. In van Dobben de Bruyn (1968) it was shown that formula (5.5) also holds for an asymmetrical two-sided DIS, provided that

$$k^+ + k^- \geq |h^+ - h^-|.$$

Note that this condition is always satisfied for a symmetric two-sided DIS, i.e. if $h^+ = h^-$ and $k^- = k^+$. Hence, formula (5.5) shows how the ARL curve of a symmetric two-sided DIS can be computed from the ARLs of two one-sided schemes.

Let us therefore consider the ARL of a one-sided DIS for detecting positive shifts in the mean of a sequence of standardized identically and independently normally distributed random variables. Page (1954) defined the following three functions

$P(s)$: the probability that a run, starting at s, ends at 0

$N(s)$: the expected length of a run, starting in s

$L(s)$: the expected run length of the one-sided DIS, which started in s.

Furthermore, let $f(\cdot)$ and $F(\cdot)$ denote the density, and the cumulative distribution function of Z_t, respectively. We can write down the following integral equation for $P(s)$

$$P(s) = F(k-s) + \int_{k-s}^{h+k-s} P(s+z-k) f(z)\,\mathrm{d}z \tag{5.6}$$

$$= F(k-s) + \int_0^h P(u) f(u-s+k)\,\mathrm{d}u, \tag{5.7}$$

where $u = s+z-k$. The first term on the right-hand side of Equation (5.6) corresponds to the case where the current outcome z causes the run to end at zero. The second term corresponds to the case where the value of the sum, after the current observation z is taken, is between 0 and h, so the run continues from $s + z - k$. Analogously, we can derive

$$\begin{aligned} N(s) &= 1 + \int_{k-s}^{h+k-s} N(s+z-k) f(z)\,\mathrm{d}z \\ &= 1 + \int_0^h N(u) f(u-s+k)\,\mathrm{d}u, \end{aligned} \tag{5.8}$$

and

$$\begin{aligned} L(s) &= 1 + L(0)F(k-s) + \int_{k-s}^{h+k-s} L(s+z-k) f(z)\,\mathrm{d}z \\ &= 1 + L(0)F(k-s) + \int_0^h L(u) f(u-s+k)\,\mathrm{d}u. \end{aligned} \tag{5.9}$$

In order to evaluate $L(s)$, $L(0)$ needs to be known. Suppose that $P(0)$ and $N(0)$ are obtained by solving (5.7) and (5.8). The number of runs before the first out-of-control signal occurs follows a geometric distribution

$$\{P(0)\}^{n-1}\{1-P(0)\} \quad \text{for } n = 1, 2, \cdots,$$

with expected value

$$\frac{1}{1-P(0)}.$$

If $N(0)^u$ is the expected number of observations in a run that ends at h, and $N(0)^l$ is the expected number of observations in a run that ends at 0, then we have

$$\begin{aligned} L(0) &= N(0)^u + \left[\frac{1}{1-P(0)} - 1\right] N(0)^l \\ &= \frac{1}{1-P(0)}\left[(1-P(0))N(0)^u + P(0)N(0)^l\right] \\ &= \frac{N(0)}{1-P(0)}. \end{aligned} \tag{5.10}$$

After substituting expression (5.10) into the integral equation (5.9), the latter can be numerically approximated using Gaussian Quadrature. Details can be found Appendix B. Integral equations for the one-sided DIS for detecting negative shifts in the mean can be derived analogously. Formula (5.5) can then be used to combine two one-sided ARLs to the ARL for a two-sided DIS.

In Figure 5.6, the ARL curve of a two-sided DIS is depicted. The reference value was set to $k = 0.5$, and the decision interval was chosen in such a way that the in control ARL equaled 370.4. This resulted in $h = 4.7749$. The resulting ARL curve is drawn together with the ARL curve of a Shewhart control chart for independent observations and the ARL of an EWMA chart for independent observations.

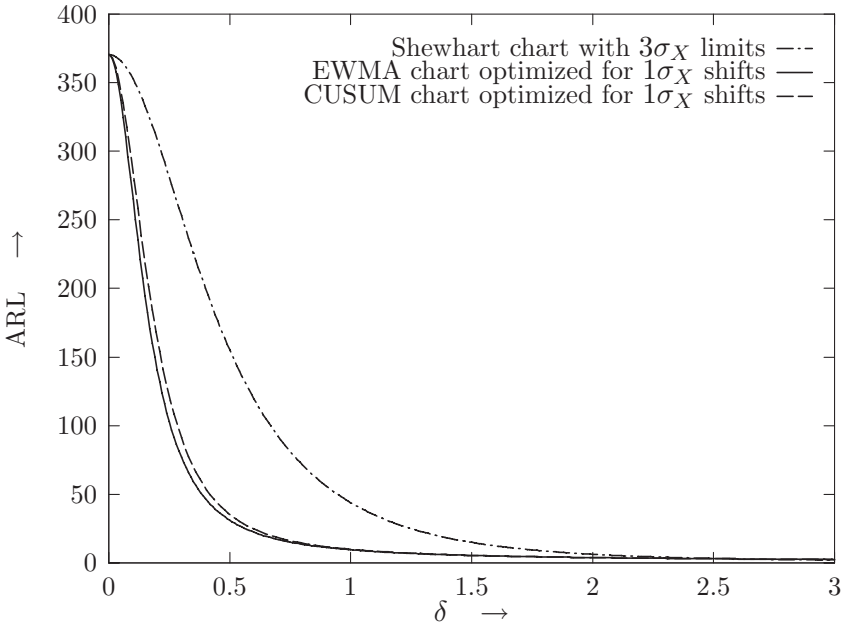


Figure 5.6: ARL curves of the Shewhart, the EWMA, and the CUSUM chart for the mean of independent observations.

Figure 5.6 shows that the ARL of the EWMA chart is smaller than the ARL of the CUSUM chart when a shift in μ occurred that is smaller than the shift that was used for the design of the EWMA chart. For larger shifts, the CUSUM ARL is slightly smaller. However, the difference is so small that this is not visible in Figure 5.6. This difference in ARL behavior is typical for a wide range of parameter values according to Lucas and Saccucci (1990).

Compared to the ARL curve of the Shewhart chart, both the CUSUM and the EWMA chart are more efficient in detecting small shifts in the mean. For larger shifts (say $\delta > 2.6$), the Shewhart control chart is slightly more efficient. For example, when $\delta = 3$, the ARL of the Shewhart chart is 2.00, the ARL of the EWMA chart is 2.51, and the ARL of the CUSUM chart is 2.49.

As in the previous two chapters, we want to investigate the effect of abandoning the assumption that the observations are independently distributed on the ARL curve of the CUSUM chart. Again, we will assume that the control chart is used to detect a shift in the mean of a sequence of observations $\{Y_t\}$, which are known to be generated by the AR(1) model (2.3) (see Subsection 2.4.2):

$$Y_t - \mu_t = \phi(Y_{t-1} - \mu_{t-1}) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}.$$

In the following section, it is investigated how the performance of a CUSUM chart that was designed for independent observations is affected by AR(1) dependence in the data.

5.2 Effect of ignoring serial correlation

In this section, we will simulate how the ARL of a CUSUM chart that was designed to detect a shift in the mean of independent data is affected by AR(1) dependence in the observations. To this end, we consider the DIS of the previous section that was designed to detect a shift of size $1\sigma_X$ as soon as possible. Using parameters $k = 0.5$ and $h = 4.7749$, we will investigate how AR(1) correlation affects the in-control ARL and the out-of-control ARL when there has been a shift of size $1\sigma_Y$ in the mean of the observations. The simulation results are presented in Table 5.1.

As in Sections 3.3 and 4.2, the ARL values are simulated assuming that the standard deviation of the observations is estimated using $\overline{MR}/d_2(2)$ which is biased for AR(1) observations. This results in control limits that are too wide in case of negative correlation. It was not feasible to simulate the in-control ARL for the cases $\phi - 0.6$ and $\phi = -0.9$. However, the simulation results for $\phi = -0.3$ (based on 100 replications) show that the in-control ARL is much higher than intended. The in-control ARL is much lower in the case of positive autocorrelation. The results show that the CUSUM chart is even more affected by serial correlation than the EWMA chart or the Shewhart chart. In the remainder of this chapter, we will

Table 5.1: ARL values of an i.i.d. CUSUM chart, applied to various AR(1) processes.

ϕ	σ_Y	$\mathrm{E}\left[\overline{MR}/d_2(2)\right]$	ARL($0\,\sigma_Y$)	ARL($1\,\sigma_Y$)
-0.9	2.2942	3.1623	—.—	—.—
-0.6	1.2500	1.5811	—.—	—.—
-0.3	1.0483	1.1952	529076.00 (55184.77)	15.77 (0.03)
0.0	1.0000	1.0000	368.09 (1.16)	9.14 (0.02)
0.3	1.0483	0.8771	21.98 (0.09)	5.15 (0.01)
0.6	1.2500	0.7906	2.51 (0.01)	1.73 (0.01)
0.9	2.2942	0.7255	1.00 (0.00)	1.00 (0.00)

discuss control charts that will be designed to detect shifts in the mean of AR(1) data.

In Section 5.3, a modified CUSUM chart is discussed that accounts for serial correlation. The sums S_L and S_H of serially correlated observations are assessed using a DIS with modified limits to account for the first-order autoregressive serial correlation. In Section 5.4, the sequence of observations $\{Y_t\}$ is monitored for a shift in the mean using a DIS for the cumulative sums of residuals. In Section 5.5, the CUSUM for modified residuals is discussed. This chapter is concluded by a comparison of the ARL behavior of these three charts in Section 5.6.

5.3 The modified CUSUM chart

In this section, we will consider the modified CUSUM chart. Analogous to the modified Shewhart chart and the modified EWMA chart, sums of untransformed AR(1) observations $\{Y_t\}$ will be assessed in the DIS of the modified CUSUM. The difference with the CUSUM for independent observations is that the value of the decision interval h is adjusted to account for the serial correlation that is present in the data. The value of h is increased

to account for the increase in the variance of the observations that are the basis for the cumulative sums. However, in Schmid (1997a), it was proved that the in-control ARL thus obtained is larger than the in-control ARL of a corresponding CUSUM for independent observations. Hence, the increase in h must be partly undone to obtain a desired in-control ARL.

The sums for the DIS of the modified CUSUM chart will be denoted by $S_{L_{Y,t}}$ and $S_{H_{Y,t}}$, and are computed as follows

$$\begin{aligned} S_{L_{Y,t}} &= \max(S_{L_{Y,t-1}} - (Y_t - \mu) - k, 0) \\ S_{H_{Y,t}} &= \max(S_{H_{Y,t-1}} + (Y_t - \mu) - k, 0). \end{aligned}$$

An out-of-control signal is issued at time t if $S_{L_{Y,t}} > h$ or if $S_{H_{Y,t}} > h$. For the design of the CUSUM DIS, we apply the rule of thumb that the reference value k is chosen equal to half the size of the shift we want to detect as fast as possible. The corresponding decision interval h is chosen such that the chart has a desired in-control ARL. For an in-control ARL of 500, h and k combinations can be found in Schmid (1997a) for selected values of ϕ. In addition, Schmid (1997a) questions whether this is an appropriate way to design the CUSUM DIS. A more correct way is to minimize the ARL for a shift of a predetermined size over all k, where h is chosen with each k to obtain a certain in-control ARL.

For the computation of the ARL of the modified CUSUM chart, we have several options. Goldsmith and Whitfield (1961), who were the first to study the effect of serial correlation on the performance of CUSUM control charts, used Monte Carlo simulations. Johnson and Bagshaw (1974) employ the theory of weak convergence of the partial sums to a Wiener process to obtain the asymptotic distribution of the run length of the one sided CUSUM when the process is in control. In a second article, Bagshaw and Johnson (1975), employed a Wiener process with drift to obtain an approximation for the distribution of the run length of a one-sided CUSUM when a shift in the mean has occurred. The latter approximation proves to be accurate for the shape of the run length. Also, this approximation is accurate for the location of the run length when negative correlation is present in the data. However, when there is positive correlation, the approximation underestimates the ARL by a considerable amount. Since we are especially interested in the ARL curves when there is positive correlation in the data, we will not use this approximation.

Another option for the approximation of the ARL of the modified CUSUM is discussed by Yashchin (1993). In this article, the ARL proper-

ties in case of serially correlated observations are evaluated by specifying a matching i.i.d. process, for which the CUSUM run length distribution is approximately the same as for the original series. The properties of the run length distribution can then be assessed by employing the Fredholm-integral approach, see the previous section, or the Markov-chain approach, see Appendix C. Yashchin discusses several procedures for choosing the matching i.i.d. process. Yashchin's approximations prove to be accurate in many, but not all cases. When the level of autocorrelation is too large, the quality of the approximations deteriorates.

Yet another option is to adapt the Fredholm-integral approach that we used in the previous section to the case of AR(1) observations, analogous to Chapter 4, where we discussed the ARL of the modified EWMA control chart. VanBrackle and Reynolds (1997) used a similar method to study the ARL of the CUSUM in case of ARMA(1,1) observations. By choosing a large enough number of quadrature points this method produces very accurate results, at the cost, however, of exponentially increasing computing time.

We decided to simulate the ARL curve of the modified CUSUM chart, since this method combines high accuracy in all cases with reasonable computing time. The ARL curves of the modified CUSUM chart are depicted in Figures 5.7 through 5.12 of Section 5.6, together with the ARL curves of the two CUSUM-type control charts that are discussed in the following sections.

5.4 The CUSUM chart of residuals

In this section, we will discuss monitoring an AR(1) process for a shift in the mean with a CUSUM chart of residuals. The sums S_L and S_H are formed using the residuals

$$e_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu).$$

The sums for the DIS of the residuals CUSUM chart will be denoted by $S_{L_{e,t}}$ and $S_{H_{e,t}}$, and are computed as follows

$$\begin{aligned} S_{L_{e,t}} &= \max(S_{L_{e,t-1}} - e_t - k, 0) \\ S_{H_{e,t}} &= \max(S_{H_{e,t-1}} + e_t - k, 0). \end{aligned}$$

An out-of-control signal is issued at time t if $S_{L_{e,t}} > h$ or if $S_{H_{e,t}} > h$.

For the design of the modified CUSUM control chart, we can repeat Johnson's approximate analogy between the SPRT and the CUSUM of the previous section for AR(1) observations. Let us consider a sample of m AR(1) observations. They are jointly normally distributed with covariance matrix

$$\boldsymbol{\Sigma} = \frac{\sigma_\varepsilon^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{m-1} \\ \phi & 1 & \phi & \cdots & \phi^{m-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{m-1} & \phi^{m-2} & \phi^{m-3} & \cdots & 1 \end{pmatrix}.$$

For the mean of the process, we formulate three hypotheses:

$$\mathrm{H}_{-1} : \boldsymbol{\mu} = \boldsymbol{\mu}_{-1} \qquad \mathrm{H}_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \qquad \mathrm{H}_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_1,$$

where $\boldsymbol{\mu}_{-1} = (\mu_0 - \delta\sigma_Y)\boldsymbol{\iota}$, $\boldsymbol{\mu}_0 = \mu_0\boldsymbol{\iota}$, and $\boldsymbol{\mu}_1 = (\mu_0 + \delta\sigma_Y)\boldsymbol{\iota}$, with $\boldsymbol{\iota}$ an $(m \times 1)$-vector with all components equal to one. For the likelihood ratio that can be used to discriminate between H_0 and H_{-1}, we have after m observations

$$\lambda_m = \frac{\dfrac{1}{(2\pi)^{m/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\dfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_0)\right)}{\dfrac{1}{(2\pi)^{m/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\dfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{-1})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{-1})\right)},$$

so that the log likelihood ratio after m observations equals

$$\log \lambda_m = \delta\sigma_Y \boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_0) + \frac{(\delta\sigma_Y)^2}{2}\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\iota}.$$

It can be shown that

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_\varepsilon^2} \begin{pmatrix} 1 & -\phi & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\phi & 1+\phi^2 & -\phi & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\phi & 1+\phi^2 & -\phi & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -\phi & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & -\phi & 1 \end{pmatrix},$$

so that

$$\begin{aligned}\log \lambda_m &= \frac{\delta\sigma_Y}{\sigma_\varepsilon^2}\left((y_1-\mu_0)-\phi(y_2-\mu_0)+e_2+(1-\phi)\sum_{i=3}^{m}e_i\right)\\ &\quad+\frac{(\delta\sigma_Y)^2}{2\sigma_\varepsilon^2}\left(2(1-\phi)+(m-2)(1-\phi)^2\right)\\ &= \frac{\delta\sigma_Y(1-\phi)}{\sigma_\varepsilon^2}\sum_{i=3}^{m}\left(e_i+\frac{\delta(1-\phi)}{2}\right)\\ &\quad+\frac{\delta\sigma_Y}{\sigma_\varepsilon^2}\left[(y_1-\mu_0)-\phi(y_2-\mu_0)+e_2+\delta(1-\phi)\right]. \quad (5.11)\end{aligned}$$

Hence, formula (5.11) shows that, apart from the first two terms, the SPRT of AR(1) observations is equivalent to monitoring a sum of residuals

$$\sum_i\left(e_i+\frac{\delta\sigma_Y(1-\phi)}{2}\right).$$

With respect to the design of the CUSUM residuals chart, this suggests choosing

$$k=\frac{\delta\sigma_Y(1-\phi)}{2}. \quad (5.12)$$

This is an attractive reference value, since we showed in Chapter 3 that only a fraction $(1-\phi)$ of a shift in the mean of AR(1) observations is transferred to the residuals. Hence, if a shift of size $\delta^*\sigma_Y$ has to be detected as fast as possible in a sequence of AR(1) observations, this can be done by monitoring the residuals of the one-step-ahead forecasts with a CUSUM chart that is designed to detect a shift of size $(1-\phi)\delta^*\sigma_Y$ as soon as possible.

Formula (5.11) is a special case of an expression that was derived by Schmid (1997a). He showed that the SPRT for the mean of the more general class of Gaussian processes is equivalent to monitoring a weighted cumulative sum of one-step-ahead prediction residuals, where the weights attached to the residuals depend on parameters of the process.

Again, if the AR(1) model is appropriate for the data, and enough data is available to assume that the model parameters are known, then the residuals will approximately be uncorrelated. In such cases, the Fredholm-integral approach of Section 5.3 can be used to approximate the ARL curve of the CUSUM chart of residuals numerically.

In Section 5.6, the ARL behavior of the CUSUM chart of residuals is compared to that of the modified CUSUM chart of Section 5.3, and of the CUSUM chart of modified residuals, which will be discussed in the next section. The CUSUM charts of residuals that are discussed in Section 5.6 are designed to detect a shift of size $1\sigma_Y$ in the mean of AR(1) observations as soon as possible. Using (5.12), the reference value of the CUSUM chart of residuals is set to

$$k = \frac{\delta}{2}\sqrt{\frac{1-\phi}{1+\phi}}\sigma_\varepsilon,$$

and the corresponding value of h is chosen with k to obtain a desired in-control ARL.

5.5 The CUSUM chart of modified residuals

In this section we consider the CUSUM chart of modified residuals. In case of a sequence of AR(1) observations $\{Y_t\}$, the modified residual at time $t = 1, 2, \cdots$, was in Section 3.6 defined as

$$u_t \equiv Y_t - \phi Y_{t-1} + \phi\hat{\mu}_t$$

where $\hat{\mu}_t$ is an estimator of μ_t, the mean of the process at time t. As discussed in Section 3.6, monitoring AR(1) data using $\{u_t\}$ has three advantages:

1. successive realizations of u_t are approximately uncorrelated;
2. the sequence $\{\mathrm{E}(u_t)\}$ is approximately equal to $\{\mu_t\}$;
3. step changes in the mean of $\{Y_t\}$ are eventually transferred fully to $\{u_t\}$.

The fact that elements of $\{u_t\}$ are approximately uncorrelated allows us to use control charts for independent observations, the behavior of which has been well-studied. The second advantage makes it easier to relate the modified residuals to the process than the unmodified one-step-ahead forecast residuals. The third advantage is the reason for the more efficient operation of the modified residuals chart, compared to the performance of the residuals charts.

The sums for the DIS of the modified residuals CUSUM chart will be denoted by $S_{L_{u,t}}$ and $S_{H_{u,t}}$, and are computed as follows

$$\begin{aligned} S_{L_{u,t}} &= \max(S_{L_{u,t-1}} - (u_t - \mu) - k, 0) \\ S_{H_{u,t}} &= \max(S_{H_{u,t-1}} + (u_t - \mu) - k, 0). \end{aligned}$$

An out-of-control signal is issued at time t if $S_{L_{u,t}} > h$ or if $S_{H_{u,t}} > h$.

For the design of the modified residuals chart we employ the rule of thumb to choose k equal to half the size of the shift we want to detect quickly. In the case of monitoring modified residuals, a step shift in the mean of the AR(1) observations is fully transferred to the modified residuals, so that we take

$$k = \frac{\delta\sigma_Y}{2}$$

as the reference value for the modified residuals CUSUM chart. The corresponding decision interval h is chosen so that a desired in-control ARL is obtained.

In the next section, the ARL performance of the modified residuals CUSUM chart is compared to the ARL behavior of the modified CUSUM chart and the residuals CUSUM chart for various levels of first-order autoregressive serial correlation. The ARL curve of the modified residuals CUSUM chart is obtained by Monte Carlo simulation.

5.6 Discussion

In this section, the CUSUM-type control charts for the mean of AR(1) data that were discussed in previous sections, are compared on the basis of their ARL behavior for selected values of ϕ. The ARL curves are depicted in Figures 5.7 through 5.12.

Goldsmith and Whitfield (1961) were the first to study the effect of serial correlation on the performance of CUSUM control charts. They studied the effect of AR(1) dependence on the ARL of a CUSUM that was assessed with a V-mask. They concluded that negative autocorrelation makes the CUSUM more sensitive, whereas positive correlation has a negative effect on the ARL performance of the CUSUM control chart. These conclusions agree with the findings of Figures 5.7 through 5.12, except for the modified CUSUM chart.

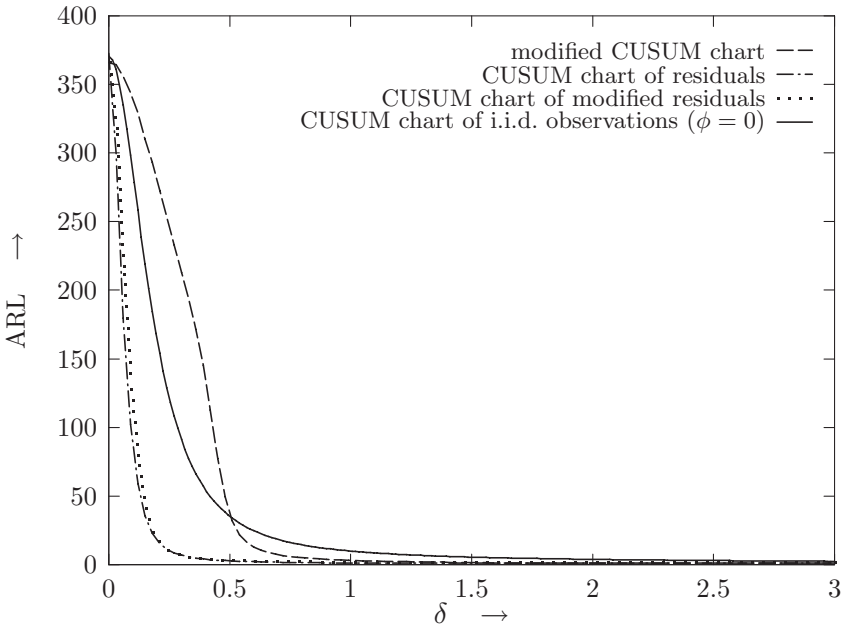


Figure 5.7: Various ARL curves for AR(1) process with $\phi = -0.9$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

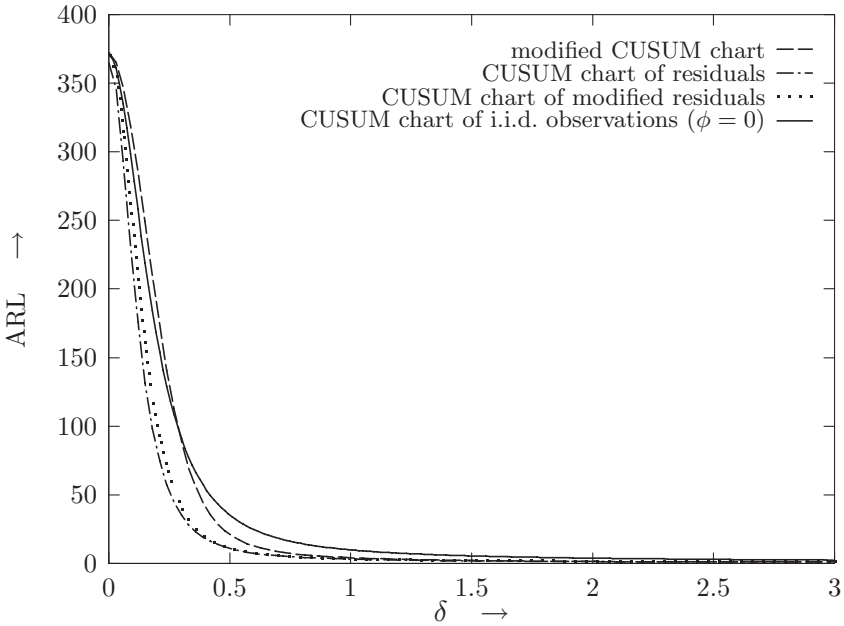


Figure 5.8: Various ARL curves for AR(1) process with $\phi = -0.6$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

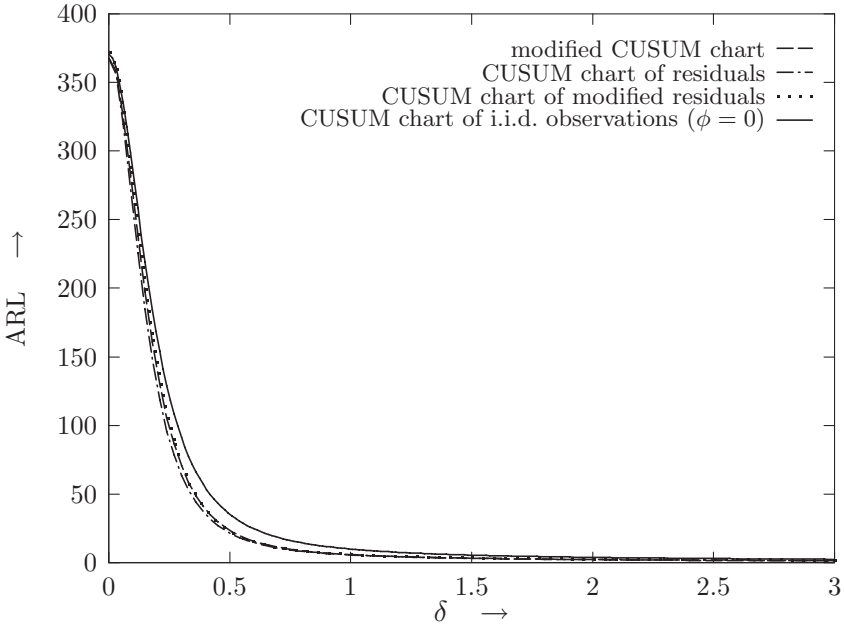


Figure 5.9: Various ARL curves for AR(1) process with $\phi = -0.3$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

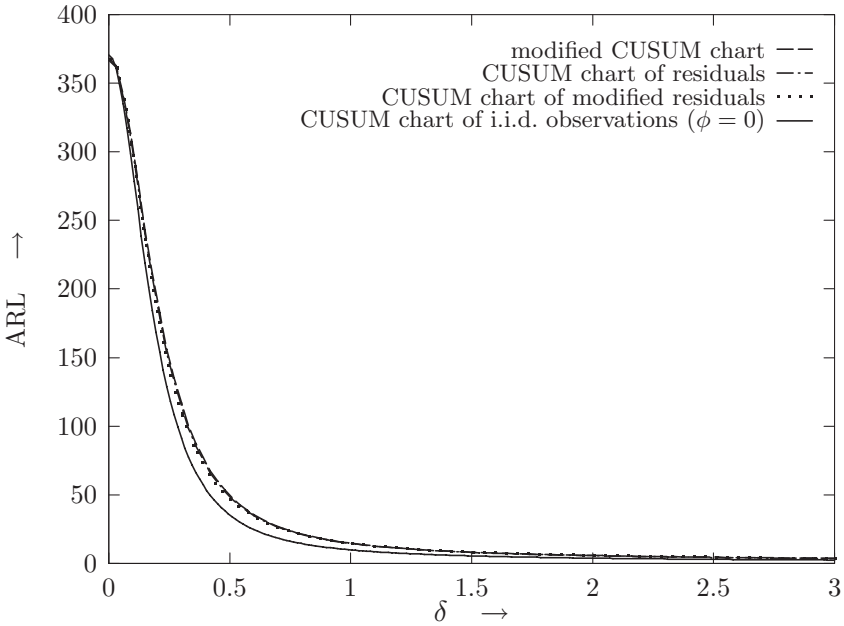


Figure 5.10: Various ARL curves for AR(1) process with $\phi = 0.3$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

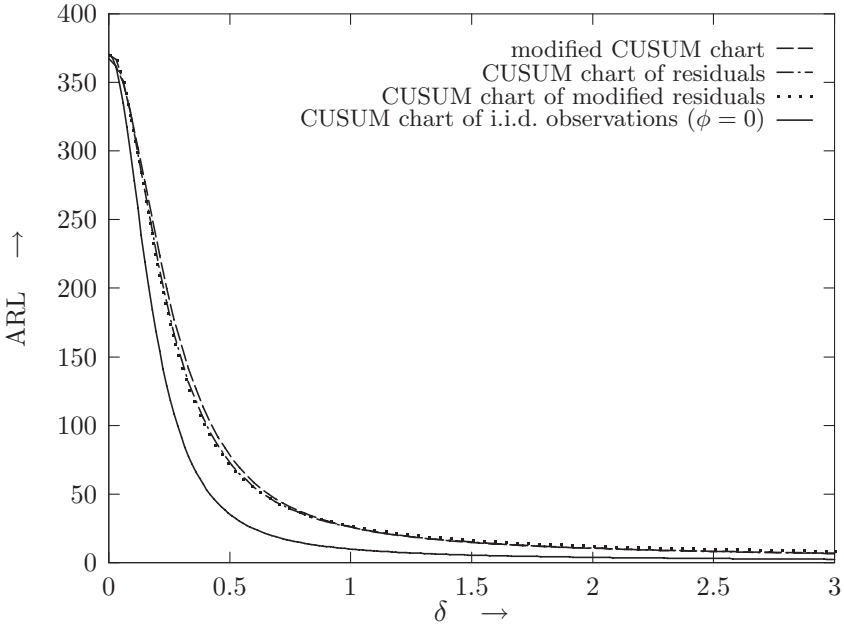


Figure 5.11: Various ARL curves for AR(1) process with $\phi = 0.6$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

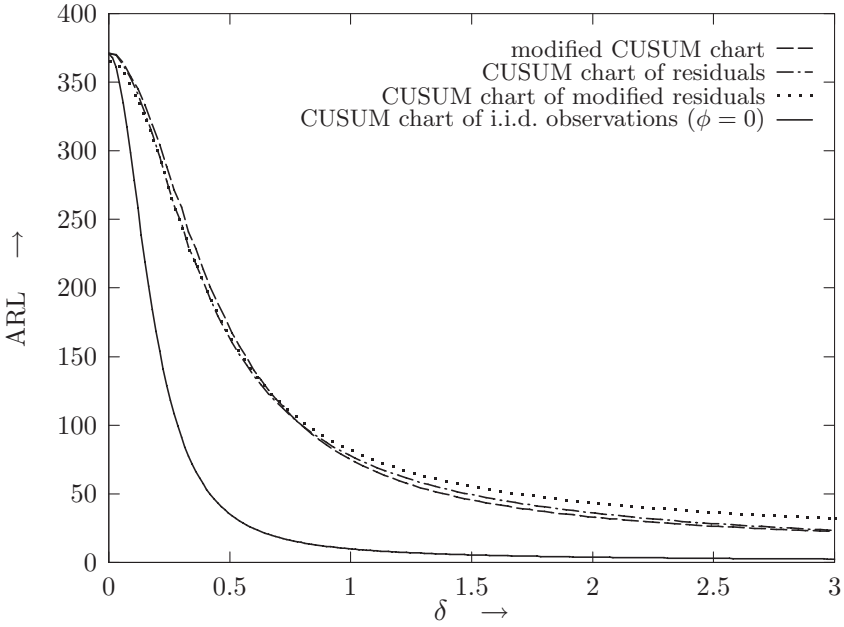


Figure 5.12: Various ARL curves for AR(1) process with $\phi = 0.9$, compared to the ARL curve for the i.i.d. case ($\phi = 0$).

The differences between the ARL curve of the CUSUM of residuals and the CUSUM of modified residuals are small. Both charts appear to perform well in all cases. In view of the slightly better results, the CUSUM chart of residuals might be preferred. For positive autocorrelation, the ARL curve of the modified CUSUM chart is very similar to that of the other two CUSUM charts. This is interesting, since we observed in the Chapters 3 and 4 that the modified Shewhart and the modified EWMA chart were performing better than the corresponding residuals charts. Hence, contrary to Shewhart- and EWMA-type control charts, there is no ARL argument in favor of the use of a modified CUSUM chart to monitor AR(1) data.

Figures 5.7 and 5.8 show that, in case of negative autocorrelation, it is also not advisable to use the modified CUSUM chart. These findings agree with the results presented in Schmid (1997a), wherein it is concluded that for negative autocorrelation, depending on the choice of h and k, the modified CUSUM is very inefficient in signalling smaller shifts than the chart was designed to detect. His results indicate that in these cases, better ARL behavior is obtained if k is not chosen equal to half the size of a shift one wants to detect. Therefore, the usual rules of thumb for the design of CUSUM-type control charts cannot be used in all cases of serially correlated data. It is our experience that these difficulties concerning the design of the CUSUM chart tend to increase with the complexity of the underlying time series model. Therefore, we agree with Schmid (1997a) that further research concerning the design and the ARL behavior of CUSUM-type control charts for other models than the AR(1) model is needed. For this reason, we decided not to tabulate ARL values of CUSUM-type control charts in Appendix A.

In addition to the foregoing argument, it may be noted that for the cases where the usual rules of thumb seem to work satisfactory (Figures 5.7 through 5.12), the ARL behavior of the CUSUM-type control charts is very similar to the ARL behavior of the EWMA-type control charts of the previous chapter. These findings agree with Schmid (1997a). Hence, it is to be expected that the ARL behavior of the EWMA control charts provides also a good indication of the ARL behavior of the CUSUM control chart in case of other time series models, a suggestion that is supported by the results of VanBrackle and Reynolds (1997).

For positive autocorrelation, the CUSUM charts appear to detect changes in the mean which are smaller than $1\sigma_Y$ slightly earlier than the EWMA charts. However, changes larger than $1\sigma_Y$ are detected faster, on average, by EWMA charts.

Chapter 6

Two worked out examples

In this chapter, we illustrate the use of control charts that were discussed in the previous chapters by two examples. The first is a real life example, based on a data set that appeared in Shewhart (1931). In his treatment of the data set, Shewhart did not take the presence of serial correlation into account. By using control charts of the previous chapters, we arrive at other conclusions than Shewhart did. The second example is a based on a simulated AR(1) series with a persistent change in the mean of the observations.

6.1 A real life example

The first book on quality control stems from the year 1931. It was written by the developer of the control chart: Dr. Walter A. Shewhart. In this very well written work, the newly developed concepts of quality control are illustrated with real-life examples. The second data set that appears in this book consists of 204 observations of the electrical resistance of a certain insulation material. In Table 6.1, these observations are reprinted in chronological order.

Shewhart decides to take subgroups of size four, and presents a control chart for the mean. The 51 subgroup averages are compared to control limits 'within which experience has shown that these observations should fall'. Since this data set is presented in one of the first chapters, Shewhart does not explain precisely how the control limits are computed. However, on page 296, Shewhart suggests computing the control limits as

$$\bar{\bar{X}} \pm 3\frac{\hat{\sigma}}{\sqrt{n}},$$

which is a formula that is very familiar to most SPC practitioners. The central line is the overall mean, which can be computed as 4,498MΩ. Estimating σ as the mean of the sample standard deviations of the subgroups, corrected by the well-known constant $c_4(4)$ to remove the bias (see for example Montgomery (1996), Table VI), results in a lower control limit of 4.006MΩ, and an upper control limit of 4,991MΩ. These values agree closely with the control limits that Shewhart depicted in the corresponding control chart. Eight of the subgroup averages fall outside the control limits, see Figure 6.2(a). Shewhart interprets these out-of-control signals as 'an indication of the existence of causes of variability which could be found and eliminated'. He reports that further research was instituted to find these causes of variability. The search was successful and a second control chart is presented, in which data points are depicted that were taken after elimination of these causes. All values remain within much tighter limits, and Shewhart concludes that 'this variation should be left to chance'.

Table 6.1: Electrical resistance of insulation in megohms

5,045	4,635	4,700	4,650	4,640	3,940	4,570	4,560	4,450	4,500	5,075	4,500
4,350	5,100	4,600	4,170	4,335	3,700	4,570	3,075	4,450	4,770	4,925	4,850
4,350	5,450	4,110	4,255	5,000	3,650	4,855	2,965	4,850	5,150	5,075	4,930
3,975	4,635	4,410	4,170	4,615	4,445	4,160	4,080	4,450	4,850	4,925	4,700
4,290	4,720	4,180	4,375	4,215	4,000	4,325	4,080	3,635	4,700	5,250	4,890
4,430	4,810	4,790	4,175	4,275	4,845	4,125	4,425	3,635	5,000	4,915	4,625
4,485	4,565	4,790	4,550	4,275	5,000	4,100	4,300	3,635	5,000	5,600	4,425
4,285	4,410	4,340	4,450	5,000	4,560	4,340	4,430	3,900	5,000	5,075	4,135
3,980	4,065	4,895	2,855	4,615	4,700	4,575	4,840	4,340	4,700	4,450	4,190
3,925	4,565	5,750	2,920	4,735	4,310	3,875	4,840	4,340	4,500	4,215	4,080
3,645	5,190	4,740	4,375	4,215	4,310	4,050	4,310	3,665	4,840	4,325	3,690
3,760	4,725	5,000	4,375	4,700	5,000	4,050	4,185	3,775	5,075	4,665	5,050
3,300	4,640	4,895	4,355	4,700	4,575	4,685	4,570	5,000	5,000	4,615	4,625
3,685	4,640	4,255	4,090	4,700	4,700	4,685	4,700	4,850	4,770	4,615	5,150
3,463	4,895	4,170	5,000	4,700	4,430	4,430	4,440	4,775	4,570	4,500	5,250
5,200	4,790	3,850	4,335	4,095	4,850	4,300	4,850	4,500	4,925	4,765	5,000
5,100	4,845	4,445	5,000	4,095	4,850	4,690	4,125	4,770	4,775	4,500	5,000

Source: Shewhart (1931), page 20, Table 2.

However, if we take a closer look at the data set in Table 6.1, it appears that the successive values exhibit serial correlation. In fact, the data set appears to be a typical example of observations that can be successfully

modelled using an AR(1) model. The sample autocorrelation function is exponentially declining, and the sample partial autocorrelation function shows a single spike at lag 1, see Figure 6.1.

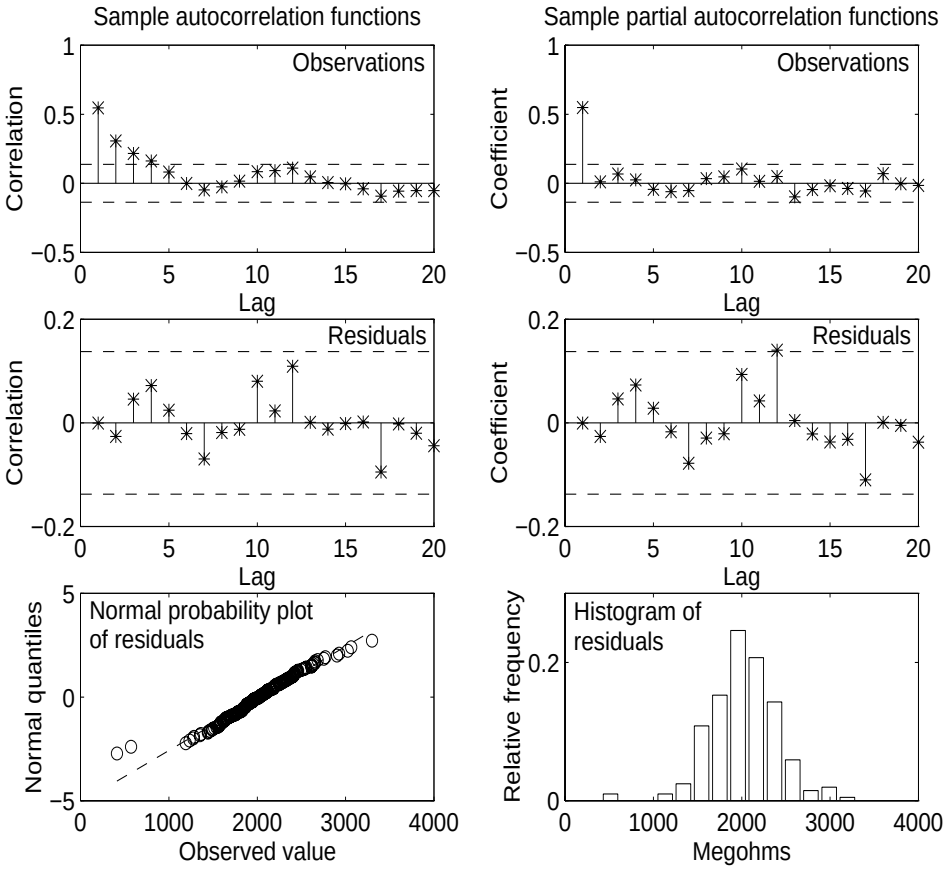


Figure 6.1: Analysis of the data in Table 6.1.

The AR parameter ϕ can be estimated as $\hat{\phi} = 0.549$. In Figure 6.1 it is shown that the residuals of this model show no significant serial correlation. A normal probability plot of the residuals indicates the presence of two or three outliers. All other observations lie more or less on a straight line, even after removal of the suspected data points. The histogram of the residuals is a little skewed to the right. Nevertheless, in our opinion, it is reasonable to assume that the residuals of this model are uncorrelated and normally distributed.

The autocorrelation function of the 51 subgroup averages shows less

convincingly that serial correlation is present. Here we observe a well-known phenomenon (see also Section 7.3): by taking subgroup averages, the serial correlation is reduced. Ignoring the serial correlation in the data seems therefore justifiable. However, Wardell, Moskowitz, and Plante (1992) warn against this kind of taking of subgroups:

> "*However, if the data are truly autocorrelated, the points on the Shewhart chart will still show runs which are essentially due to correlation resulting from common causes rather than any special cause*".

This statement is in agreement with Figure 6.2, where five control charts corresponding to the data in Table 6.1 are depicted. Figure 6.2(a) shows a control chart of subgroup averages, as proposed by Shewhart (1931). In Figures 6.2(b)–(d), a modified Shewhart control chart, a Shewhart chart of residuals, and a Shewhart chart of modified residuals are depicted, respectively. In Figure 6.2(e), an EWMA chart of modified residuals is presented. We have omitted other EWMA charts and CUSUM-type control charts, since their results turned out to be very similar to the results of the EWMA chart of modified residuals.

The control limits in Figures 6.2(b)–(e) are adjusted so as to have an approximate in-control ARL of 370.4. This was presumably Shewhart's intention.

When comparing the width of the control limits of Figures 6.2(a)–(d) (the four Shewhart-type control charts), we observe that the control limits of Figures 6.2(b)–(d) are much wider than the control limits of Figure 6.2(a). This is to be expected, since the control chart of Figure 6.2(a) is constructed for monitoring subgroup averages, whereas the control charts in Figures 6.2(b)–(d) monitor statistics of individual observations.

However, the estimator of σ_Y that is used for Figure 6.2(a) is biased in case of small subsamples of serially correlated data, so that the width of the control limits is not correctly determined. The problem encountered here is similar to the problems of Sections 3.3, 4.2 and 5.2, concerning the bias in $\overline{MR}/d_2(2)$ as an estimator of σ_Y.

The control limits in Figure 6.2(a) are computed using the mean of the subgroup standard deviations divided by $c_4(4)$ as an estimator for σ_Y. In Section 7.5 of the next chapter, it is established that in case of serially correlated data, S^2 is a biased estimator for σ_Y^2. Using these results, it can be shown that the mean of the subgroup standard deviations underestimates σ_Y in case of positive autocorrelation. This bias disappears for large subgroup sizes. In the present case, where we have $n = 4$ and $\hat{\phi} = 0.549$, σ_Y is

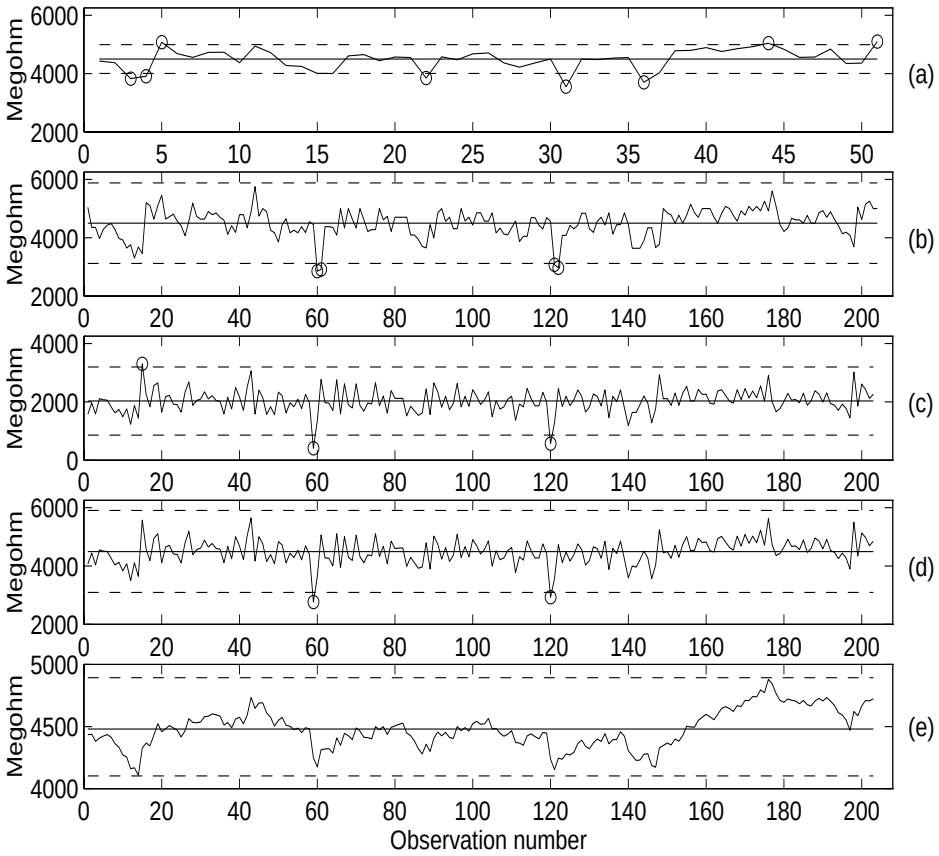


Figure 6.2: Five control charts corresponding to Table 6.1.
(a)=Chart for subgroup averages,
(b)=modified Shewhart,
(c)=Shewhart chart of residuals,
(d)=Shewhart chart of modified residuals,
(e)=EWMA chart of modified residuals.

underestimated, which results in control limits that are too tight. Consequently, too many false out-of-control signals are generated in the control chart for subgroup averages of Figure 6.2(a).

The Shewhart-type control charts of Figures 6.2(b)–(d) all show fewer out-of-control signals. Around observations 60 and 121, the presence of a special cause of variation may be suspected. Note that the modified Shewhart chart generates two signals around both of these observation numbers, whereas the Shewhart chart of residuals and the Shewhart chart of modified residuals produce only one out-of-control signal at both of these points. This illustrates the fact that the probability of an out-of-control signal on residuals-based control charts is high at the first observation following a shift in the mean, whereas this probability is much smaller at later observations, see also Section 3.5. Therefore, it may be wise to add observations 61 and 122 to the list of possible out-of-control situations. Based on an out-of-control signal on the Shewhart chart of residuals (Figure 6.2(c)), we have reason to suspect observation 16, too.

The EWMA control chart of modified residuals of Figure 6.2(e) does not produce any out-of-control signals, whereas the Shewhart-type control charts in Figures 6.2(b)–6.2(d) all signal shifts in the mean at the same observations. However, these out-of-control signals seem to relate to 'spike'-shifts in the mean. Shewhart-type control charts are to be preferred for this kind of shifts, since on these charts only the current observation is monitored. On EWMA charts and CUSUM charts, the effect of a 'spike'-shift is smoothed out, so that these charts are insensitive to changes in the mean of short duration, see also Lin and Adams (1996). However, as we will see in the following example, the EWMA of modified residuals is much more efficient in detecting persistent step changes in the mean than Shewhart-type control charts.

6.2 A simulated example

In this section, we will illustrate the use of the control charts that were considered in previous chapters in the case of a persistent step change in the mean of AR(1) observations. To this end, we simulated 150 observations of an AR(1) process with $\phi = 0.6$. The expectation of the first 79 observations is 2; at observation 80, a shift to $\mu = 3.25$ is introduced, which corresponds to a shift in the mean of the process of $1\sigma_Y$. In Figure 6.3, a modified Shewhart chart, a Shewhart chart of residuals, a Shewhart chart of modified residuals and an EWMA chart of modified residuals are used to monitor

these observations.

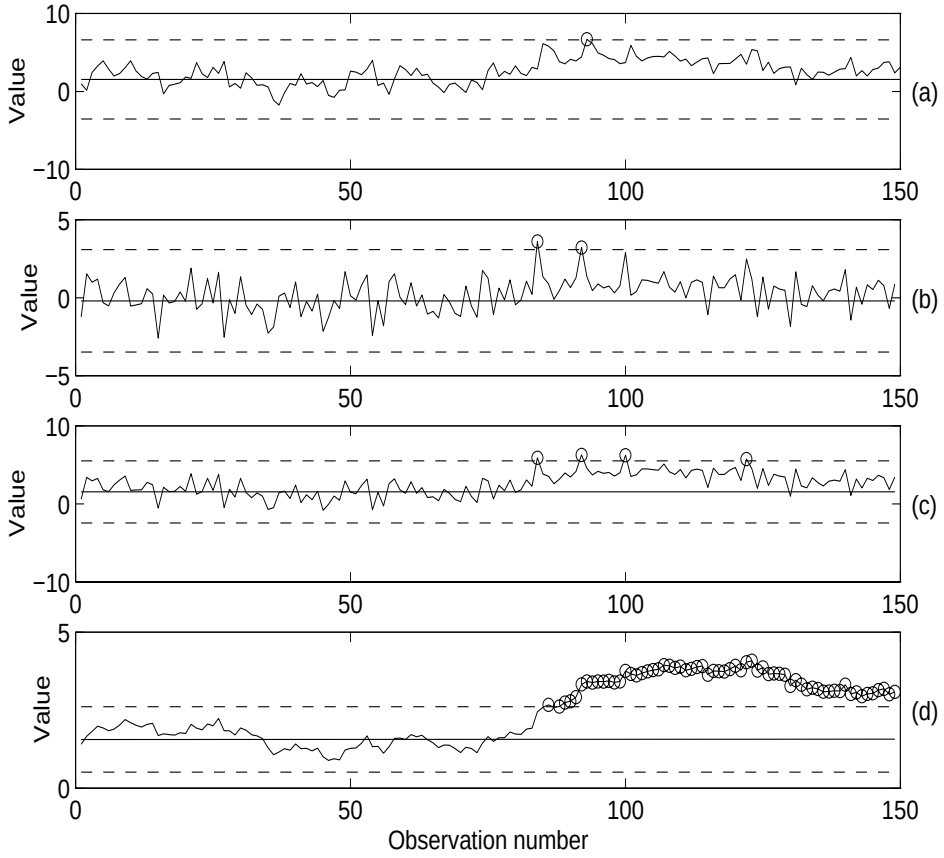


Figure 6.3: Control charts for a simulated AR(1) series.
(a)=modified Shewhart,
(b)=Shewhart chart of residuals,
(c)=Shewhart chart of modified residuals,
(d)=EWMA chart of modified residuals.

Figure 6.3(a) shows that the modified Shewhart chart signals at observation 93. The Shewhart chart of residuals in Figure 6.3(b) signals at the 84th and the 92nd residual. For the Shewhart chart of modified residuals in Figure 6.3(c) a value of $\lambda = 0.1$ is used for the EWMA. Out-of-control signals are observed at 84th, the 92nd, 100th, and the 122nd residual. The EWMA chart of modified residuals signals at observation 87, and every observation thereafter. The control limits are in all cases chosen such that

Chapter 7

Control charts for the spread

In the previous chapters, we considered control charts for the mean of individual serially correlated measurements. In Section 2.3, we mentioned that a control chart for the mean is in most cases accompanied by a control chart for the dispersion of the process outcomes. In this chapter, control charts for the spread of serially correlated data will be discussed. In the first section, we will focus on control charts for the spread of individual observations. In this case, it is not possible to estimate the spread on a given time point. The moving range control chart and so-called omnibus methods are discussed that are designed to circumvent this problem. In Sections 7.2 through 7.8, we will consider the case where the serially correlated data is subgrouped into samples, so that it becomes feasible to monitor estimators of the within-sample variance for a shift in the spread of the process.

As in the previous chapters, the discussion will be focused on a special case of serially correlated processes, viz. the AR(1) process. As argued earlier, the reason for only considering this type of serially correlated processes is that the AR(1) process is frequently encountered in practice. Furthermore, it may serve as an approximation to other time series models. However, the theory is easily adapted to include other time series models as well.

7.1 Control charts for the spread when $n = 1$

7.1.1 Introduction

In the previous chapters, we considered control charts for the mean of individual serially correlated measurements. There the variance of the process was assumed to be constant. The discussion focused on the efficiency of

detecting a shift in the mean of the process. In this section, we will assume that the mean of the process is constant. We only consider special causes that shift the variance of the process from σ_Y^2 to $b\sigma_Y^2$, where $b > 1$.

In the case of independently distributed individual observations, there is some controversy over the question of whether or not to use a separate chart for the spreading. There is a group of authors, such as Duncan (1986), Wheeler and Chambers (1990), Wetherill and Brown (1991) and Montgomery (1996), who recommend to use both a control chart for the mean, and a control chart for a measure of spread.

There are, however, other authors such as Nelson (1982, 1990) and Roes, Does and Schurink (1993) who argue that all the information is already contained within the X-chart. For example, when an increase in the variance of a process occurs, the limits on the control chart for the mean will be too tight. Hence, the probability of crossing one of the control limits of a control chart for the mean is increasing with $b > 1$. Therefore, the question has been raised whether a control chart for the spread adds extra power to a control chart for the mean when a shift in the spread of the process is to be detected.

For the case of individual observations, a chart for the spread could be based on the *moving range* (MR). The moving range at time t is defined as

$$MR_t = |X_t - X_{t-1}|$$

where $\{X_t\}$ is the sequence of independently distributed individual observations that is to be monitored. Roes, Does, and Schurink (1993) computed the conditional probability (assuming independence of the observations) of observing a signal on the MR-chart, given that an X-chart for the mean does not signal. For selected shifts in the mean and shifts in the variance, the probability

$$\mathrm{P}(MR_t \text{ signals } |E)\,, \tag{7.1}$$

is considered, where E is the event that the X-chart for the mean does not signal on time t and time $t-1$. In Table 7.1, the probabilities given by (7.1) are tabulated for various shifts in the mean and in the variance.

The values of $\mathrm{P}(MR_t \text{ signals } |E)$ in Table 7.1 differ from those of Table 7 of Roes, Does, and Schurink (1993). Roes, Does, and Schurink (1993) neglected the lower control limit of the MR-chart in the computation of $\mathrm{P}(MR_t \text{ signals } |E)$. The effect of ignoring the lower control limit is largest

Table 7.1: Conditional probabilities of an out-of-control signal on the *MR*-chart.

Shift in the mean	P(MR_t signals $\|E$)	P(X_t signals on X chart)
$\delta = 0.0$	0.00158	0.0020
$\delta = 0.5$	0.00191	0.0050
$\delta = 1.0$	0.00305	0.0183
$\delta = 1.5$	0.00523	0.0559
$\delta = 2.0$	0.00846	0.1378
$\delta = 2.5$	0.01241	0.2775
$\delta = 3.0$	0.01662	0.4641
Shift in the spread	**P(MR_t signals $\|E$)**	**P(X_t signals on X chart)**
$b = 1.0$	0.00158	0.0020
$b = 1.5$	0.00336	0.0394
$b = 2.0$	0.00492	0.1223
$b = 2.5$	0.00592	0.2164
$b = 3.0$	0.00654	0.3020

for the in-control situations. Roes, Does, and Schurink (1993) report P(MR_t signals $|E$) = 0.00058 for $\delta = 0.0$ and for $b = 1$. The difference becomes smaller for larger shifts in the mean or the variance.

The probabilities P(MR_t signals $|E$) are small for the out-of-control situations. Therefore, in Roes, Does and Schurink (1993) it is concluded that the contribution of the *MR*-chart to the power of discovering an out-of-control situation is small.

We will arrive at the same conclusion, using another argument. The fact that all the values of P(MR_t signals $|E$) are small could be the result of a poor design of the control charts, see Amin and Ethridge (1998). For example, the control limits of the *MR*-chart may be too wide. To properly assess the added value of the *MR*-chart, values of P(MR_t signals $|E$) for the out-of-control situations must be compared to the value of the same probability, evaluated in the in-control situation. It is the increase in power for detection of out-of-control situations, *in comparison to the in-control situation*, that makes a control chart sensitive in signalling special causes of variation.

The probability of an out-of-control signal on an X-chart when there

has been a shift in the mean of size $3\sigma_X$ is about 230 times as large as the in-control probability of an out-of-control signal. Given a shift in the mean of $3\sigma_X$, the conditional probability of a signal on the *MR*-chart is increased with a factor of about 10, compared to the in-control situation.

When the standard deviation of the process shifts from σ_X to $3\sigma_X$, the probability of an out-of-control signal on the X-chart is increased with a factor of about 150, in comparison to the in-control situation. The conditional probability of an out of control signal on the *MR*-chart is increased by a factor smaller than 4.

Hence, in agreement with Roes, Does, and Schurink (1993), we conclude that the *MR*-chart adds very little to the power of an X-chart when there has been a shift in the mean, and more surprisingly, the contribution to the power is even less when there has been a shift in the process spread.

The approach of Roes, Does and Schurink (1993) was criticized by Adke and Hong (1997). They consider the probabilities

$$\mathrm{P}\,[MR \text{ chart signals between } t+1 \text{ and } t+n \mid X \text{ chart does not signal between } t+1 \text{ and } t+n]$$

for various values of n. For $n = 2$, these probabilities are compared to those computed by Roes, Does, and Schurink (1993). The differences arise due to the fact that Roes, Does and Schurink (1993) assume that a shift in one of the process parameters occurs between X_{t-1} and X_t, whereas Adke and Hong (1998) assume that both X_{t-1} and X_t are drawn from an out-of-control distribution. In Amin and Ethridge (1998) it was suggested that the differences are also caused by the fact that Adke and Hong used a two-sided *MR*-chart, whereas Roes, Does and Schurink (1993) essentially used a one-sided *MR*-chart. However, the probabilities computed by Adke and Hong (1997) are also based on a one-sided *MR*-chart. In Table 7.2, conditional probabilities of a signal on different *MR*-charts are tabulated for various shifts in the variance, for the case where only X_t is out of control and for the case where both X_{t-1} and X_t are out of control.

The second column of Table 7.2 displays the probabilities of an out-of-control signal on a one-sided *MR*-chart when only X_t is out of control. In the third column the probabilities corresponding to a two-sided *MR*-chart are tabulated for the case where only X_t is out of control. The fourth column contains the probabilities of a signal on a one-sided *MR*-chart when both X_{t-1} and X_t are out of control. In the fifth column the probabilities

corresponding to a two-sided *MR*-chart are displayed for the case that both X_{t-1} and X_t are out of control.

Table 7.2: Conditional probabilities of a signal on a *MR*-chart at time t, given that X_{t-1} and X_t do not signal on the X-chart.

	X_t out of control		X_{t-1} and X_t out of control	
b	one-sided *MR*	two-sided *MR*	one-sided *MR*	two-sided *MR*
1.0	0.00058	0.00158	0.00058	0.00158
1.5	0.00254	0.00336	0.00939	0.01011
2.0	0.00420	0.00492	0.02257	0.02320
2.5	0.00525	0.00592	0.03294	0.03354
3.0	0.00590	0.00654	0.04009	0.04068

The probabilities in Table 7.2 show that $\mathrm{P}(MR_t < \mathrm{LCL}_{MR})$ is not extremely small, and cannot be neglected when computing the probability of a signal on a two-sided *MR*-chart.

Let us consider the difference between the probability of a signal on the *MR*-chart in the in-control situation and the probability of a signal in an out-of-control situation. This difference, which is a measure for the added power of the *MR*-chart, becomes smaller when a lower control limit is added to a one-sided *MR*-chart. Therefore, when it is decided to use both an X-chart and an *MR*-chart to monitor independent individual observations, it is advisable to use a one-sided *MR*-chart.

The foregoing analysis can also be performed for serially correlated observations. However, the probabilities of an out-of-control signal depend on the state of the process. Instead of comparing single probabilities as in the i.i.d. case described above, functions of the state of the process must be compared. This does not provide a clear argument whether or not to use an *MR*-chart in addition to an X-chart in the case of serially correlated individual observations. In addition, for the case of independently distributed individual observations, we argued that using an X-chart alone is nearly as effective as using both an X-chart and an *MR*-chart. Amin and Ethridge (1998) arrive at the same conclusion, by comparing the ARL of the X-chart alone to the ARL of the combined X-*MR* procedure. When the observations of the process are serially correlated, it is not to be expected either that a control chart based on moving ranges adds extra power to a control chart for the mean. Therefore, we advise to use only a control chart

for the mean to monitor a sequence of serially correlated observations.

However, Amin and Ethridge (1998) mention other considerations that justify the use of a combined procedure. For this reason, control charts for the spread of individual observations are briefly discussed in the next two subsections. In the sections following Section 7.2, we discuss control charts for the spread when the data is subgrouped first.

7.1.2 The moving range chart when $n = 1$

In the previous subsection, we discussed the *MR*-chart for independently distributed individual observations. In this subsection, individual AR(1) observations $\{Y_t\}$ are considered, with

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z},$$

where the ε_t are independently distributed as $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_{\varepsilon,t})$. Note that the mean of the process μ is not indexed by t, since we assume that the process is in control with respect to the mean. Instead, to indicate that a shift in the spread may occur, the variance of the disturbances is indexed by t. The moving range at time t is computed as

$$MR_t = |Y_t - Y_{t-1}|.$$

Suppose that the process is not only in control with respect to the mean, but also with respect to the spread ($\sigma^2_{\varepsilon,t}$ is constant for all t). Then $Y_t - Y_{t-1}$ follows a normal distribution with expectation zero and variance $2(1 - \phi)\sigma^2_Y$. For the expectation of MR_t we note that $MR_t^2/(2(1-\phi)\sigma^2_Y)$ follows a χ^2 distribution with one degree of freedom. Next we are interested in evaluating the expectation of Z^α, where Z follows a χ^2 distribution with one degree of freedom.

$$\begin{aligned}
\mathrm{E}(Z^\alpha) &= \int_0^\infty z^\alpha \frac{e^{-\frac{1}{2}z}}{\sqrt{2z}\,\Gamma(\frac{1}{2})}\,\mathrm{d}z \\
&= 2^\alpha\,\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\frac{1}{2})}\int_0^\infty \left(\frac{1}{2}\right)^{\alpha+\frac{1}{2}} \frac{z^{\alpha-\frac{1}{2}}e^{-\frac{1}{2}z}}{\Gamma(\alpha+\frac{1}{2})}\,\mathrm{d}z \\
&= 2^\alpha\,\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\frac{1}{2})}.
\end{aligned}$$

Hence, using $\alpha = 1/2$, we obtain the following well-known result for $\mathrm{E}(MR_t)$:

$$\mathrm{E}(MR_t) = \frac{2}{\sqrt{\pi}}\sqrt{1-\phi}\,\sigma_Y,$$

see also Cryer and Ryan (1990). The factor $2/\sqrt{\pi}$ is usually denoted by $d_2(2)$. Note that in the case $\phi = 0$, $MR_t/d_2(2)$ is an unbiased estimator of σ_Y. However, for $\phi \neq 0$, $MR_t/d_2(2)$ is biased for σ_Y by a factor of $\sqrt{1-\phi}$. For negative ϕ, σ_Y is overestimated, whereas for $\phi > 0$, σ_Y is underestimated.

When the mean of $MR_t/d_2(2)$ is used as an estimator for σ_Y, the bias does not disappear. This explains the most commonly encountered problems when monitoring serially correlated data for a shift in the mean. When there is positive autocorrelation, σ_Y is underestimated, and the control limits become too tight. Hence, false alarms are generated. For negative autocorrelation, the control limits are unnecessary wide, so that significant shifts in the process mean may go undetected (see for example Maragah and Woodall (1992)). In Subsection 7.5, we will see that estimating σ_Y on the basis of S^2 also leads to underestimation for negative ϕ, and to overestimation for positive ϕ. Therefore, determining the width of the control limits based on S^2 will lead to the same complaints. However, the bias in S^2 disappears when the number of observations is sufficiently large.

For the variance of MR_t we have

$$\begin{aligned}
\mathrm{Var}\,(MR_t) &= \mathrm{E}\left(|Y_t - Y_{t-1}|^2\right) - \left(\mathrm{E}(MR_t)\right)^2 \\
&= 2(1-\phi)\,\sigma_Y^2 - \frac{4}{\pi}(1-\phi)\,\sigma_Y^2 \\
&= 2(1-\phi)\left(1 - \frac{2}{\pi}\right)\sigma_Y^2.
\end{aligned}$$

For the covariance between MR_t and MR_{t+k} we firstly note that

$$\begin{aligned}
&\mathrm{Cov}\,(Y_t - Y_{t-1}, Y_{t+k} - Y_{t+k-1}) \\
&\quad = \mathrm{Cov}\,(Y_t, Y_{t+k}) - \mathrm{Cov}\,(Y_t, Y_{t+k-1}) \\
&\qquad - \mathrm{Cov}\,(Y_{t-1}, Y_{t+k}) + \mathrm{Cov}\,(Y_{t-1}, Y_{t+k-1}) \\
&\quad = -\phi^{k-1}\,(1-\phi)^2\,\sigma_Y^2,
\end{aligned}$$

so that we have for ρ_k, the correlation between $(Y_t - Y_{t-1})$ and $(Y_{t+k} - Y_{t+k-1})$

$$\rho_k = -\frac{1}{2}\phi^{k-1}(1-\phi). \tag{7.2}$$

With (A5) of Cryer and Ryan (1990) (see also page 92 of Johnson and Kotz (1972)) we obtain

$$\mathrm{Cov}(MR_t, MR_{t+k}) = \frac{4(1-\phi)}{\pi}\left(\sqrt{1-\rho_k^2} + \rho_k \arcsin(\rho_k) - 1\right)\sigma_Y^2,$$

where ρ_k is given in (7.2). Thus, $\{\rho_{MR,k}\}$, the autocorrelation function of the moving ranges of AR(1) data, can be computed as

$$\rho_{MR,k} = \frac{2}{\pi - 2}\left(\sqrt{1-\rho_k^2} + \rho_k \arcsin(\rho_k) - 1\right),$$

where ρ_k is given in (7.2). In Table 7.3, the autocorrelations between MR_t and MR_{t+k} are determined for lags $k = 1, 2, \cdots, 10$ for selected values of ϕ.

Table 7.3: Autocorrelations between MR_t and MR_{t+k} for selected values of ϕ.

$\phi =$	-0.9	-0.6	-0.3	0	0.3	0.6	0.9
$k = 1$	0.8809	0.5989	0.3852	0.2239	0.1084	0.0352	0.0022
$k = 2$	0.6928	0.2060	0.0334	0	0.0097	0.0126	0.0018
$k = 3$	0.5507	0.0732	0.0030	0	0.0009	0.0045	0.0014
$k = 4$	0.4401	0.0262	0.0003	0	0.0001	0.0016	0.0012
$k = 5$	0.3529	0.0094	0.0000	0	0.0000	0.0006	0.0009
$k = 6$	0.2837	0.0034	0.0000	0	0.0000	0.0002	0.0008
$k = 7$	0.2284	0.0012	0.0000	0	0.0000	0.0001	0.0006
$k = 8$	0.1842	0.0004	0.0000	0	0.0000	0.0000	0.0005
$k = 9$	0.1486	0.0002	0.0000	0	0.0000	0.0000	0.0004
$k = 10$	0.1201	0.0001	0.0000	0	0.0000	0.0000	0.0003

For positive ϕ, the autocorrelations of successive moving ranges are small. This can be explained by noting that for values of ϕ close to one, MR_t approximately equals $|\varepsilon_t|$, and the disturbances are assumed to be

independent. For strong positive autocorrelation, the moving ranges can be treated as statistics that are not serially correlated. This facilitates the design and the evaluation of the ARL of the *MR*-chart, since the ARL can then be approximated (given lower control limit LCL_{MR} and upper limit UCL_{MR}) by

$$\mathrm{ARL}_{MR} \approx \frac{1}{1 - \mathrm{P}(\mathrm{LCL}_{MR} < MR_t < \mathrm{UCL}_{MR})}.$$

Since the distribution of MR_t is not known, Monte Carlo simulation can be used to obtain the probability $\mathrm{P}(\mathrm{LCL}_{MR} < MR_t < \mathrm{UCL}_{MR})$. The control limits can be determined by requiring a certain in-control ARL and a certain out-of-control ARL.

In the foregoing, the *MR*-chart for individual serially correlated observations is discussed. However, the chart can also be applied to residuals of a fitted time series model. If the time series model is appropriate for the data, the residuals will be approximately uncorrelated. Furthermore, if the time series model and its parameters are known, a shift in the variance of the serially correlated data is transferred fully to a shift in the variance of the residuals, so that the main objection to the use of residuals charts for the mean does not apply to residuals charts for the spread. Therefore, the results of the *MR*-chart of residuals can be interpreted as in the case of independently distributed observations. In our view, the use of an *MR*-chart of residuals is to be preferred to the use of an *MR*-chart of correlated observations.

7.1.3 Omnibus control charts when $n = 1$

In Domangue and Patch (1991), a so-called omnibus EWMA control chart is proposed. The control chart is based on plotting statics

$$A_t = \lambda \left| \frac{Y_t - \mu}{\sigma_Y} \right|^{\alpha} + (1 - \lambda) A_{t-1}. \tag{7.3}$$

Domangue and Patch (1991) have evaluated the ARL behavior of the omnibus EWMA control chart for $\alpha = 0.5$ and $\alpha = 2$. The control chart is shown to react sensitively to (gradual or sudden) changes in the mean, and to (gradual or sudden) changes in the spread of the observations, and combinations thereof. In the same article, the performance of the omnibus EWMA control chart is compared to the behavior of other charts, such as various CUSUM schemes and combinations of X- and R-charts.

One of the CUSUM schemes considered is a so-called omnibus CUSUM chart, proposed by Hawkins (1981) and Healy (1987). It involves monitoring a sum

$$S_t = \max\left(0, \left|\frac{Y_t - \mu}{\sigma_Y}\right|^{\alpha} - k + S_{t-1}\right),$$

where k is the reference value. An out-of-control signal is given at time t if $S_t > h$ for a decision interval h. This procedure can also be used to detect shifts in the mean or in the spread, or combinations thereof. In the ARL computations of Domangue and Patch (1991), it is shown that the performance of the omnibus EWMA control chart is comparable to the performance of the omnibus CUSUM control chart.

In MacGregor and Harris (1993), a special case of the omnibus EWMA control chart is considered, viz. $\alpha = 2$. The resulting chart is called the Exponentially Weighted Mean Square (EWMS) control chart. It is shown that the quantities A_t are weighted sums of χ^2 random variables. Furthermore, an approximating distribution function for A_t is discussed. The statistics A_t are approximately distributed as $\chi^2(\nu)/\nu$, where the number of degrees of freedom ν depends upon the exponential weighting parameter λ, the correlation structure of the Y_t's and the degrees of freedom associated with A_t. MacGregor and Harris use this approximation to design the EWMS control chart.

In addition, MacGregor and Harris (1993) discuss a variant to the EWMS control chart, where μ is replaced by an estimate for μ_t in formula (7.3). They call the resulting statistic the Exponentially Weighted Moving Variance (EWMV). In the article, control limits for the EWMV chart are derived for the case of independent observations and for the case of autocorrelated observations. The EWMS chart is shown to respond both to changes in the mean and the variance. The EWMV chart only responds to changes in the variance.

In case of serially correlated data, the control charts of this subsection should preferably be applied to the residuals of an appropriate time series model.

7.2 Subgrouping serially correlated data

Let us again consider a sequence of observations $\{Y_t\}$, whose elements are serially correlated. In the previous section we discussed the difficulties that arise when such observations are to be monitored with a control chart for

the spread. In the remaining sections of this chapter we consider the case where these problems are circumvented by creating subgroups first.

Assume that from the sequence $\{Y_t\}$, samples of size n are drawn. Between every two samples k realizations of $\{Y_t\}$ are not observed. Figure 7.1 illustrates the way samples are drawn.

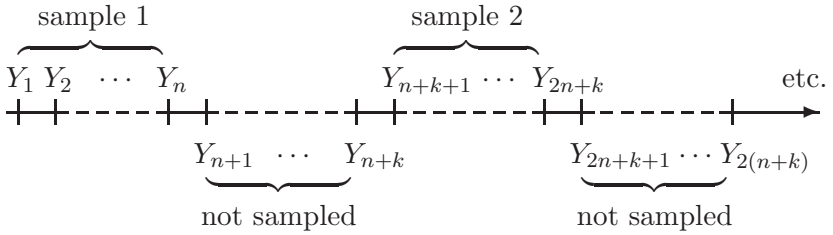


Figure 7.1: The method of sampling.

It is assumed that the underlying measurements $Y_1, Y_2, \cdots$ (see Figure 7.1) are generated by an AR(1) model:

$$Y_t - \mu_t = \phi(Y_{t-1} - \mu_{t-1}) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z},$$

where the ε_t are independently distributed as $\mathcal{N}(0, \sigma^2_{\varepsilon_t})$. The index t in μ_t and $\sigma^2_{\varepsilon_t}$ is used to indicate that a special cause of variation may shift the unconditional mean or variance of the process.

Furthermore, we will assume that the samples are rational subgroups (see Subsection 2.3.1); the samples are drawn in such a way that shifts in the mean or in the spread of the process only occur *between* samples, and *not within* samples. This allows us to assume that within each sample i, the unconditional means $\mu_{(n+k)(i-1)+1}, \cdots, \mu_{(n+k)(i-1)+n}$ are constant and equal to, say μ^*_i.

Likewise, the unconditional variances $\sigma^2_{\varepsilon_{(n+k)(i-1)+1}}, \cdots, \sigma^2_{\varepsilon_{(n+k)(i-1)+n}}$ are assumed to be constant within sample i and equal to, say $\sigma^2_{\varepsilon^*_i}$.

After creation of subgroups, it becomes feasible to construct control charts for the spread of the process. However, subgrouping the data has also consequences for the control chart of the mean of the process. Therefore, in Section 7.3, we will focus on control charts for the mean of samples of AR(1) observations. In Sections 7.4 through 7.7, control charts for the spread of subgrouped AR(1) data are discussed. In Section 7.8 the ARL performance of these control charts is evaluated.

7.3 Control charts for batch means

In this section we consider control charts based on the mean of samples of AR(1) data. The mean of the ith sample is denoted by $\bar{Y}_i$. Note that the relation between $\bar{Y}_i$ and the underlying observations is

$$\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{(i-1)(n+k)+j}.$$

Several authors have investigated the properties of the series $\{\bar{Y}_i\}$. For the case $k = 0$, Kang and Schmeiser (1987) refer to Anderson (1976) to show that if $\{Y_t\}$ is a stationary ARMA(p,q) process, then the sequence of sample means $\{\bar{Y}_i\}$ is a stationary ARMA(p,$\bar{q}$) process, with $\bar{q} = p - \lfloor (p-q)/n \rfloor$, where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x, and n is the sample size. For the case of AR(1) observations, $p = 1$, and $q = 0$, so that the sequence of sample means is an ARMA(1,1) process when the first element of a sample is drawn directly after the last element of the previous sample ($k = 0$). Alwan and Radson (1992b) consider the case $k > 0$, and show that the sequence of sample means is also an ARMA(1,1) process when there are k unobserved AR(1) realizations between successive samples.

In the in-control situation, we have $\mu_i^* = \mu$ for samples $i = 1, 2, \cdots$, and $\sigma^2_{\varepsilon_i^*} = \sigma^2_\varepsilon$ for $i = 1, 2, \cdots$. Using the results of the cited references, the sequence of sample means of AR(1) observations obeys the following dynamic structure

$$\bar{Y}_i = (1 - \bar{\phi})\mu + \bar{\phi}\bar{Y}_{i-1} + \bar{\varepsilon}_i - \bar{\theta}\bar{\varepsilon}_{i-1} \tag{7.4}$$

where $\{\bar{\varepsilon}_i\}$ is a sequence of i.i.d. $\mathcal{N}(0, \sigma^2_{\bar{\varepsilon}})$ disturbances, $\bar{\phi}$ is the AR-parameter, and $\bar{\theta}$ is the MA-parameter.

It is now of interest how the parameters $\bar{\phi}$, $\bar{\theta}$ and $\sigma^2_{\bar{\varepsilon}}$ of model (7.4) relate to the parameters ϕ and σ^2_ε of the model for the underlying individual observations, and n and k. In general, it is very difficult to determine the relationships between the parameters of an ARMA(p,q) model and the parameters of the ARMA(p,$\bar{q}$) model of the aggregated time series. However, Alwan and Radson (1992b) tracked down closed form formulas for the case of underlying AR(1) observations. They show that

$$\bar{\phi} = \phi^{n+k},$$

where it is to be noted that $\bar{\phi}$ does not depend on σ_ε^2. The relation between $\bar{\theta}$ and $\sigma_{\bar{\varepsilon}}^2$ of the aggregated model and ϕ, σ_ε^2, n, and k is derived as follows. The variance of $\bar{Y}_i$ is evaluated in terms of ϕ, σ_ε^2, n, and k:

$$\begin{aligned}\mathrm{Var}(\bar{Y}_i) &= \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{1}{n}\left[1+\frac{2}{n}\sum_{j=1}^{n-1}(n-j)\phi^j\right] \\ &= \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{1}{n^2}\left[n(1+2C_2)-2\phi C_1\right],\end{aligned}$$

where C_1 is defined as $\sum_{j=1}^{n-1} j\phi^{j-1}$, and $C_2 = \sum_{j=1}^{n-1}\phi^j$; notation borrowed from Alwan and Radson (1992b). This expression is equated to the variance of $\bar{Y}_i$, expressed in parameters of the aggregated model:

$$\mathrm{Var}(\bar{Y}_i) = \frac{1+\bar{\theta}^2-2\bar{\theta}\bar{\phi}}{1-\bar{\phi}^2}\sigma_{\bar{\varepsilon}}^2.$$

Expressing the covariance between successive means in terms of ϕ, σ_ε^2, n, and k leads to:

$$\begin{aligned}\mathrm{Cov}(\bar{Y}_i,\bar{Y}_{i+1}) &= \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\phi^{n+k}}{n^2}\left[n+\sum_{j=1}^{n-1}j\phi^{j-n}+\sum_{j=1}^{n}(n-j)\phi^j\right] \quad (7.5)\\ &= \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\phi^{n+k}}{n^2}\left[n(1+C_2)+\left(\frac{1}{\phi^{n-1}}-\phi\right)C_1\right].\end{aligned}$$

Next, this expression is equated to

$$\mathrm{Cov}(\bar{Y}_i,\bar{Y}_{i+1}) = \frac{(1-\bar{\phi}\bar{\theta})(\bar{\phi}-\bar{\theta})}{1-\bar{\phi}^2}\sigma_{\bar{\varepsilon}}^2.$$

Using $\bar{\phi}=\phi^{n+k}$, we have two equations in two unknowns. Solving for $\bar{\theta}$ and $\sigma_{\bar{\varepsilon}}^2$ leads to the solutions

$$\begin{aligned}\bar{\theta}_1 &= \frac{-K_1(1+\phi^{2(n+k)})+2K_2\phi^{n+k}+K_3}{2(K_2-\phi^{n+k}K_1)}\\ \bar{\theta}_2 &= \frac{-K_1(1+\phi^{2(n+k)})+2K_2\phi^{n+k}-K_3}{2(K_2-\phi^{n+k}K_1)},\end{aligned}$$

where, as in Alwan and Radson (1992b),

$$K_1 = \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{1}{n^2}C_3\left[n(1+2C_2)-2\phi C_1\right],$$

$$K_2 = \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\phi^{n+k}}{n^2}C_3\left[n(1+C_2)+\left(\frac{1}{\phi^{n-1}}-\phi\right)C_1\right],$$

$$K_3 = \sqrt{(\phi^{2(n+k)}-1)\left[(2K_2-\phi^{n+k}K_1)^2-K_1^2\right]},$$

and

$$C_3 = 1-\phi^{2(n+k)}.$$

It can be shown that the two ARMA(1,1) models corresponding to $\bar{\theta}_1$ and $\bar{\theta}_2$ have the same autocorrelation function. However, only $\bar{\theta}_1$ satisfies the invertibility condition (see for example Harvey (1993)). The two solutions for $\sigma_{\bar{\varepsilon}}^2$ corresponding to $\bar{\theta}_1$ and $\bar{\theta}_2$ are

$$\sigma_{\bar{\varepsilon},1}^2 = \frac{K_1(1+\phi^{2(n+k)})-2K_2\phi^{n+k}+K_3}{2-2\phi^{2(n+k)}}$$

$$\sigma_{\bar{\varepsilon},2}^2 = \frac{K_1(1+\phi^{2(n+k)})-2K_2\phi^{n+k}-K_3}{2-2\phi^{2(n+k)}}.$$

From these formulas it can be seen that $\sigma_{\bar{\varepsilon}}^2$ does depend on all four parameters, and that $\sigma_{\bar{\varepsilon}}^2$ is proportional to σ_ε^2. In Table 7.4, we computed the parameters $\bar{\phi}$ and $\bar{\theta}$ for selected values of ϕ for the cases $n=5, k=0$, $n=5, k=5$, and $n=5, k=10$.

Table 7.4: ARMA(1,1) parameters $\bar{\phi}$ and $\bar{\theta}$ as functions of ϕ, n, and k.

	$n=5,\ k=0$		$n=5,\ k=5$		$n=5,\ k=10$	
	$\bar{\phi}$	$\bar{\theta}$	$\bar{\phi}$	$\bar{\theta}$	$\bar{\phi}$	$\bar{\theta}$
$\phi=-0.9$	-0.5905	0.0103	0.3487	-0.0045	-0.2059	0.0024
$\phi=-0.6$	-0.0778	0.0793	0.0060	-0.0060	-0.0005	0.0005
$\phi=-0.3$	-0.0024	0.0563	0.0000	-0.0001	0.0000	0.0000
$\phi=0.3$	0.0024	-0.0735	0.0000	-0.0002	0.0000	0.0000
$\phi=0.6$	0.0778	-0.1769	0.0060	-0.0129	0.0005	-0.0010
$\phi=0.9$	0.5905	-0.2457	0.3487	-0.0849	0.2059	-0.0440

The values for $\bar{\theta}$ differ from those presented in Table 1 of Alwan and Radson (1992b). This is due to an error in their expression for the covariance between $\bar{Y}_i$ and $\bar{Y}_{i-1}$; compare formula (7.6) to equation (13) of Alwan and Radson (1992b). As a result, K_2 is also erroneously defined. In addition, an error slipped into the definition of K_1; compare the definition above to the definition of K_1 in Appendix I of Alwan and Radson (1992b). This results in imaginary solutions for $\bar{\phi}$ in some cases. This does not happen if the computations are performed as indicated above.

Note from Table 7.4 that both the absolute value of $\bar{\phi}$ and the absolute value of $\bar{\theta}$ decrease considerably when n or k is increased. Hence, taking sample means reduces the serial correlation in the statistic that is to be monitored in the control chart. This effect of subgroup taking is also observed by Runger and Willemain (1995). These authors also consider the case of monitoring a weighted sample mean, where the weights are chosen in such a way that successive weighted means are uncorrelated. This procedure was introduced by Bischak, Kelton, and Pollock (1993).

From Table 7.4, it can also be concluded that the MA parameter $\bar{\theta}$ is much smaller throughout than the AR parameter $\bar{\phi}$, so that the series $\{\bar{Y}_i\}$ will, in most cases, be very similar to an AR(1) series. Alwan and Radson (1992b) have encountered this phenomenon frequently in practice.

The analysis above reduces monitoring sample means of AR(1) data to monitoring ARMA(1,1) data. This means that the control charts of Chapters 3 through 5 can be applied, provided that the control chart is adapted to account for ARMA(1,1) dependence in the data. Appendix A might offer guidance on which control chart to use, since the ARL behavior for some of these charts for ARMA(1,1) models is tabulated there.

In the following section, control charts for the spread of subgrouped serially correlated data are discussed.

7.4 The moving range chart when $n > 1$

In this subsection, we will investigate the properties of the *MR*-chart for AR(1) data that are grouped into subsamples. Consider again a sequence of AR(1) observations $\{Y_t\}$, generated by (2.3)

$$Y_t - \mu_t = \phi(Y_{t-1} - \mu_{t-1}) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}.$$

Suppose that subgroups of size n are formed in such a way that there are k observations between two successive samples. Thus, if the first sample

consists of observations $\{Y_1, Y_2, \cdots, Y_n\}$, the second sample consists of the observations $\{Y_{n+k+1}, Y_{n+k+2}, \cdots, Y_{2n+k}\}$, and so on.

Within each sample, $n-1$ moving ranges are computed. From the ith sample, which consists of the observations

$$Y_{(n+k)(i-1)+1}, Y_{(n+k)(i-1)+2}, \cdots, Y_{(n+k)(i-1)+n},$$

the moving ranges $MR_{i,2}, \cdots, MR_{i,n}$ are computed as

$$MR_{i,j} = |Y_{(n+k)(i-1)+j} - Y_{(n+k)(i-1)+j-1}| \qquad \text{for } j = 2, \cdots, n.$$

The ith average moving range is then computed as

$$\overline{MR}_i = \frac{1}{n-1} \sum_{j=2}^{n} MR_{i,j}.$$

These statistics are plotted in the $\overline{MR}$ control chart, and compared to limits $\mathrm{LCL}_{\overline{MR}}$ and $\mathrm{UCL}_{\overline{MR}}$.

Using the results of Subsection 7.1.2, we obtain the following properties of $\overline{MR}_i$. For the expected value of $\overline{MR}_i$ we have

$$\begin{aligned} \mathrm{E}(\overline{MR}_i) &= \frac{1}{n-1} \sum_{j=2}^{n} \mathrm{E}(MR_{i,j}) \\ &= \frac{2}{\sqrt{\pi}} \sqrt{1-\phi}\, \sigma_Y. \end{aligned}$$

Note that $\overline{MR}_i / d_2(2)$ is biased as an estimator for σ_Y by a factor of $\sqrt{1-\phi}$, and that the bias does not disappear as the sample size increases. For the variance of $\overline{MR}_i$ we have

$$\begin{aligned}
\mathrm{Var}(\overline{MR}_i) &= \frac{1}{(n-1)^2}\left(\sum_{j=2}^{n}\mathrm{Var}(MR_{i,j}) + \right. \\
&\qquad \left. + \; 2\sum_{j=2}^{n-1}\sum_{l=j+1}^{n}\mathrm{Cov}(MR_{i,j}, MR_{i,l})\right) \\
&= \frac{2(1-\phi)}{\pi(n-1)^2}\,\sigma_Y^2\left((n-1)(\pi-2) + \right. \\
&\qquad \left. + \; \sum_{l=1}^{n-2} 4(n-l-1)\left(\sqrt{1-\rho_l^2} + \rho_l \arcsin(\rho_l) - 1\right)\right)
\end{aligned}$$

where ρ_l is defined in Equation (7.2) of Section 7.1.2. The expression for $\mathrm{Var}(\overline{MR}_i)$ of AR(1) data does not simplify much further as it does for the i.i.d. case, see Cryer and Ryan (1990). This is due to $\rho_l \neq 0$ for all l if $\phi \neq 0$. However, the expression for $\mathrm{Var}(\overline{MR}_i)$ above is relatively easy to compute numerically in practical cases.

For the covariance between $\overline{MR}_i$ and $\overline{MR}_{i+m}$ we have

$$\begin{aligned}
&\mathrm{Cov}(\overline{MR}_i, \overline{MR}_{i+m}) \\
&= \frac{1}{(n-1)^2}\sum_{j=2}^{n}\sum_{l=2}^{n}\mathrm{Cov}(MR_{i,j}, MR_{i+m,l}) \\
&= \frac{4(1-\phi)}{\pi(n-1)^2}\,\sigma_Y^2 \times \\
&\qquad \times \sum_{l=-n+2}^{n-2}(n-1-|l|)\left(\sqrt{1-\rho_{q-l}^2} + \rho_{q-l}\arcsin(\rho_{q-l}) - 1\right)
\end{aligned}$$

where $q = (n+k)m$, and ρ_j is defined in Equation (7.2). The correlation between $\overline{MR}_i$ and $\overline{MR}_{i+m}$ is denoted by $\rho_{\overline{MR},m}$ and is defined as

$$\rho_{\overline{MR},m} = \frac{\mathrm{Cov}(\overline{MR}_i, \overline{MR}_{i+m})}{\mathrm{Var}(\overline{MR}_i)}.$$

The correlations $\rho_{\overline{MR},m}$ are extremely small for choices of n and k that will be used in practice (say 5 and 10 respectively). The correlations are the largest if the smallest possible sample size is chosen, which is $n = 2$, and when there are no unobserved observations between the samples, i.e. $k = 0$. Table 7.5 tabulates the correlations between successive moving ranges for $n = 2$, $k = 0$ for selected values of ϕ and the lag m ranging from 1 to 10.

Table 7.5: Autocorrelations between $\overline{MR}_i$ and $\overline{MR}_{i+m}$ for selected values of ϕ, with $n = 2$ and $k = 0$.

$\phi =$	-0.9	-0.6	-0.3	0	0.3	0.6	0.9
$m = 1$	0.6928	0.2060	0.0334	0	0.0097	0.0126	0.0018
$m = 2$	0.4401	0.0262	0.0003	0	0.0001	0.0016	0.0012
$m = 3$	0.2837	0.0034	0.0000	0	0.0000	0.0002	0.0008
$m = 4$	0.1842	0.0004	0.0000	0	0.0000	0.0000	0.0005
$m = 5$	0.1201	0.0001	0.0000	0	0.0000	0.0000	0.0003
$m = 6$	0.0784	0.0000	0.0000	0	0.0000	0.0000	0.0002
$m = 7$	0.0513	0.0000	0.0000	0	0.0000	0.0000	0.0001
$m = 8$	0.0336	0.0000	0.0000	0	0.0000	0.0000	0.0001
$m = 9$	0.0220	0.0000	0.0000	0	0.0000	0.0000	0.0001
$m = 10$	0.0144	0.0000	0.0000	0	0.0000	0.0000	0.0000

Note that the first five rows of Table 7.5 equal respectively the second, fourth, sixth, eighth and tenth row of Table 7.3. The correlations between successive $\overline{MR}_i$ decrease quickly with k and also with n. For practical purposes, the successive correlations may be treated as uncorrelated statistics. As in the case for $n = 1$, this facilitates the design and the evaluation of the ARL of the $\overline{MR}$ control chart. The ARL will be approximated (given lower control limit $\mathrm{LCL}_{\overline{MR}}$ and upper limit $\mathrm{UCL}_{\overline{MR}}$) by

$$\mathrm{ARL}_{\overline{MR}} \approx \frac{1}{1 - \mathrm{P}\big(\mathrm{LCL}_{\overline{MR}} < \overline{MR}_{i,j} < \mathrm{UCL}_{\overline{MR}}\big)}.$$

Again, the probability $\mathrm{P}\big(\mathrm{LCL}_{MR} < \overline{MR}_{i,j} < \mathrm{UCL}_{MR}\big)$ will be evaluated by Monte Carlo simulation. The control limits are determined by requiring a certain in-control ARL and a certain out-of-control ARL. In Subsection 7.8, the ARL curves of the moving range chart are compared to other schemes for monitoring the spread using subgrouped AR(1) observations.

7.5 The S^2-chart

In this subsection, we will discuss the S^2-chart for the spread of AR(1) observations. Consider a sequence of AR(1) observations $\{Y_t\}$, generated by (2.3)

$$Y_t - \mu_t = \phi(Y_{t-1} - \mu_{t-1}) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}.$$

Suppose that subgroups of size n are formed in such a way that there are k observations between two successive samples.

The sequence of observed sample variances $\{S_i^2\}$ is monitored by a lower control limit LCL_{S^2} and an upper control limit UCL_{S^2}. The width of these limits depend on the standard deviation of the S_i^2, which is proportional to σ_Y^2. If the control limits are chosen such that in the in-control case $\mathrm{P}(S_i^2 < LCL_{S^2}) = \mathrm{P}(S_i^2 > UCL_{S^2})$, then the control limits will not be centered around $\mathrm{E}(S_i^2)$ due to skewness of the distribution of S_i^2.

Consider the case where the process is in control with respect to the mean, that is, assume $\mu_t = \mu$ for all t. Then the expectation of the ith sample variance equals

$$\begin{aligned}
\mathrm{E}(S_i^2) &= \mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_{(n+k)(i-1)+j} - \bar{Y}_i\right)^2\right) \\
&= \mathrm{E}\left(\frac{1}{n-1}\left\{\sum_{j=1}^{n}\left(Y_{(n+k)(i-1)+j} - \mu\right)^2 - n\left(\bar{Y}_i - \mu\right)^2\right\}\right) \\
&= \frac{n}{n-1}\left(\sigma_Y^2 - \mathrm{Var}(\bar{Y}_i)\right),
\end{aligned}$$

where

$$\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{(n+k)(i-1)n+j}.$$

For the variance of $\bar{Y}_i$ we have from Anderson (1971)

$$\begin{aligned}
\mathrm{Var}(\bar{Y}_i) &= \frac{\sigma_Y^2}{n}\left[1 + 2\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\phi^j\right] \\
&= \frac{\sigma_Y^2}{n^2(1-\phi)^2}\left(n - 2\phi - n\phi^2 + 2\phi^{n+1}\right),
\end{aligned}$$

so that we obtain

$$\mathrm{E}(S_i^2) = \sigma_Y^2 \left(\frac{n}{n-1}\right)\left[1 - \frac{n - 2\phi - n\phi^2 + 2\phi^{n+1}}{n^2\,(1-\phi)^2}\right]. \tag{7.6}$$

Hence, for finite samples, the sample variance is a biased estimator for σ_Y^2. The bias disappears for $n \to \infty$. In Figure 7.2, the behavior of the factor of σ_Y^2 in Equation (7.6) is plotted against the sample size n for various ϕ.

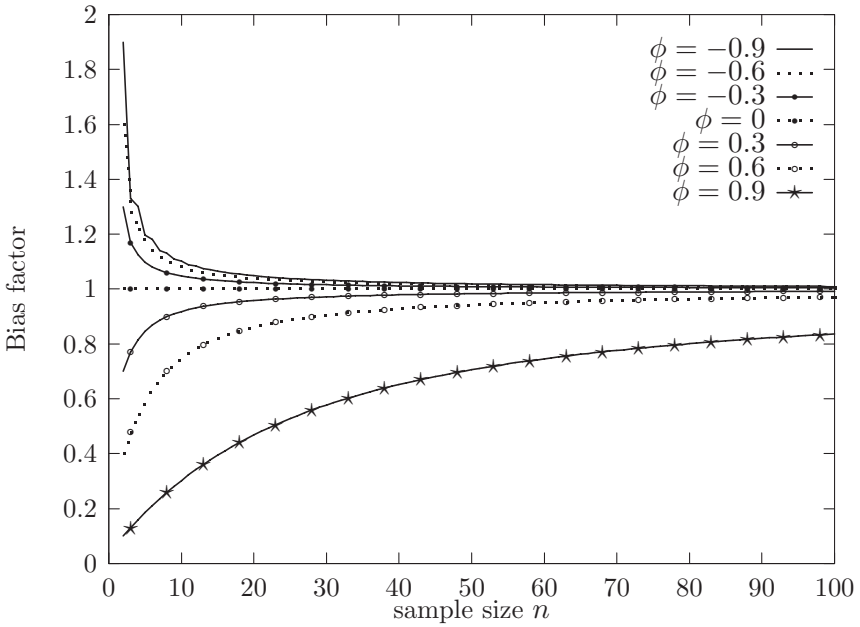


Figure 7.2: Behavior of the bias factor in $\mathrm{E}(S_j^2)$ against n for various ϕ.

For the uncorrelated case $\phi = 0$, S_i^2 is an unbiased estimator of σ_Y^2. However, for negative ϕ the process variance is overestimated, whereas for $\phi > 0$ the variance is underestimated. Note that this bias does not disappear if the process variance is estimated as the average of a number of sample variances. For a common subgroup sample size of $n = 5$, the process is overestimated by 20% if $\phi = -0.9$, by 16% if $\phi = -0.6$, and by 10% if $\phi = -0.3$. With this sample size σ_Y^2 is underestimated by 15% if $\phi = 0.3$, by 40% if $\phi = 0.6$, and by 81% if $\phi = 0.9$! For large positive autocorrelation, the bias factor converges very slowly to one. Sample sizes larger than $n = 40$ are required to obtain a reasonable accurate estimate.

The bias in S_i^2 as an estimator for σ_Y^2 poses no direct problems for the construction of an S^2-chart for the spread. It is the first purpose of the control chart to detect changes in the process variance. This can very well be done using biased statistics, as long as the level of the control chart is not interpreted as the size of the variance of the process.

Next the covariance of S_i^2 and S_{i+m}^2 will be derived. Without loss of generality, we assume that $\mu_t = \mu = 0$. Define

$$\boldsymbol{Y} = \begin{pmatrix} Y_{(n+k)(i-1)+1} \\ Y_{(n+k)(i-1)+2} \\ \vdots \\ Y_{(n+k)(i-1)+n} \\ Y_{(n+k)(i+m-1)+1} \\ Y_{(n+k)(i+m-1)+2} \\ \vdots \\ Y_{(n+k)(i+m-1)+n} \end{pmatrix}.$$

The variance matrix of $\boldsymbol{Y}$ is

$$\mathrm{Var}(\boldsymbol{Y}) = \begin{pmatrix} A & B \\ B' & A \end{pmatrix}$$

with

$$A = \frac{\sigma_\varepsilon^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \cdots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \cdots & 1 \end{pmatrix},$$

and

$$B = \frac{\sigma_\varepsilon^2}{1-\phi^2}\phi^{(n+k)m}\begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \frac{1}{\phi} & 1 & \phi & \cdots & \phi^{n-2} \\ \frac{1}{\phi^2} & \frac{1}{\phi} & 1 & \cdots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\phi^{n-1}} & \frac{1}{\phi^{n-2}} & \frac{1}{\phi^{n-3}} & \cdots & 1 \end{pmatrix}.$$

Then $(n-1)S_i^2$ can be written as $(P\boldsymbol{Y})'P\boldsymbol{Y} = \boldsymbol{Y}'P\boldsymbol{Y}$, where

$$P = \begin{pmatrix} M & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

with $\mathbf{0}$ an $(n \times n)$ matrix with all elements equal to zero, and

$$M = I_n - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}',$$

where I_n is the $(n \times n)$ identity matrix and $\boldsymbol{\iota}$ is an $(n \times 1)$ vector with all elements equal to one. Analogously, $(n-1)S_{i+m}^2$ can be written as $(Q\boldsymbol{Y})'Q\boldsymbol{Y} = \boldsymbol{Y}'Q\boldsymbol{Y}$, where

$$Q = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & M \end{pmatrix}.$$

Using Corollary 4.1 from Magnus and Neudecker (1979), we have

$$\begin{aligned} (n-1)^2 \operatorname{Cov}(S_i^2, S_{i+m}^2) &= 2 \operatorname{tr}\left[P \operatorname{Var}(\boldsymbol{Y})\, Q \operatorname{Var}(\boldsymbol{Y})\right] \\ &= 2 \operatorname{tr}\begin{pmatrix} MBMB' & MBMA \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &= 2 \operatorname{tr}(MB(BM)') \\ &= 2\sum_{p=1}^{n}\sum_{q=1}^{n}(MB)_{pq}^2, \end{aligned}$$

where $(MB)_{pq}$ is the element in the pth row and qth column of MB.

For $(n-1)^2 \operatorname{Var}(S_i^2) = (n-1)^2 \operatorname{Var}(S_{i+m})^2$ we have, using again Corollary 4.1 from Magnus and Neudecker (1979),

$$\begin{aligned} \operatorname{Var}(\boldsymbol{Y}'P\boldsymbol{Y}) &= 2\operatorname{tr}\left[(P\operatorname{Var}(\boldsymbol{Y}))^2\right] + 4\mu^2 \boldsymbol{e}_{2n}' P \operatorname{Var}(\boldsymbol{Y}) P \boldsymbol{e}_{2n} \\ &= 2\operatorname{tr}\left[(MA)^2\right]. \end{aligned}$$

Hence for $\rho_{S^2,m}$, the correlation coefficient between S_i^2 and S_{i+m}^2, we have

$$\rho_{S^2,m} = \frac{\operatorname{tr}\left[MB(BM)'\right]}{\operatorname{tr}\left[(MA)^2\right]}.$$

The value of $\rho_{S^2,m}$ is decreasing in n, k, and m, and is increasing with $|\phi|$. However, for practical choices of n and k, the values of the autocorrelation function $\{\rho_{S^2,1}, \rho_{S^2,2}, \cdots\}$ are very close to zero. As an illustration, in Table 7.6 we tabulated the case where samples of size $n = 5$ are drawn, and where there are $k = 10$ unobserved AR(1) realizations between two samples. The correlations are tabulated for lags $m = 1, \cdots, 5$ for the case where there is high negative autocorrelation, $\phi = -0.8$, and for the case when there is high positive autocorrelation, $\phi = 0.8$.

Table 7.6: Autocorrelations between S_i^2 and S_{i+m}^2 for $n = 5$, $k = 10$.

	$\phi = -0.8$	$\phi = 0.8$
$m = 1$	0.00328040956151	0.00533538388005
$m = 2$	0.00000406095034	0.00000660488533
$m = 3$	0.00000000502721	0.00000000817645
$m = 4$	0.00000000000622	0.00000000001012
$m = 5$	0.00000000000001	0.00000000000001

From Table 7.6 it can be observed that the correlation between successive sample variances is minimal, even in the case of severe autocorrelation (either negative or positive). Therefore, the correlation between successive sample variances will be neglected in the construction of an S^2-chart.

Treating the sample variances as independent statistics allows us to determine the ARL for the S^2-chart assuming control limits LCL_{S^2} and UCL_{S^2} by

$$ARL \approx \frac{1}{1 - \mathrm{P}\left(\mathrm{LCL}_{S^2} < S_i^2 < \mathrm{UCL}_{S^2}\right)}.$$

The control limits LCL_{S^2} and UCL_{S^2} are determined in the usual way by requiring a certain in-control ARL and a certain out-of-control ARL. To this end, the distribution of S_i^2 needs to be known. Schmid (1995a) remarks that the usual results for quadratic forms do no apply here, due to the fact that $P\,\mathrm{Var}(\boldsymbol{Y})$ and $Q\mathrm{Var}(\boldsymbol{Y})$ are not idempotent. To overcome this problem, Schmid (1995a) uses Laguerre expansion of the distribution function of S_i^2 to obtain the ARL of the S^2-chart. Based on ARL requirements, he obtains upper control limits that account for serial correlation in the data.

The results of Schmid (1995a) show that, compared to the case $\phi = 0$, the control limits become wider for $\phi < 0$. This is to be expected in the light of Figure 7.2, where it is shown that S_i^2 overestimates σ_Y^2 for finite subsamples. Conversely, the control limits of the S^2-chart are tighter for the case $\phi > 0$, due to underestimation of S_i^2.

In Section 7.8, we will simulate the ARL behavior of the S^2-chart. This will be compared to the ARL behavior of the other charts for the spreading of subgrouped serially correlated data that are discussed in this chapter. Note that the S chart is not separately discussed. The properties of this control chart are comparable to the properties of the S^2-chart.

7.6 The R-chart

In Alwan and Radson (1992a), the effect of positive first-order autocorrelation on the range of a subsample is investigated. The subsample range of sample i is defined as

$$\begin{aligned} R_i &= \max\left[Y_{(n+k)(i-1)+1}, Y_{(n+k)(i-1)+2}, \cdots, Y_{(n+k)(i-1)+n}\right] + \\ &\quad - \min\left[Y_{(n+k)(i-1)+1}, Y_{(n+k)(i-1)+2}, \cdots, Y_{(n+k)(i-1)+n}\right]. \end{aligned}$$

In the case of independent observations, the mean of a subsample range increases with the sample size. Simulation studies that were performed by Alwan and Radson (1992a) show that the effect of positive autocorrelation on the mean of R_i is small, compared to the effect of the sample size n. This indicates that $R_i/d_2(n)$ is a biased estimator for σ_Y, since the latter is increasing with $|\phi|$. In Table 7.7, we simulated the mean of $R_i/d_2(n)$ for $\phi = -0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9$, using sample sizes $n = 2, 3, 4, 5, 6$. The

number of unobserved AR(1) realizations between two samples was taken to be $k = 0$, since the results are not influenced by different values of k.

Table 7.7: Simulated values for $\mathrm{E}[R_i/d_2(n)]$ for selected values of ϕ and n.

	σ_Y	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
$\phi=-0.9$	2.2942	3.1552 (0.0075)	2.4435 (0.0049)	2.2230 (0.0041)	2.1145 (0.0036)	2.0557 (0.0033)
$\phi=-0.6$	1.2500	1.5717 (0.0038)	1.3881 (0.0025)	1.3242 (0.0020)	1.2933 (0.0017)	1.2759 (0.0016)
$\phi=-0.3$	1.0483	1.1975 (0.0029)	1.1286 (0.0019)	1.1040 (0.0015)	1.0893 (0.0013)	1.0809 (0.0012)
$\phi=0$	1.0000	1.0062 (0.0024)	1.0011 (0.0017)	0.9994 (0.0014)	1.0009 (0.0012)	1.0008 (0.0011)
$\phi=0.3$	1.0483	0.8772 (0.0021)	0.9184 (0.0015)	0.9418 (0.0013)	0.9604 (0.0012)	0.9711 (0.0011)
$\phi=0.6$	1.2500	0.7867 (0.0019)	0.8587 (0.0015)	0.9091 (0.0013)	0.9458 (0.0012)	0.9768 (0.0011)
$\phi=0.9$	2.2942	0.7288 (0.0017)	0.8144 (0.0014)	0.8914 (0.0014)	0.9505 (0.0013)	1.0033 (0.0013)

Table 7.7 shows that $R_i/d_2(n)$ is biased for σ_Y. For negative ϕ, σ_Y is overestimated in most cases. For positive ϕ, the bias is negative. Except for $\phi = -0.9$, the bias reduces with larger n. For negative ϕ, the bias reduction is faster than for positive ϕ. However, the results for $\phi = -0.9$ show that the bias is positive for $n = 2, 3$, and σ_Y is underestimated for $n = 4, 5, 6$. Note that the entries in the column corresponding to $n = 2$ agree with the analytical results in the second column of Table 3.1.

Alwan and Radson (1992a) also performed simulation studies to determine the effect of AR(1) dependence in the observations on the correlation of successive subsample ranges. The results show that the simulated first-order autocorrelations are within the standard 95% confidence interval, for practical choices of n and k, and $\phi > 0$. Especially for larger values of k, the correlation between successive subsample ranges are indistinguishable from zero.

Utilizing these results, we decide to design the R-chart as if the R_i's were independent. The ARL for the R-chart assuming control limits LCL_R

and UCL_R is then computed as

$$ARL \approx \frac{1}{1 - \mathrm{P}(\mathrm{LCL}_R < R_i < \mathrm{UCL}_R)}.$$

The probability in this expression will be evaluated by Monte Carlo simulation. The control limits LCL_R and UCL_R are again determined by making two requirements on ARL curve.

7.7 The residuals chart

Residuals charts have already been discussed in Chapters 3, 4 and 5. In those applications, residuals of a fitted time series model are monitored for a change in the mean of the correlated observations. However, a change in the variance of the correlated observations will also be transferred to the variance of residuals. This motivates the use of a control chart for the spread of residuals to monitor the variance of serially correlated observations.

Suppose that we have AR(1) observations available, generated by

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \varepsilon_t \qquad \text{for } t \in \mathbb{Z}.$$

The expectation of the observations is assumed to be constant and equal to μ for all t. The AR parameter ϕ is also assumed to be constant. However, at time T, the variance of the disturbances may change from, say, $\sigma^2_{\varepsilon,0}$ to $\sigma^2_{\varepsilon,1}$. The residual e_t is computed as

$$(Y_t - \mu) - \phi(Y_{t-1} - \mu) = \varepsilon_t,$$

where the last equality only holds if μ and ϕ are known. A shift in the variance of the disturbances is thus transferred to a shift in the variance of the residuals of the same size . Hence, a residuals control chart for the spread is unaffected by the serial correlation in the data. Furthermore, since the residuals are (approximately) uncorrelated, the standard control charts for monitoring the variance of a process can be used and interpreted without further alterations. In the following section, the ARL behavior of an R-chart based on residuals is compared to the ARL performance of the $\overline{MR}$-chart, the S^2-chart, and the R-chart, respectively.

7.8 ARL comparison

In this section, the ARL behavior of the control charts for the spreading of subgrouped data is evaluated. In Table 7.8(a), simulated ARL values of the $\overline{MR}$ chart for subgrouped AR(1) data are tabulated for $\phi = -0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9$. The values in the first row correspond to the case where the variance is in control, i.e. the variance of $\{Y_t\}$ is σ_Y. The entries in the second row are the simulated ARL values when the variance of $\{Y_t\}$ has shifted from from σ_Y to $2\sigma_Y$. In the third row, the effect of a shift from σ_Y to $3\sigma_Y$ on the ARL is simulated. Tables 7.8(b), (c) and (d) are set up similarly. Table 7.8(b) contains the simulated ARL values of the S^2-chart. In Table 7.8(c), the simulated ARL behavior of the R-chart for subgrouped AR(1) data is tabulated. The results of the simulations for an R-chart of residuals are presented in Table 7.8(d).

All entries of Table 7.8 are computed assuming that subgroups of size $n = 5$ are formed, and that there are $k = 10$ unobserved AR(1) realizations between successive samples. The values are based on 10,000 replications. The bracketed numbers are the simulated standard errors.

From the results of Table 7.8, we conclude that of the four control charts considered, the R-chart based on residuals has the best ARL performance for all values of ϕ. Recall from Chapters 3, 4 and 5 that a shift in the mean of serially correlated data is, in general, not transferred to a shift of the same size in the mean of the residuals. A residuals chart for the spread of serially correlated data does not have such a disadvantage. If the appropriate time series model and its process parameters are known, a shift in the standard deviation of the process will be transferred to a shift of the same size in the standard deviation of the residuals.

Another argument in favor of the use of residuals-based control charts for the spread is that such charts are easy to design. If the fitted time series model is appropriate for the data, the residuals will be approximately uncorrelated, so that the control limits do not depend on the parameters of the model. The width of the control limits depends only on σ_ε, the standard deviation of the disturbances of the time series model, and the desired ARL behavior of the control chart. Hence, multipliers of σ_ε that result in desired ARL behavior of the control chart have to be established only once. After that, control limits for monitoring the spread of any type of serially correlated process can be found by multiplying the appropriate multiplier with the standard deviation of the disturbances of the process.

In the light of the foregoing, we recommend to monitor the spread of a serially correlated process with a residuals-based control chart.

Table 7.8: Simulated ARL values of various control charts for the spread of subgrouped serially correlated data.

(a) $\overline{MR}$ chart of subgrouped data

$\phi =$	-0.9	-0.6	-0.3	0.0	0.3	0.6	0.9
$b = 1$	374.8 (3.7)	376.9 (3.8)	367.7 (3.6)	367.6 (3.7)	360.8 (3.7)	359.2 (3.6)	374.7 (3.8)
$b = 2$	85.4 (0.9)	35.0 (0.3)	25.9 (0.2)	19.6 (0.2)	17.0 (0.2)	15.7 (0.2)	14.3 (0.1)
$b = 3$	37.2 (0.4)	13.7 (0.1)	9.3 (0.1)	7.2 (0.1)	6.0 (0.1)	5.5 (0.0)	5.0 (0.0)

(b) S^2-chart of subgrouped data

$\phi =$	-0.9	-0.6	-0.3	0.0	0.3	0.6	0.9
$b = 1$	359.6 (3.6)	377.1 (3.8)	375.7 (3.8)	383.1 (3.8)	366.0 (3.7)	366.8 (3.7)	364.4 (3.6)
$b = 2$	79.1 (0.8)	26.4 (0.3)	14.8 (0.1)	11.7 (0.1)	14.1 (0.1)	21.3 (0.2)	26.8 (0.3)
$b = 3$	33.3 (0.3)	9.2 (0.1)	5.0 (0.0)	4.1 (0.0)	4.9 (0.0)	7.3 (0.1)	9.6 (0.1)

(c) R-chart of subgrouped data

$\phi =$	-0.9	-0.6	-0.3	0.0	0.3	0.6	0.9
$b = 1$	358.7 (3.6)	379.7 (3.8)	369.8 (3.7)	364.1 (3.6)	368.8 (3.8)	369.4 (3.7)	376.3 (3.7)
$b = 2$	57.3 (0.6)	22.4 (0.2)	15.2 (0.1)	12.8 (0.1)	14.8 (0.1)	20.2 (0.2)	26.5 (0.3)
$b = 3$	22.1 (0.2)	7.9 (0.1)	5.2 (0.0)	4.5 (0.0)	5.1 (0.0)	7.0 (0.1)	9.2 (0.1)

(d) R-chart of residuals of subgrouped data

$\phi =$	-0.9	-0.6	-0.3	0.0	0.3	0.6	0.9
$b = 1$	370.8 (3.8)	370.8 (3.6)	363.9 (3.7)	364.1 (3.6)	365.8 (3.6)	362.5 (3.6)	367.4 (3.6)
$b = 2$	12.8 (0.1)	13.0 (0.1)	12.9 (0.1)	12.8 (0.1)	12.7 (0.1)	12.9 (0.1)	12.7 (0.1)
$b = 3$	4.5 (0.0)	4.4 (0.0)	4.5 (0.0)	4.5 (0.0)	4.5 (0.0)	4.5 (0.0)	4.4 (0.0)

Chapter 8

Philips case

In the previous chapters, we discussed various aspects of monitoring serially correlated data. In the present chapter, we will discuss a quality improvement project in which we ourselves were involved. Not all the topics that are brought to the attention of the reader have a direct relation with the contents of this thesis. However, in Section 8.4 it is described how monitoring serial correlation was handled in practice. We find it useful to provide a more or less complete discussion of the problems that an SPC practitioner might encounter in practice. One of these is the problem of monitoring serially correlated data with a control chart.

8.1 Introduction

Philips Semiconductors Stadskanaal is a leading supplier of diodes. Customers of Philips Semiconductors Stadskanaal are among others the automotive industry, the communications sector and manufacturers of consumer electronics. These customers are producers themselves, whose product quality is partly determined by the quality of the diodes. Therefore, Philips' customers are demanding with respect to the quality of the diodes. They require reliable, well functioning diodes that are easy to process.

In most cases the customers solder the diodes on to a printed circuit board, so that ease of processing is to a large extent determined by solderability of the diodes. In order to ensure solderability of the diodes, Philips Stadskanaal is applying a protective tin/lead layer to the connection points of the diodes.

Insufficient layer thickness or wrong composition of the layer has been the cause of customer complaints in the past. Philips Stadskanaal is there-

fore looking for ways to improve the process of applying the tin/lead layer, the objective being a better solderability.

In the last five years, Philips Stadskanaal has acquired valuable experience in process improvement through successful application of *Statistical Process Control* (SPC) techniques (see Does, van Oord and Trip (1994)). The key to this success may be found in the approach that was chosen towards implementation of SPC.

At Philips Stadskanaal SPC techniques are implemented by so-called *Process Action Teams* (PAT's). A PAT is constituted as follows: operators are important members because they are heavily involved in the process. The team is chaired by a responsible technical engineer, and may be complemented by a quality engineer, a service mechanic and/or a developer. A neutral outsider with profound knowledge of and experience with *Statistical Process Control* completes the team. A PAT receives a clear mission what to improve, and the means to realize their plans (see Does, Roes and Trip (1999)).

A PAT was started to improve the process step of tin-plating diodes. In this chapter, we will go through some of the achievements of this PAT. In the following subsections, we will introduce respectively the product, the production process, and the data that are gathered during the project. In the sections thereafter we will describe more or less chronologically the developments around this process step.

8.1.1 The product

A diode is an important electrical component possessing the special property of conducting current in only one direction, whereas it has a high resistance in the reverse direction. Diodes are used in all kinds of electrical circuits in for instance TV sets, computers, automotive ignition systems, telecommunication apparatus, power supplies for X-ray generators, and a great variety of consumer electronics.

Philips Semiconductors Stadskanaal makes four different types of glass-encapsulated diodes. Each type is made using a different manufacturing process. In this chapter we will restrict ourselves to one type of diodes, the so-called *Surface Mounted Implosion Diodes* (SMID's). An exploded view of a SMID is depicted in Figure 8.1.

One of the striking things in Figure 8.1 is that the connection points of the diode are flanges rather than leads. This makes this type of diodes suitable for surface mounting, which explains the first part of the name of this type of diodes. Surface mounted diodes are smaller than leaded diodes

Figure 8.1: A surface mounted implosion diode.

and easier to process in automated industry. Philips Semiconductors Stadskanaal produces both types of diodes. However, the tin/lead layer is more critical for solderability of surface-mounted diodes than for solderability of leaded diodes, so that we will restrict ourselves to surface-mounted products.

8.1.2 The production process

The crystal is the heart of a diode. It is made of a small slice of the semiconducting material silicon. By contaminating both sides of a silicon wafer, one side with phosphorus, the other with boron, the silicon conducts current in one direction and blocks current in the other direction. The contaminations are brought into the silicon wafer by a diffusion process. Subsequently, several crystals are produced from one wafer.

The crystal of the diode of Figure 8.1 is placed between two studs/flange pairs. An implosion process follows that tightly fits a glass body around the crystal and the studs. The implosion process takes place in a vacuum so that the edge of the crystal only contacts the nonconductor glass. This prevents charge carriers from traveling through the diode in any other way than through the crystal. The word 'implosion' is part of the name of the diode in Figure 8.1 to distinguish it from diodes that are produced using other methods to apply a nonconducting body.

The studs are made of molybdenum, the flanges are made of copper. Molybdenum is a metal that has a good conductivity. When oxidized, it has the additional property that it adheres well to the glass body. The adherence of the molybdenum studs to the glass body provides mechanical strength and a hermetic sealing of the diode. The latter is very important since exposing the crystal to open air ruins the electrical properties of the diode.

Whereas oxidation of the molybdenum studs is helpful, this is not the case for oxidation of the copper flanges. One of the properties of oxidized copper is that the adherence to tin is bad. The customers of the type of diodes of Figure 8.1 solder the diodes with tin, so that solderability is adversely influenced by the oxidation of the copper flanges. To overcome this problem, the oxidized copper is removed, and a tin/lead layer that prevents the flanges from coming into contact with oxygen is applied. In the remainder of this chapter, we will focus on this production step.

After tin-plating, a 100% inspection follows. Several electrical characteristics of *each* diode are measured, and it is decided to accept or to reject the diode. For some types of diodes, a classification in sub-types is made. The tin/lead layer is not only important for solderability but also facilitates the inspection. If it were absent, a potential difference would arise between copper-oxide on the surface of the flange and the pure copper inside the flange. This would introduce a large measurement error.

In the following process step, a white band is painted on the glass body of the accepted diodes (see Figure 8.1). The placement of the band indicates the cathode side of the diode. The electrical characteristics can be deduced from a code that is also printed on the glass body.

As a final process step, the diodes are packed in such a way that the customers can process the diodes in an automated way.

8.1.3 Relevant quality characteristics of the tin/lead layer

In the remainder of this chapter, we confine ourselves to only one of the process steps, namely applying a tin/lead layer. As discussed in the previous subsection, the tin/lead layer is important for the customers, since it determines the solderability of the diodes. The layer also facilitates inspection, which is the process step directly following the tin-plating of the diodes. The quality of the tin/lead layer is determined by two characteristics: solderability and composition.

The first quality characteristic is difficult to ascertain, since it is the net result of a complex of determinants such as thickness of the layer,

composition of the layer, pollution of the layer with organic materials, and the like. Therefore, "*a good solderability*" is translated into a requirement on a more ascertainable characteristic of the layer: viz. layer thickness.

If the diodes were soldered right after production, any thickness of the tin/lead layer would provide sufficient protection against oxidation. However, the diodes are shipped to factories and warehouses all over the world, where they may be kept in stock for some time. If the thickness of the tin/lead layer is not sufficient, diffusion of copper atoms in the tin/lead layer will result in copper atoms reaching the surface of the layer, and these atoms will oxidize. This phenomenon is one of the causes of bad solderability. For this reason, a lower specification limit is set for the thickness of the layer. A minimal thickness of 1μm (1×10^{-6}m) proves to be sufficient to warrant good solderability after a two years stay in any warehouse (provided certain conditions concerning relative humidity and temperature are met).

Excessive thickness may cause problems during the 100% inspection in the one but last step of the production. Diodes with a tin/lead layer that is too thick (say more than 50μm) will not pass the sieve that is used to prevent crooked diodes from entering the measuring apparatus. However, a strict upper specification limit is not used in practice, and will therefore not be considered. The target thickness of the tin/lead layer is 10μm.

The second quality characteristic, the composition of the layer, concerns the ratio of tin and lead in the tin/lead layer. This characteristic is directly observable on the products. If the portion of lead in the tin/lead layer is too large, the layer will be too soft. Putting a diode with too soft a layer through the measuring apparatus will cause some of the tin/lead layer to be worn off. These grindings smudge the glass body of subsequent diodes, which may lead to leakage of current over the body. Such diodes are rejected, although the electrical characteristics may have been perfect before entering the measuring apparatus. To guard against waste caused by this phenomenon, the lower specification limit of the tin portion in the tin/lead layer is set to 77%. The target value is 80%. In the next section we will describe how the tin/lead layer is applied.

8.1.4 The tin-plating process

The protective layer is applied to the diode by electrogalvanizing the connection points with a tin/lead alloy. A schematic view of the tin-plating process is depicted in Figure 8.2.

The tin-plating takes place in a conducting chemical bath. A rotating

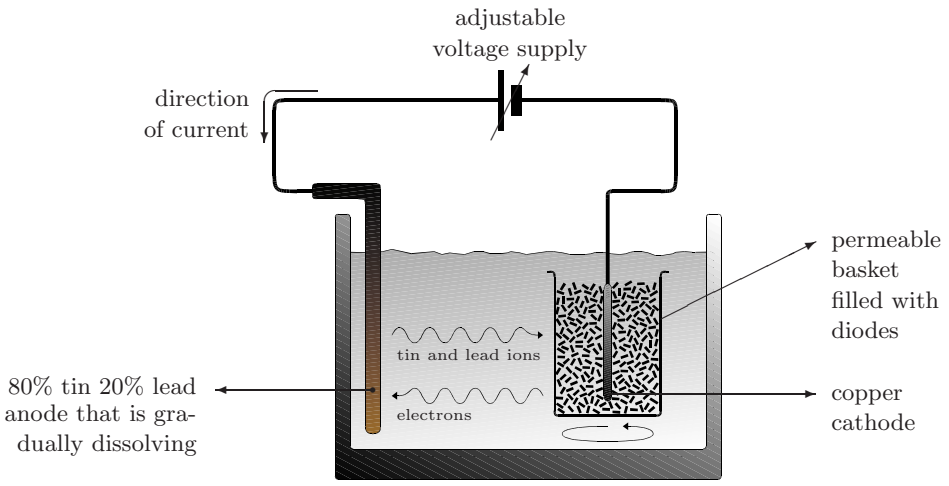


Figure 8.2: A schematic view of the tin-plating process.

basket that is permeable with respect to the tin-plating liquid is filled with diodes and placed into the bath. The anode side of the voltage supply consists of a 80% tin, 20% lead alloy that is gradually dissolving in the bath liquid. The dissolved metal ions precipitate on diodes contacting the cathode side of the voltage supply. That is, on the cathode side we have the following chemical reactions:

$$\mathrm{Sn}^{2+} + 2\mathrm{e}^{-} \longrightarrow \mathrm{Sn}$$

and

$$\mathrm{Pb}^{2+} + 2\mathrm{e}^{-} \longrightarrow \mathrm{Pb},$$

where Sn is the chemical symbol for tin, Pb is the chemical symbol for lead, and electrons are denoted by e^{-}.

As a result of this reaction on the cathode side of the voltage supply, the concentration of metal ions in the liquid will decrease. By using an anode of a 80% tin, 20% lead alloy (see Figure 8.2) it is possible to replenish the metal ions. With current running through the bath, electrons of tin and lead molecules in the anode are sent away into the direction of the voltage supply and the remaining metal ions dissolve in the liquid. Hence, the reverse of the chemical reaction that takes place at the cathode will

take place at the anode. The net effect is that tin and lead molecules are transferred from the anode of the voltage supply to diode flanges.

Important parameters of the process that influence the quality of the tin/lead layer are among others the temperature of the bath, the current density, the length of the time span the diodes stay in the bath, the voltage and the chemical composition of the bath. In order to maintain a high quality of the tin/lead layer, controlling each of these parameters is required. In practice it turned out to be very difficult to devise a proper control strategy that could deal with exhaustion of the bath due to the production process. This is the problem the PAT focused on. For the combined effect of other parameters we will assume that they do not have a systematic effect on the tin/lead layer.

The chemical composition of the tin-plating bath changes over time. This is the result of exhaustion of the bath due to running the production process, and evaporation of the tin-plating liquid. In order to maintain a good quality of the tin/lead layer, the concentrations of the components of the bath must fall between certain limits. The bath is a mixture of the following five components:

1. acid;
2. an Sn^{2+} solution;
3. a Pb^{2+} solution;
4. brightener;
5. formalin.

The acid is thought of as an important component since it takes care of the conductivity of the bath. The two metal solutions are also considered to be important since the Sn^{2+} and Pb^{2+} ions present in them precipitate on the diodes to form the tin/lead layer. The remaining components, brightener and formalin, were thought to play a certain part in the tin-plating process. However, it turned out that their importance was underrated (see Subsection 8.3.4). The brightener takes care of a smooth layer, whereas the formalin helps to control the proportions of tin and lead in the resulting layer.

During the production process, the composition and the volume of the tin-plating liquid changes. Changes in the composition are due to chemical reactions that take place in the production process. Furthermore, some components (e.g. formalin) evaporate faster than others, resulting in a

change in their relative proportions. The overall volume of the bath decreases due to evaporation and to the dragging out of the tin-plating liquid together with the diodes. As a result, the tin-plating bath needs to be replenished regularly.

For each of the components, a lower limit is set on the concentration. If the concentration of one of the components falls below its limit, the quality of the tin/lead layer deteriorates. Likewise, on the concentrations of acid, Sn^{2+} solution, and Pb^{2+} solution an upper limit is set.

If the composition of the bath is such that one or more of the concentrations fall outside the limits, action is required. Fortunately, highly concentrated solutions for each of the components are available so that concentrations can be raised by adding an appropriate mix of these solutions to the bath. The bath can be diluted using demineralized water.

Before the PAT was started the bath used to be replenished with fixed additions of the separate components, independent of the actual composition of the bath. This replenishment strategy was not satisfactory, since it resulted in a high level of chemicals being used, whereas the bath did not appear to be stable. The instability of the bath affected the quality of the tin/lead layers on the diodes in such a way that customer complaints were received. A new replenishment strategy was, therefore, adopted that used the actual contents of the bath to determine the additions. To this end, the bath was analyzed every day. Based on the results of the analysis, a computer program determined what additions were necessary to ensure that the bath was fit for production again. The new strategy resulted in a large reduction in the amounts of chemicals being used. The quality of the tin/lead layer however remained unsatisfactory.

Therefore, further improvement was necessary. At this point the author got involved in the problem. The first improvement he was able to suggest concerned the way in which the additions were computed. In the next section this problem is formulated as a *Linear Programming* (LP) problem, for which an optimal solution (in the sense of minimal costs of additions) can be found.

8.2 Computing additions to the tin-plating bath

8.2.1 Introduction

In the previous section it was discussed that the chemical composition of the tin-plating bath changes over time. In order to maintain a good quality of the tin/lead layer, the concentrations of the components of the bath

must fall between certain limits. If the composition of the bath is such that one or more of the concentrations fall outside the limits, action is required. Fortunately, there are highly concentrated solutions for each of the components which can be added in case its concentration has become too low. If the concentration of one of the components is too high, demineralized water can be added to lower the concentration.

So far the problem seems quite straightforward: concentrations that are too low can be raised by adding some of the appropriate solutions. Concentrations that are too high can be lowered by thinning the bath. However, computing how much is needed of every solution is not as straightforward as it may seem at first sight. The problem is that all five concentrations change if one of the components is added. This may cause trouble if we compute the additions for each of the components one by one. This can be illustrated by means of an example.

Suppose that we have computed how much of the first four components must be added so that the requirements are met. Then it may happen that the addition of the fifth component lowers the concentrations of the first four in such a way that some of these fall below their lower limits. This means that an additional amount of such components is required. But these additions in their turn lower the concentrations of all other components, which again may lead to extra additions, etc.

Obviously, this "one-by-one" strategy brings with it the danger of recommending large additions, even in situations where it would have been possible to recondition the bath using smaller additions. Furthermore, the strategy completely ignores the costs of the additions to be made. In this section, the replenishment problem is formulated as a *Linear Programming* model that has minimization of the costs of the additions as its objective, whereas requirements with regard to the concentrations serve as restrictions. The input of the model is a chemical analysis of the bath, and the output is a prescription of how much of each component must be added so that the bath meets all requirements. The output will be an *optimal* solution to the replenishment problem in the sense that it is not possible to find a cheaper set of additions that will bring all the concentrations within their limits.

Minimization of costs of the additions is not the only advantage of this approach. As a result of too large additions, it regularly happened that tin-plating liquid had to be drained off to make place for the recommended additions. However, the liquid contains heavy metals, and must be purified before it is allowed to be drained off. Hence, besides the high cost associated with adding too much of expensive chemicals to the bath, additional costs

are charged by the purifying department for draining off the surplus of tin-plating liquid. The *Linear Programming* approach discussed in this section allows us to take these costs into account, too. The practical examples that are discussed at the end of this section show that draining off old tin-plating liquid can be avoided in some cases. Besides cost reduction, this is also beneficial for the environment.

This section is organized as follows: in Subsection 8.2.2, we will introduce the mathematical notation that is used in the remainder of this section; this in its turn will be followed by a mathematical formulation of the problem in Subsection 8.2.3. The Simplex algorithm will be used to solve the *Linear Programming* model for which a starting point is required. In Subsection 8.2.4, a standard starting point is derived that can be used for every instance of the problem. In Subsection 8.2.5 we will compare some of the results of our model to the results of a computer program that was used previously at Philips Semiconductors Stadskanaal. A short conclusion in Subsection 8.2.6 ends this section.

8.2.2 Preliminaries

In Table 8.1, an overview is presented of the mathematical notation that will be used in the remainder of this section.

Table 8.1: Overview of the notation.

component	observed concentration	lower, upper limit	addition in liters	concentration of addition	cost of addition
acid	m_1 (gr/l)	l_1, u_1	x_1	d_1 (gr/l)	c_1 (Dfl/l)
Sn^{2+} solution	m_2 (gr/l)	l_2, u_2	x_2	d_2 (gr/l)	c_2 (Dfl/l)
Pb^{2+} solution	m_3 (gr/l)	l_3, u_3	x_3	d_3 (gr/l)	c_3 (Dfl/l)
brightener	m_4 (gr/l)	l_4	x_4	d_4 (ml/l)	c_4 (Dfl/l)
formalin	m_5 (gr/l)	l_5	x_5	d_5 (gr/l)	c_5 (Dfl/l)
demin. water			x_6		c_6 (Dfl/l)
drain off			x_7		c_7 (Dfl/l)
acid in Sn^{2+} sol.				d_* (gr/l)	

As the input of our model, we have the measurements of the concentrations of the bath components. The observed concentration of component i, ($i = 1, \cdots, 5$), is denoted by m_i.

For all of the components, a lower limit is set on the concentration. On the concentration of acid, Sn^{2+} solution, and Pb^{2+} solution also an upper limit is set. We denote the lower limit for component i by l_i and the upper limit for component i by u_i.

Highly concentrated solutions are available for each of the five components. These solutions can be added in order to raise the corresponding concentrations in the tin-plating liquid. This results in five *decisions*: for each component, we have to determine the size of the addition, thereby taking into account that, after replenishment, all requirements on the concentrations must be met. With decision i we associate a *decision variable* x_i that indicates the number of liters we add from component i for $i = 1, \cdots, 5$. If the concentration of one of the components is too high, demineralized water can be added. This calls for a sixth decision variable x_6 indicating the number of liters of demineralized water we decide to add. Furthermore, in cases where the bath volume is high, some tin-plating liquid may have to be drained off before making the additions associated with $x_1, x_2, \cdots, x_6$ in order to prevent the bath from overflowing. To this end, we define x_7 as the number of liters we decide to drain off *before* making the additions. We emphasize 'before' because the number of liters to be drained off was not taken in consideration in the old situation. The new bath volume was reported, whether the bath was overflowing or not. So, *after* the additions were computed it was known how many liters had to be drained off. This is in general not equal to the number of liters to be drained off *before* making additions, since draining off affects the additions to be made.

The concentrations of the highly concentrated solutions are also needed in the computations. We denote them by $d_1, \cdots, d_5$. Acid is available as a separate component. However, it is also the solvent of the Sn^{2+} solution. The mathematical model that will be developed in the next section has to take into account that with every addition of Sn^{2+} solution, acid will be added to the bath, too. We let d_* denote the concentration of acid in the Sn^{2+} solution.

With each of the decision variables, costs are associated. For the first six variables this is simply the price per liter, whereas for the cost associated with x_7 the cost price of purifying one liter of tin-plating liquid is used. We denote these costs coefficients by $c_1, \cdots, c_7$.

8.2.3 Mathematical formulation of the problem

Our objective is to minimize the costs of replenishing the bath, under the restriction that the concentrations fall between predetermined limits. With

the symbols introduced in the preceding section we can express our objective as follows

$$\text{minimize } c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + c_5x_5 + c_6x_6 + c_7x_7.$$

It is also possible to formulate the restrictions on the concentrations in mathematical terms. If we let V_{old} denote the volume of the bath before replenishing, our new bath volume will be $V_{\text{old}}+x_1+x_2+x_3+x_4+x_5+x_6-x_7$. Since the new bath volume may not exceed a certain maximum bath volume $b_{\max} = 600$ liters, we have as a first constraint

$$V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7 \leq b_{\max}.$$

The new bath volume must exceed a certain minimal value $b_{\min} = 450$ liters, leading to the following constraint:

$$V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7 \geq b_{\min}.$$

The tin-plating liquid contains $m_2(V_{\text{old}} - x_7)$ grammes of tin directly after an amount of x_7 liters is drained off and just before any additions are made. We make the decision to add d_2x_2 grammes of tin, so that after replenishment the bath contains $m_2(V_{\text{old}} - x_7) + d_2x_2$ grammes of tin. If we divide this by the new bath volume, we have the concentration of tin after replenishment. Since this concentration must exceed l_2, we arrive at the following constraint

$$\frac{m_2(V_{\text{old}} - x_7) + d_2x_2}{V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7} \geq l_2$$

which can be rewritten as

$$\begin{aligned} l_2x_1 + (l_2 - d_2)x_2 + l_2x_3 + l_2x_4 + l_2x_5 + \\ + l_2x_6 + (m_2 - l_2)x_7 \leq (m_2 - l_2)V_{\text{old}}. \end{aligned}$$

Analogously, we find the following constraint associated with the upper limit of the tin concentration in the bath

$$\begin{aligned} -u_2x_1 + (d_2 - u_2)x_2 - u_2x_3 - u_2x_4 - u_2x_5 - u_2x_6 + \\ + (u_2 - m_2)x_7 \leq (u_2 - m_2)V_{\text{old}}. \end{aligned}$$

Following exactly the same line of reasoning, we find constraints associated with the lower and upper limit of the concentration of lead, and a constraint associated with the lower limit of the concentration of brightener.

So far, we skipped the constraints associated with limits of the concentration of acid and formalin. The reason for this is that these constraints are different from the constraints above.

For the constraints on the concentration of acid, the difference is due to the fact that there are two ways to add acid to the bath. First, like all other components, we have acid available as a separate component for making additions to the tin-plating liquid. But secondly, acid is also the solvent of the Sn^{2+} solution, so that with each addition of Sn^{2+} solution, acid is added to the bath, too. Therefore, besides the addition of acid, which we denoted by x_1, we have to take the amount of Sn^{2+} solution added (x_2) into account when controlling the concentration of acid.

The number of grammes of acid we decide to add to the tin-plating liquid can be written as $d_1x_1 + d_*x_2$. Just before any additions are made, $m_1(V_{\text{old}} - x_7)$ grammes of acid are present in the bath, so that after addition we can write the concentration of the acid as

$$\frac{m_1(V_{\text{old}} - x_7) + d_1x_1 + d_*x_2}{V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7}.$$

The lower limit on the concentration of acid then leads to the following constraint

$$\begin{aligned}(l_1 - d_1)x_1 \;+\; (l_1 - d_*)x_2 \;+\; l_1x_3 \;+\; l_1x_4 \;+\; l_1x_5 \;+ \\ +\; l_1x_6 \;+\; (m_1 - l_1)x_7 \;\leq\; (m_1 - l_1)V_{\text{old}},\end{aligned}$$

whereas the upper limit can be expressed as

$$\begin{aligned}(d_1 - u_1)x_1 + (d_* - u_1)x_2 - u_1x_3 - u_1x_4 - u_1x_5 - u_1x_6 + \\ +\; (u_1 - m_1)x_7 \;\leq\; (u_1 - m_1)V_{\text{old}}.\end{aligned}$$

The constraint associated with the lower limit of formalin differs from other constraints due to the fact that no measurement device is available to determine m_5, the concentration of formalin. In practice, the concentration of formalin is estimated at its lower limit and enough formalin is added to ensure that its concentration does not fall below the lower limit if additions of the other components are made.

If the concentration of formalin were observable we would have had the constraint

$$l_5x_1 + l_5x_2 + l_5x_3 + l_5x_4 + (l_5 - d_5)x_5 + l_5x_6 + (m_5 - l_5)x_7 \leq (m_5 - l_5)V_{\text{old}},$$

but by setting $m_5 = l_5$, this reduces to

$$l_5x_1 + l_5x_2 + l_5x_3 + l_5x_4 + (l_5 - d_5)x_5 + l_5x_6 \leq 0.$$

Note that this constraint is independent of $V_{\text{old}} - x_7$, the volume after draining off and before adding. It only depends on the volume of the additions. If the unknown concentration of formalin satisfies $m_5 > l_5$, this constraint is more restrictive than the constraint we would have had if m_5 was available, and less restrictive if $m_5 < l_5$. For this reason, combined with the fact that formalin evaporates quickly, the process engineers prescribed a minimal addition of 0.5 liters of formalin. Hence, we include the following constraint in our model

$$x_5 > 0.5.$$

To complete the model, a few more constraints must be added. The first is that it is not possible to drain off more tin-plating liquid than the old volume of the bath:

$$x_7 \leq V_{\text{old}}.$$

And finally, we cannot add or drain off negative amounts. That is, we add the following nonnegativity constraints

$$x_i \geq 0 \qquad \text{for } i = 1, 2, \cdots, 7.$$

We have formulated the replenishment problem as a problem of minimization of a linear objective function, subject to linear constraints. Summarized, the complete model is:

$$
\begin{array}{lrrrrrrrll}
\min & c_1x_1 & + c_2x_2 & + c_3x_3 & + c_4x_4 & + c_5x_5 + & c_6x_6 & + c_7x_7 & & \\
\\
\text{s.t.} & - x_1 & - x_2 & - x_3 & - x_4 & - x_5 & - x_6 & + x_7 & \leq & V_{\text{old}} - b_{\min} \\
& x_1 & + x_2 & + x_3 & + x_4 & + x_5 & + x_6 & - x_7 & \leq & b_{\max} - V_{\text{old}} \\
\\
& (l_1 - d_1)x_1 + & (l_1 - d_*)x_2 & + l_1x_3 & + l_1x_4 & + l_1x_5 + & l_1x_6 + & (m_1 - l_1)x_7 & \leq & (m_1 - l_1)V_{\text{old}} \\
& (d_1 - u_1)x_1 + & (d_* - u_1)x_2 & - u_1x_3 & - u_1x_4 & - u_1x_5 - & u_1x_6 + & (u_1 - m_1)x_7 & \leq & (u_1 - m_1)V_{\text{old}} \\
\\
& l_2x_1 + & (l_2 - d_2)x_2 & + l_2x_3 & + l_2x_4 & + l_2x_5 + & l_2x_6 + & (m_2 - l_2)x_7 & \leq & (m_2 - l_2)V_{\text{old}} \\
& - u_2x_1 + & (d_2 - u_2)x_2 & - u_2x_3 & - u_2x_4 & - u_2x_5 - & u_2x_6 + & (u_2 - m_2)x_7 & \leq & (u_2 - m_2)V_{\text{old}} \\
\\
& l_3x_1 & + l_3x_2 + & (l_3 - d_3)x_3 & + l_3x_4 & + l_3x_5 + & l_3x_6 + & (m_3 - l_3)x_7 & \leq & (m_3 - l_3)V_{\text{old}} \\
& - u_3x_1 & - u_3x_2 + & (d_3 - u_3)x_3 & - u_3x_4 & - u_3x_5 - & u_3x_6 + & (u_3 - m_3)x_7 & \leq & (u_3 - m_3)V_{\text{old}} \\
\\
& l_4x_1 & + l_4x_2 & + l_4x_3 + & (l_4 - d_4)x_4 & + l_4x_5 + & l_4x_6 + & (m_4 - l_4)x_7 & \leq & (m_4 - l_4)V_{\text{old}} \\
\\
& l_5x_1 & + l_5x_2 & + l_5x_3 & + l_5x_4 + & (l_5 - d_5)x_5 + & l_5x_6 & & \leq & 0 \\
\\
& & & & & - x_5 & & & \leq & -0.5 \\
\\
& & & & & & & x_7 & \leq & V_{\text{old}}
\end{array}
$$

$$x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7 \geq 0.$$

In this form, the replenishment problem is a standard *Linear Programming* (LP) problem, that can be solved by using for example the *Simplex Algorithm*. The Simplex Algorithm is described in every textbook on Operations Research (see e.g. Sierksma (1996)). An optimal solution of the problem is a set of values for the decision variables $x_1, x_2, \cdots, x_7$, such that the costs of replenishment are minimal while all restrictions are fulfilled.

8.2.4 A starting point for the Simplex Algorithm

Each time additions need to be computed, the LP model may differ from previous models because of different results of the analysis $(m_1, \cdots, m_5)$ and a different volume of the bath (V_{old}). Hence, assuming that the lower and upper limits and the concentrations of the additions remain unchanged, the model has six parameters.

If the model is solved by means of the Simplex Algorithm, a starting point (a so-called *basic feasible* solution) is needed. Generating a basic feasible solution can be done for example by means of the Big M method, or by the first phase of the Two Phase Simplex Method (see e.g. Sierksma (1996)). However, some computing efficiency could be gained if we had a starting point for the Simplex Algorithm that is independent of the measurements. We could just plug in our 'standard' basic feasible solution, and start the Simplex Algorithm right away, whatever the measurements are.

Such a measurement independent basic feasible solution exists if it is possible to create a new bath out of the separate components. This becomes clear if we realize that this corresponds to draining off the whole bath, i.e. setting x_7 equal to V_{old}, and create a completely new bath that satisfies the requirements. If such a solution exists, it is independent of the measurements since we started with an empty bath. In mathematical terms, this means choosing $x_7 = V_{\text{old}}$. The objective function can then be written as

$$\text{minimize } c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + c_5x_5 + c_6x_6,$$

where we suppressed the term c_7V_{old}, since this is a constant and therefore does not influence the optimal solution of the problem. The set of constraints reduces to

$$
\begin{array}{rrrrrrcl}
c_1x_1 & +c_2x_2 & +c_3x_3 & +c_4x_4 & +c_5x_5+ & c_6x_6 & & \\
 & & & & & & & \\
-x_1 & -x_2 & -x_3 & -x_4 & -x_5 & -x_6 & \leq & -b_{\min} \\
x_1 & +x_2 & +x_3 & +x_4 & +x_5 & +x_6 & \leq & b_{\max} \\
 & & & & & & & \\
(l_1-d_1)x_1+ & (l_1-d_*)x_2 & +l_1x_3 & +l_1x_4 & +l_1x_5+ & l_1x_6 & \leq & 0 \\
(d_1-u_1)x_1+ & (d_*-u_1)x_2 & -u_1x_3 & -u_1x_4 & -u_1x_5- & u_1x_6 & \leq & 0 \\
 & & & & & & & \\
l_2x_1+ & (l_2-d_2)x_2 & +l_2x_3 & +l_2x_4 & +l_2x_5+ & l_2x_6 & \leq & 0 \\
-u_2x_1+ & (d_2-u_2)x_2 & -u_2x_3 & -u_2x_4 & -u_2x_5- & u_2x_6 & \leq & 0 \\
 & & & & & & & \\
l_3x_1 & +l_3x_2+ & (l_3-d_3)x_3 & +l_3x_4 & +l_3x_5+ & l_3x_6 & \leq & 0 \\
-u_3x_1 & -u_3x_2+ & (d_3-u_3)x_3 & -u_3x_4 & -u_3x_5- & u_3x_6 & \leq & 0 \\
 & & & & & & & \\
l_4x_1 & +l_4x_2 & +l_4x_3+ & (l_4-d_4)x_4 & +l_4x_5+ & l_4x_6 & \leq & 0 \\
 & & & & & & & \\
l_5x_1 & +l_5x_2 & +l_5x_3 & +l_5x_4+ & (l_5-d_5)x_5+ & l_5x_6 & \leq & 0 \\
 & & & & & & & \\
 & & & & -x_5 & & \leq & -0.5
\end{array}
$$

$$x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6 \geq 0,$$

Solving this model provides a 'standard' basic feasible solution to the replenishment problem that is independent of the measurements. This solution can be interpreted as the cheapest way to compose an admissible new tin-plating bath out of its separate components.

Note that the inclusion of the trivial constraint $x_7 \leq V_{\text{old}}$ in the original model (which is not relevant for the optimal solution) makes such a solution a *basic* feasible solution.

8.2.5 Comparison of results

In this section we compare some solutions arrived at through the old method with the corresponding solutions of the LP model. To this end, we use four *real* analyses that were handed to us by a responsible technical engineer, whom we asked for a few measurements that are typical for the process. For these analyses, we will compare the output of the method currently used with the LP solution. The method currently used basically computes the additions one by one. In Subsection 8.2.1, we discussed why this approach may result in additions that are too large. Solving the LP model was done with the computer program PCProg. The size of the model is such that computing time is negligible (less than a second on a 486 personal computer).

The analysis of July 10, 1995

In the analysis of the tenth of July, the concentration of acid fell between its upper and lower limits, but the concentrations of Sn^{2+} solution, Pb^{2+} solution, and brightener were too low. The bath volume was 470 liters, only 20 liters over the minimal bath volume, so that a considerable addition was possible before the bath would overflow.

Table 8.2: Comparison of replenishing strategies for July 10, 1995.

component	**old method**	**old method***	**LP solution**
acid	114.6 l	71.4 l	59.16 l
Sn^{2+} solution	135.0 l	84.1 l	74.99 l
Pb^{2+} solution	26.9 l	16.8 l	15.36 l
brightener	27.3 l	17.0 l	16.00 l
formalin	3.5 l	2.2 l	1.67 l
demineralized water	0.0 l	0.0 l	0.00 l
drain off	0.0 l	177.3 l	37.18 l
new bath volume	777.3 l	484.2 l	600.00 l
total costs:	Dfl. 8317.48	Dfl. 6853.01	Dfl. 3088.33

*draining off before making additions allowed

In Table 8.2, the results of the old method and the LP solution are tabulated. Note that there are two columns for the solution of the old method. In the first column, indicated by 'results of old method' the rough results of the old method can be found. From this we see that, since the maximum bath volume $b_{\max} = 600$ liters, the bath would overflow with 177.3 liters. This suggestion is therefore not admissible. Furthermore, since this method does not take draining off into account, the additions are computed on the basis of a high volume bath. Apart from any computing inefficiencies, this is another reason why the additions are so high.

The second column presents the results of running the program with the same initial concentration measurements, but now some tin-plating liquid is allowed to be drained off first. Usually the operators drained off an amount equal to the surplus of the first suggestion (in this case 177.3 liters). When draining off before making additions is allowed, both the amounts of chemicals used and the total costs are reduced. Total costs decrease with approximately 1500 Dutch guilders. However, an even greater cost

reduction is possible if we use the LP-solution, which is tabulated in the last column of Table 8.2. The main part of the cost reduction stems from finding a combination of the decision variables so that the number of liters that is to be drained off is kept at a minimum. As a result, the new bath volume is maximal, while the new bath volume of the old method is close to its minimal value of $b_{\min} = 450$ liters.

The analysis of August 29, 1995

The concentration of Sn^{2+} solution in the tin-plating bath on the 29th of August was a little too high. All other concentrations were acceptable. The bath volume was 575 liters, so that only small additions would be allowed. Fortunately, the bath had to be thinned only a little bit. In Table 8.3 we present the results of the two strategies. Note that in this table only one column is presented for the output of the old method since draining off does not appear to be necessary.

Table 8.3: Comparison of replenishing strategies for August 29, 1995.

component	**old method**	**LP solution**
acid	0.0 l	0.00 liters
Sn^{2+} solution	0.0 l	0.00 liters
Pb^{2+} solution	2.2 l	0.00 liters
brightener	4.4 l	0.00 liters
formalin	0.4 l	0.50 liters
demineralized water	6.0 l	2.90 liters
drain off	0.0 l	0.00 liters
new bath volume	588.0 l	578.40 liters
total costs:	Dfl. 111.15	Dfl. 0.69

In an absolute sense, the differences between the methods are less spectacular than in the previous case, but the cost reduction is enormous if expressed in percentages. The total costs of the LP solution are only **0.6%** of the total costs of the old method. The LP solution adds the obligatory 0.5 liters of formalin, and some demineralized water. The old method suggests to add more water, together with Pb^{2+} solution and brightener. The last two additions explain the main part of the cost difference between the

methods. Note that the old method apparently has no lower limit on the addition of formalin.

The analysis of September 27, 1995

On the 27th of September, the tin-plating bath met all but one of the requirements: the concentration Sn^{2+} solution was too low. The concentration of Pb^{2+} solution was found exactly at its lower limit, so that with any addition of another component, addition of Pb^{2+} solution would become necessary. The bath volume was 475 liters, so that quite large additions were possible. In Table 8.4 the results of the two methods can be found. Since the output of the old method initially results in an overflowing bath, again a second column is presented in which there is allowance for draining off the surplus of the first solution before computing additions. The last column contains the solution of the LP model.

Table 8.4: Comparison of replenishing strategies for Sept. 27, 1995.

component	**old method**	**old method***	**LP solution**
acid	67.2 l	61.0 l	32.16 l
Sn^{2+} solution	80.0 l	72.5 l	43.82 l
Pb^{2+} solution	15.0 l	13.6 l	6.76 l
brightener	4.9 l	4.5 l	0.00 l
formalin	2.0 l	1.8 l	0.84 l
demineralized water	0.0 l	0.0 l	0.00 l
drain off	0.0 l	44.1 l	0.00 l
new bath volume	644.1 l	584.2 l	558.60 l
total costs:	Dfl. 3149.17	Dfl. 2961.59	Dfl. 1032.04

*draining off before making additions allowed

The LP solution is approximately 2000 Dutch guilders cheaper than the solution of the old method. The cost reduction is caused by a reduction in the use of chemicals. The effect of this is twofold; not only do we have a reduction in the costs of the additions, it also becomes unnecessary to drain off tin-plating liquid, which is very expensive.

The analysis of October 19, 1995

The analysis of the tin-plating bath of the final example indicated that the concentration of Sn^{2+} solution was too low, while the concentration Pb^{2+} solution was too high. The other concentrations fell within their limits. The bath volume was 575 liters, so that there was not much room for making additions. According to Table 8.5 both methods do not appear to be able to find additions such that draining off can be avoided.

Table 8.5: Comparison of replenishing strategies for October 19, 1995.

component	**old method**	**old method***	**LP solution**
acid	24.6 l	23.1 l	12.99 l
Sn^{2+} solution	33.9 l	31.9 l	20.44 l
Pb^{2+} solution	0.0 l	0.0 l	0.00 l
brightener	0.0 l	0.0 l	0.00 l
formalin	0.0 l	0.0 l	0.50 l
demineralized water	0.9 l	0.9 l	2.69 l
drain off	0.0 l	36.4 l	11.62 l
new bath volume	636.4 l	596.5 l	600.00 l
total costs:	Dfl. 1675.07	Dfl. 1628.83	Dfl. 743.03

*draining off before making additions allowed

In this case too, the LP solution results in a considerable cost reduction of replenishment.

8.2.6 Conclusions

In this section, we discussed an alternative way of computing additions for replenishing a chemical bath that needs regular adjustment. To illustrate what cost reductions were possible in practice, we compared the replenishments that were computed by the method currently used to the corresponding LP solutions, for four typical conditions of the bath. The total costs of the method currently used were 13,252.87 guilders if no draining off before adding was allowed, and 11,558.58 guilders if draining off was allowed. In contrast, the total costs of the LP solution were 4,864.09 guilders for these four days. This means that a cost reduction of 63% and 58%, respectively is possible! Philips Stadskanaal has several tin-plating baths, which are

replenished on a regular basis.

The methodology described in this section is easily adjusted to accommodate similar replenishment problems. When all the components are separately available, the corresponding constraints follow immediately. However, if some of the components can only be added by adding a mix of components, as was the case with the Sn^{2+} solution, the constraints can easily be adapted to deal with this situation. If some of the components are separately available, but also in some pre-mixed form (which may be cheaper), the model can consider all of these as possible additions, and choose the cheapest combination.

Depending on the number of components, the LP-model is typically small. With the aid of modern computer equipment such a model is solvable within a second. However, it is possible to gain a little computation time if a standard starting solution for the Simplex Algorithm is used.

We would like to emphasize that this LP-model is only a tool for computing the cheapest way to create a bath that satisfies certain predefined limits. The cost reductions stem only from a more efficient method of computing the additions, not from a radical change in the replenishment strategy.

However, as the quality improvement project progressed, we obtained further insight in the process by studying measurements that were taken from the production process. Based on the observed data, our proposal was essentially to part with the daily replenishments. The arguments for this piece of advice are described in the remainder of this chapter. The PAT decided to experiment with this drastic change in the replenishment strategy, and the results were surprisingly good. As a result, the LP method, so far, has not been used on the production floor of the department where the SMID's are produced. However, in departments where other types of diodes are manufactured the method has been successfully implemented.

8.3 Adjusting the tin-plating bath

In Chapter 2 we referred to Nolan and Provost (1990)) for the following definition of a process: "*a set of causes and conditions that repeatedly come together to transform inputs into outcomes*", see Figure 8.3.

For the process under consideration, one of the inputs is the composition of the tin-plating bath. Relevant outputs are the thickness and the composition of the tin/lead layer. A lot of effort is put into maintaining a stable tin-plating bath with the objective of reducing variation in the

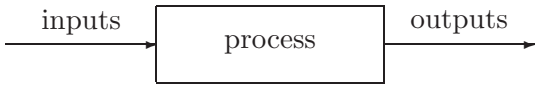


Figure 8.3: Block diagram of a process.

outputs of the process. In this section, we will investigate if such efforts can be justified by studying process data. In Subsection 8.3.1, we describe the data that is available. In Subsection 8.3.2, we investigate whether inputs data has a relation with data on the output side of the process. The measurement error of the observations of the process is discussed in Subsection 8.3.3. In Subsection 8.3.4, it is argued why the presumed relationships between inputs and outputs are not reflected in the data. The conclusion of this investigation and its implications for running the process are discussed in Subsection 8.3.5.

8.3.1 The data

Two kinds of measurements on the tin-plating process are available. In addition to measurements of the inputs of the tin-plating process, we have output measurements of the products at the end of the tin-plating process at our disposal.

The input measurements consist of a chemical analysis of the tin-plating bath. By means of titration the concentrations of acid, Sn^{2+} and Pb^{2+} are measured. The concentration of brightener can only be roughly determined, and there is no measurement device available for establishing the concentration of formalin. The analysis is performed once a day by the operators. The results form the input for a computer program that computes the necessary additions in order to obtain a bath that is fit for applying a good tin/lead layer.

The limits that are used by this program are *process limits*: during production, the concentration measurements are not supposed to exceed these limits. Hence, a replenishment is only prescribed if the observed concentrations are out of specifications. Consequently, the process was functioning outside its process limits half of the time according to the measurements. Fortunately but surprisingly, the quality of the tin/lead layers did not deteriorate in such situations. This gives good reason to doubt the strategy that was used to control the composition of the bath.

Moreover, the value of the process limits is hard to explain. The supplier of the tin-plating chemicals prescribes a certain ideal composition of

the tin-plating bath. However, in the course of time these limits were adjusted, partly based on experience of the operators, and partly based on superstition.

Furthermore, compared to the variation in the measurements, the limits do not seem to be equally tight for all of the components. This is illustrated by Figures 8.4 and 8.5, where respectively the concentrations of Sn^{2+} solution and Pb^{2+} solution are depicted. The graphs reflect the results from consecutive analyses over a three weeks period.

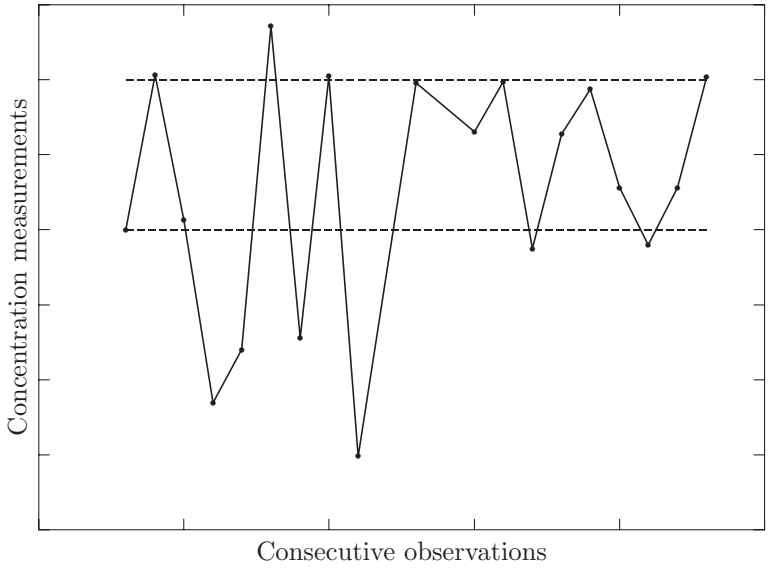


Figure 8.4: The behavior of Sn^{2+} solution in time.

From these figures we see that the process limits for the Sn^{2+} solution are tighter compared to the variation in the measurements than the limits for Pb^{2+} solution. It is to be expected that the concentration of Sn^{2+} solution needs to be adjusted more often than the concentration of Pb^{2+} solution.

Another peculiarity of Figures 8.4 and 8.5 is that the phenomenon of bath exhaustion is not noticeable from these figures. Instead of concentrations jumping up and down more or less randomly, we would have expected a trend or some other form of deterministic behavior (possibly depending on replenishments).

The available output measurements consist of samples of size 10 that are taken from every batch of diodes. A batch may contain up to 31000 diodes and about 60 batches are processed each day in three shifts. From each

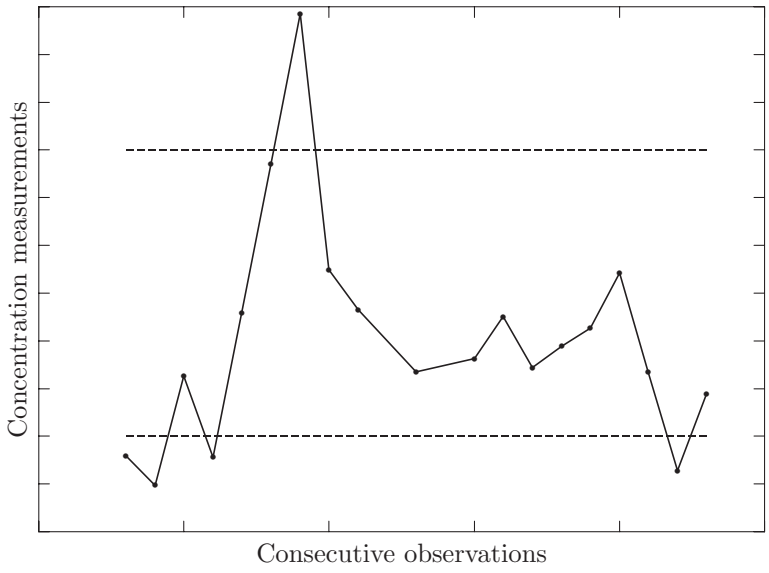


Figure 8.5: The behavior of Pb^{2+} solution in time.

of the sampled diodes, the thickness and the composition of the tin/lead layer are measured. The frequency of sampling and the sample size were set some years ago, and none of the people involved remembered the exact reasons for this sampling strategy. The operators spent considerable time taking these measurements. During the PAT meetings, we first decided to take a sample every other batch, and later on to reduce the sample size to five as well.

A sample size of 10 observations suffices to obtain an impression of the performance of the process. This makes the data fit for *process monitoring*. However, the data were used for *acceptance sampling*. On the basis of a sample of size 10 it was decided to reject or accept a batch of 31000 diodes! The following decision rule was used: if the thickness of the tin/lead layer of one or more of the sampled diodes did not exceed 3μm, or if the sample mean was smaller than 4.85μm, it was decided to tin-plate the whole batch again. Otherwise, the batch was accepted and passed on to the inspection department where the following production step takes place. This *sampling plan* is not sufficient to ensure a high quality level of the outgoing batches. For example, assuming independence and normality of the observations and a (realistic) standard deviation equal to 4, a batch with 5% of the tin/lead layers smaller than 1μm passes this test with a probability of about 25%.

Indeed, if one of these tests rejects the batch, then there are good reasons to suspect that something is wrong. However, to ensure an outgoing quality level of only a few defective parts per million, the sample size must be increased drastically.

In the next subsection, we will investigate whether there are some sensible relationships between inputs and outputs visible in the data.

8.3.2 The relation between input and output measurements

In this subsection, we consider a data set of 106 observations of bath analyses and corresponding product measurements. The observations were successively taken on a daily basis in the last five months of 1995.

The bath analyses are approximately equally spaced, and have values for the concentration acid, Sn^{2+} solution, Pb^{2+} solution and brightener. Recall from the previous subsection that the concentration of brightener can only be roughly determined, whereas the concentration of formalin is not measured at all. The volume of the bath, on the other hand, is available. The output measurements are taken from the last produced batch before analyzing and replenishing the bath. A sample of ten diodes is randomly drawn from the batch and the average thickness and the average composition of the tin/lead layers are recorded.

After investigating several (lagged) relationships between inputs and outputs, suggested by the chemical reactions taking place in the bath, we reached the conclusion that none of the presumed relationships was reflected in the data. To be more precise, we were not able to find any significant relation between input and output measurements that had a practical importance.

Note that this might have been expected from Figures 8.4 and 8.5, where we observed that the bath was frequently operating outside its process limits, while the tin/lead layers were satisfactory.

As an illustration of the lack of correlation, we depicted in Figure 8.6 a scatter plot of the ratio of the Sn^{2+} /Pb^{2+} concentrations in the bath, against the Sn^{2+} /Pb^{2+} ratios of the product measurements. Contrary to everyone's expectations, no strong correlation is present.

From the arguments above, we realized that attempting to control the quality characteristics of tin/lead layers on the basis of the bath measurements is asking for trouble. This conclusion meant a breakthrough in the attempts of the PAT to improve the quality of the tin/lead layers, since it opened the way for other means of controlling the output of the process.

However, before developing a new control mechanism, it had to be understood what caused the lack of correlation between inputs and outputs. To this end the accuracy of the measurements was evaluated first. The results are summarized in the following subsection. Secondly, chemical experts were invited to explain why graphs similar to Figure 8.6 did not show some expected relationship.

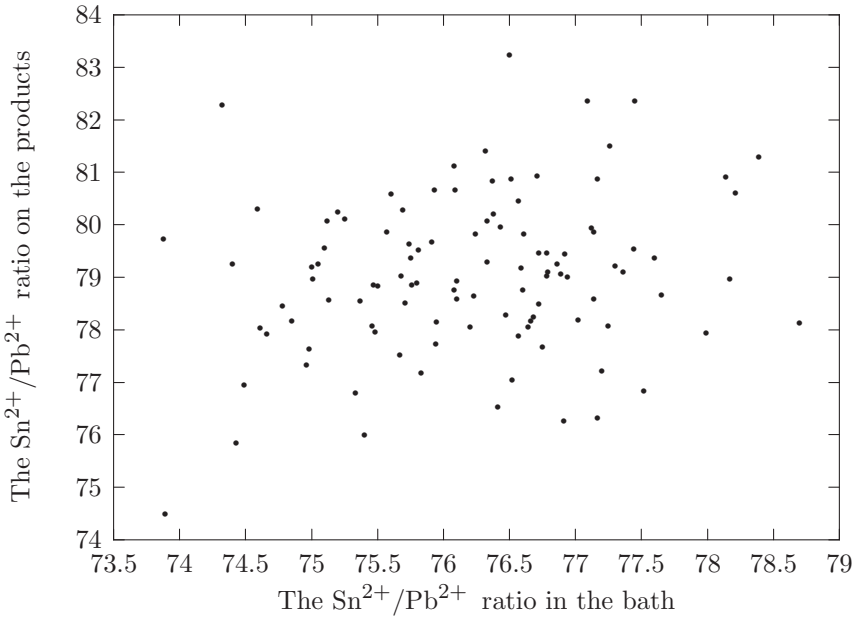


Figure 8.6: The relation between Sn^{2+}/Pb^{2+} in the bath and on the product.

8.3.3 Measurement errors

Based on the methodology described in Does, Roes and Trip (1999), or Banens et al. (1994), an R&R (Repeatability & Reproducibility) study for the product measurements had already been performed. It turned out that the 'gage R&R' was 1.120 for the thickness of the tin/lead layers, while the 'gage R&R' was 2.346 for the concentration measurements. The 'gage R&R' is computed as 5.15×(total measurement variance), and under the assumption of normally distributed observations it is the width of a 99% confidence interval. To judge the accuracy of the measurement device, the 'gage R&R' is compared to the width of the specification limits. In our

case, the product measurements were sufficiently accurate.

Considering the bath measurements, we reached another conclusion. To determine the concentrations of the bath components, operators were asked to take exactly 1 ml of bath liquid using a pipette, whereafter an automated titration device determined how much (in mol) of each component was present in the liquid. The results of these measurements were imported into an online computer, which divided these numbers by 1 ml and presented the outcomes as concentration measurements. Clearly, the accuracy of this procedure is largely determined by the ability of the operators to pipette exactly 1 ml of bath liquid. To investigate this, operators were asked to pipette at least five times 1 ml of demineralized water. To determine the pipetted quantity, it was weighted on a very accurate pair of scales. The results for 21 operators are depicted in Figure 8.7.

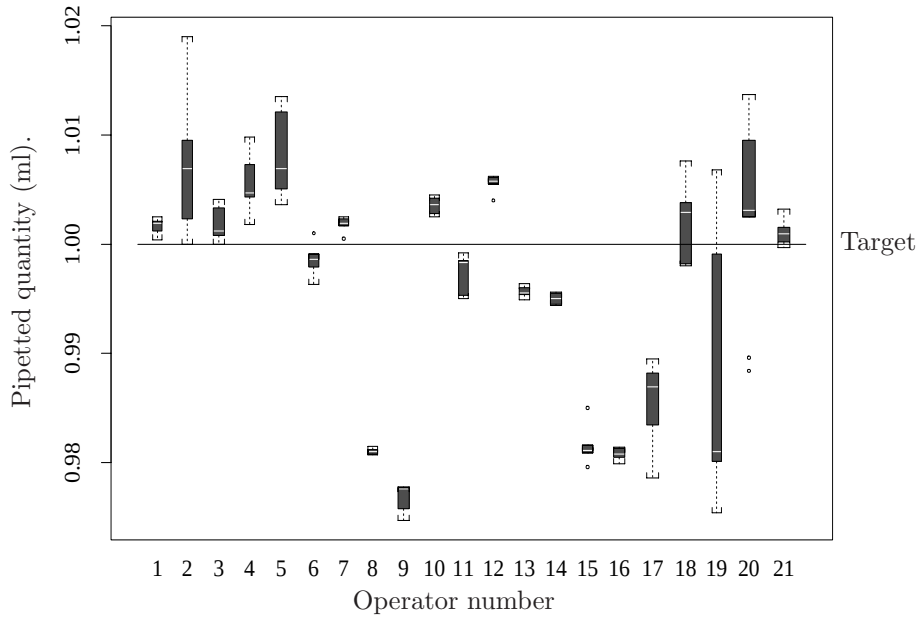


Figure 8.7: Box plots of pipetting results.

In the Box plots of Figure 8.7 the median of the observations per operator is indicated by a white line in the box. The length of the boxes indicates the interquartile range, and the connected hooks (if present) show the minimum and the maximum values. Observations falling outside the range of 1.5×(interquartile range) are considered as outliers, and are plotted as disconnected dots. The width of the boxes is proportional to the square root

of the number of observations per operator. The target value of 1 ml is indicated by a solid line.

Figure 8.7 shows that there is considerable variation between operators. Often, the target value of 1 ml is not even in the range of the observations. Some operators are able to take the amount of water with high precision, but in most of these cases the level is off-target. In other cases the range is quite large, which is not always the result of a larger number of observations.

These observations were taken in a laboratory at the end of an operator course. It is therefore to be expected that the results on the work floor will be even less accurate. Furthermore, Figure 8.7 only shows the variation due to pipetting. There is also additional variation, for example variation due to measurement error of the titration device. Hence, it is to be expected that the total variation in the analysis measurements is even higher than the variation shown in Figure 8.7. However, if we use the results of this experiment, and assume that not water but bath liquid was pipetted, and that the true concentration values were all exactly on target (i.e. in the middle of specification limits), then we are able to compute what the effect of the pipetting inaccuracy would have been on the analysis results. For the concentration of Sn^{2+} , this is depicted in Figure 8.8.

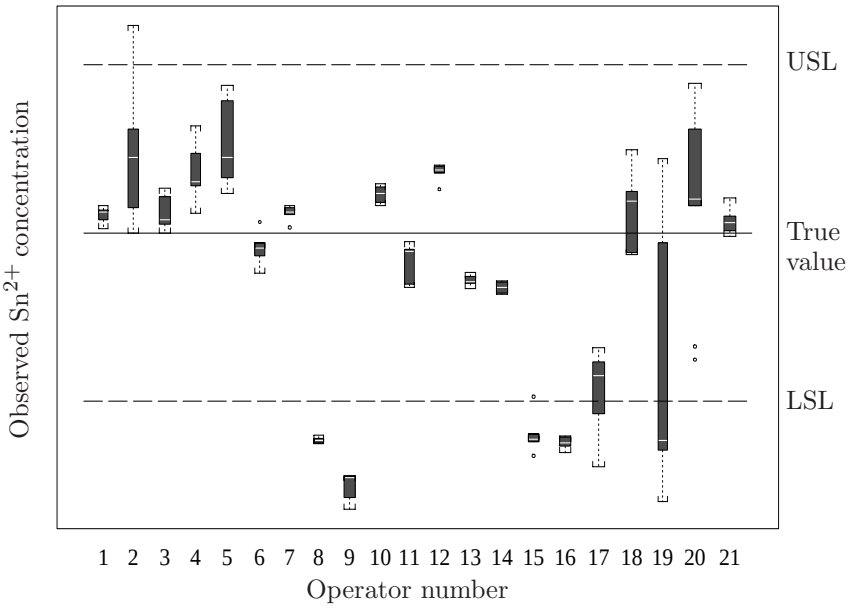


Figure 8.8: Box plots of Sn^{2+} measurement inaccuracies.

From Figure 8.8 it may be concluded that the measurement error due to pipetting is unacceptably high. The analysis results of a bath, the true concentration values of which are exactly in the middle of the specification limits (an ideal bath), can lead to the absurd conclusion that the bath is not fit for production. The analyses tell us more about the large measurement error than about the contents of the bath. The conclusion is that controlling on the basis of analysis results essentially comes down to controlling on the basis of measurement error. Such control actions generally increase the variation in the outcomes of the process. This type of overacting was called *tampering with the process* by Deming.

8.3.4 Process mechanics

In our search for an explanation for the lack of correlation between observed process inputs and process outputs, we also discovered that the assumptions concerning the manner in which the inputs affect the outputs were not entirely correct. Since such assumptions form the basis for any control strategy, this gives another explanation for the malfunctioning of the current replenishment strategy.

Recall from Section 8.1.4 that the tin-plating bath consists of five components: acid, an Sn^{2+} solution, a Pb^{2+} solution, brightener, and formalin. The current strategy calls for action if one or more of the concentrations of Sn^{2+} solution, Pb^{2+} solution, or acid is out of range. These three parameters are thought to have a great impact on the composition and the thickness of the tin/lead layers. The concentration of brightener is only very roughly determined, whereas determining the concentration of formalin is not possible at all. Hence, controlling basically takes place on the first three parameters. However, in brainstorming sessions together with chemical engineers, it turned out that the thickness and the composition of the layer were mainly determined by the two other parameters. Brightener not only takes care of a shiny tin/lead layer, it also influences the thickness of the layer. The concentration of formalin is an important determinant of the composition of the tin/lead layers.

The other components are not unimportant: the acid takes care of a good conductivity of the bath, while the concentrations of the Sn^{2+} and Pb^{2+} solutions must be large enough to ensure that the chemical reactions as described in Section 8.1.4 can take place. However, none of these does directly influence the characteristics of the tin/lead layer on the product.

Furthermore, it was assumed that due to the primary chemical reactions in production, the concentrations of Sn^{2+} and Pb^{2+} solution would

change. If we take another look at the chemical reactions, this does not make sense. For every Sn^{2+} or Pb^{2+} ion that precipitates on a diode, another one dissolves from the anode (see Figure 8.2). So, apart from secondary chemical reactions, there is no reason to assume that the concentration of Sn^{2+} solution or the concentration of Pb^{2+} solution will change with the volume of production.

8.3.5 Conclusion

From the discussion in this chapter, it is easy do deduce why the existing control strategy did not lead to the desired results. Firstly, control actions took place on the basis of relatively unimportant process parameters. And secondly, the measurements of these parameters are highly inaccurate, so that the control actions are not based on process information, but rather on 'noise'. As a result, the effort that was put into stabilizing the bath, caused the bath to destabilize. This type of overacting is called *tampering*, and is a well-known phenomenon in the field of statistical quality control. It is one of the main arguments for using SPC techniques. In the next section we will see that it is possible to obtain a more stable process by simply not reacting on input measurements. A control strategy is proposed that is based on leaving the process alone, except in situations where there is statistical evidence of the presence of an additional, abnormal source of variation.

8.4 A new control strategy

In the previous section we evaluated the control strategy that was used up until now. We reached the conclusion that this strategy, due to *tampering*, has a destabilizing effect on the bath. Furthermore, we argued that from the chemical reactions discussed in Section 8.1.4, it is not clear why replenishments are necessary at all.

Therefore, we experimented with producing without replenishments (except for formalin for reasons of quick evaporation, and brightener). Not surprisingly, the variation in the measurements of the tin/lead layers decreased. During the period of not replenishing, we closely monitored the product measurements to see whether this strategy had the expected effect. Details can be found in Sections 8.4.1 and 8.4.2. After four weeks of applying the new strategy, it was observed that one of the quality parameters, the composition of the tin/lead layer, was drifting away. From a reliable analysis of the bath, performed by the laboratory, the level of acid and

Sn^{2+} solution turned out to be too low for a stable process. Very likely, this was due to secondary chemical reactions in the bath.

Based on the experiment described above, we dropped the daily analysis by the operators and devised a new replenishment strategy, based on *not* replenishing except in situations where there is statistical evidence of something abnormal affecting the process (and except for daily formalin and brightener additions). However, the experiment showed that four weeks without replenishing was too long.

Once a week the bath is filtered to clean the tin-plating liquid from precipitations, caused by secondary reactions. This was a natural moment for a weekly reliable bath analysis, performed by a laboratory employee. On the basis of this analysis, appropriate replenishments can be determined with the aid of the LP model, discussed in section 8.2, to make sure that the concentrations are within their limits.

During the week, the product measurements are monitored with the aid of a control chart, to check whether there is evidence of an additional, abnormal source of variation. How this is done will be discussed in the following subsections.

8.4.1 Monitoring the process

In Figures 8.9 and 8.10 the product measurements of the period March 4, 1996 through March 31, 1996 are depicted. This was the period in which it was decided to experiment with producing without replenishing.

Figure 8.9 shows the successive means of thickness measurements, and Figure 8.10 shows the successive means of Sn^{2+}/Pb^{2+} ratios. Remember that a sample of size 10 is taken from every batch of approximately 31,000 diodes.

Despite the agreement not to make additions, it can be suspected from both figures that around observation 300 something unusual has happened to the process. Both the layer thickness and composition values seem to be affected by some special cause of variation.

Figure 8.10 also shows some other interesting behavior. The mean composition values are, on average, slowly decreasing. This trend is only interrupted by the same sudden jump upwards around observation 300. We will discuss these phenomena in more detail in Subsection 8.4.4.

Figure 8.11 shows a scatter plot of layer thickness and composition. No strong relationship seems present. The correlation coefficient between thickness and composition measurements can be computed as 0.21. We therefore decided to monitor the process with two univariate control charts.

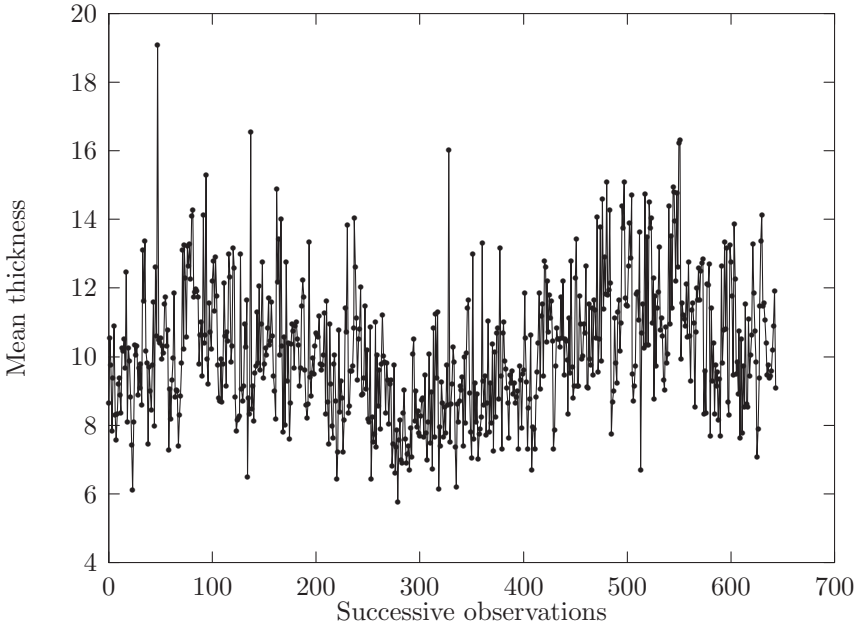


Figure 8.9: Successive mean thickness observations.

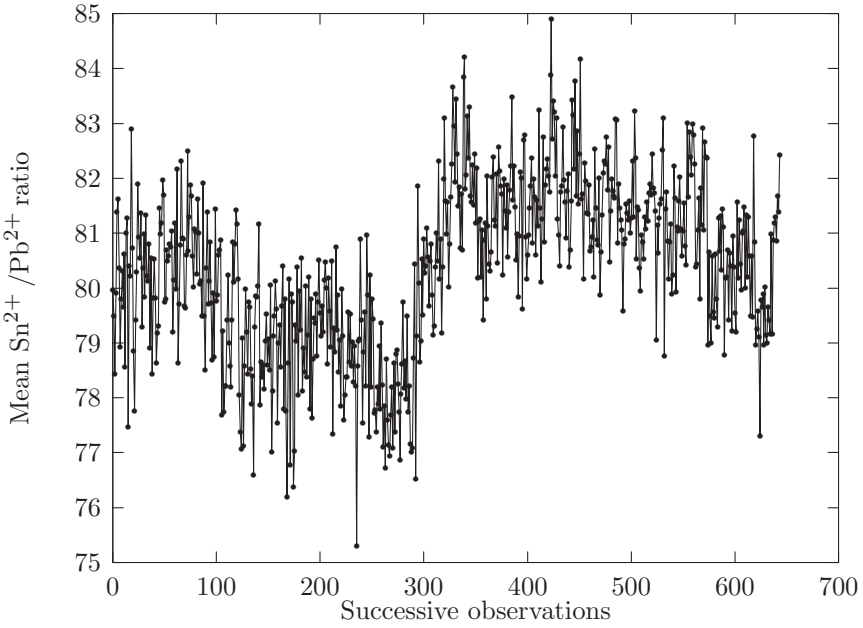


Figure 8.10: Successive mean Sn^{2+} /Pb^{2+} ratio observations.

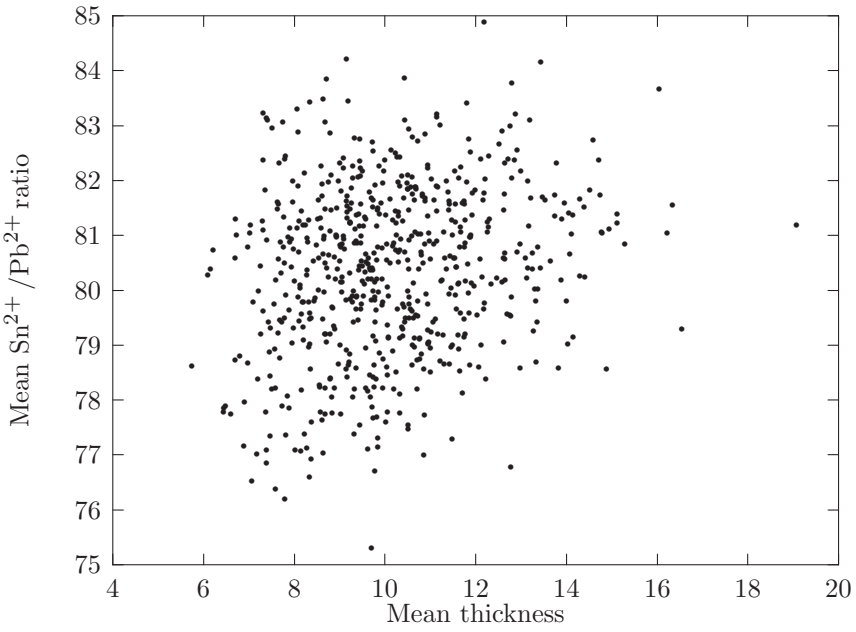


Figure 8.11: Scatter plot of mean Sn^{2+} /Pb^{2+} ratios and mean layer thickness.

In Chapter 1, it was discussed that the application of standard control charts hinges on two important assumptions: normality and independence of successive observations. Since serial correlation distorts various tools for checking normality of the data (such as normal probability plots), the correlation structure of the data is explored first, and the exploration of normality is deferred to Subsection 8.4.3.

A useful tool to explore the correlation structure of the data is the sample autocorrelation function. Figures 8.12 and 8.13 depict the sample autocorrelation functions for mean layer thickness, and mean layer composition, respectively.

It can be shown (see e.g. Anderson (1971)) that a fixed number of sample autocorrelation coefficients of white noise are asymptotically normally and independently distributed with zero means and standard deviations equal to $1/\sqrt{N}$, where N is the number of observations, in our case equal to 644. This can be used to judge the sample autocorrelation coefficients.

The dashed lines in Figures 8.12 and 8.13 are drawn at $1.96 \times 1/\sqrt{N}$ as an approximate 95% confidence interval for individual sample autocor-

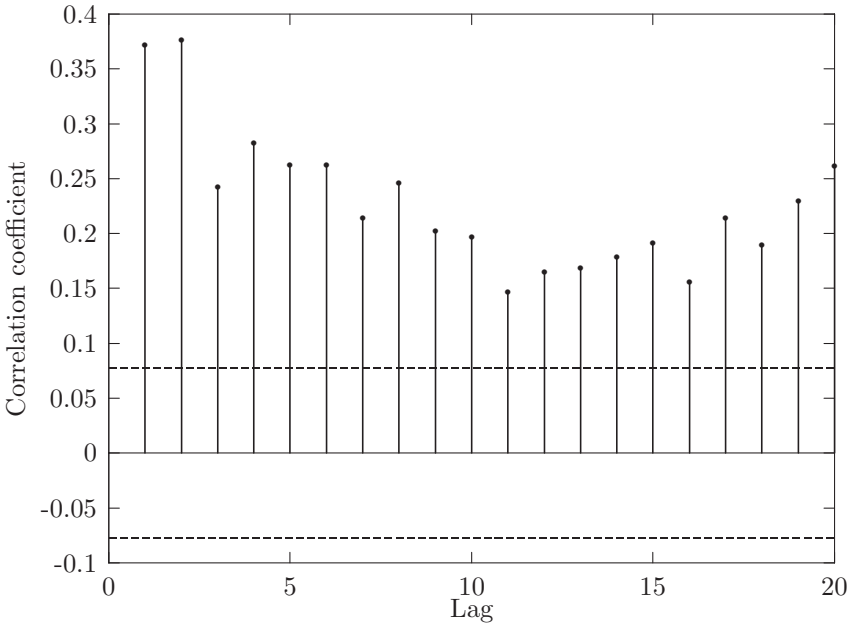


Figure 8.12: Autocorrelation function of mean thickness values.

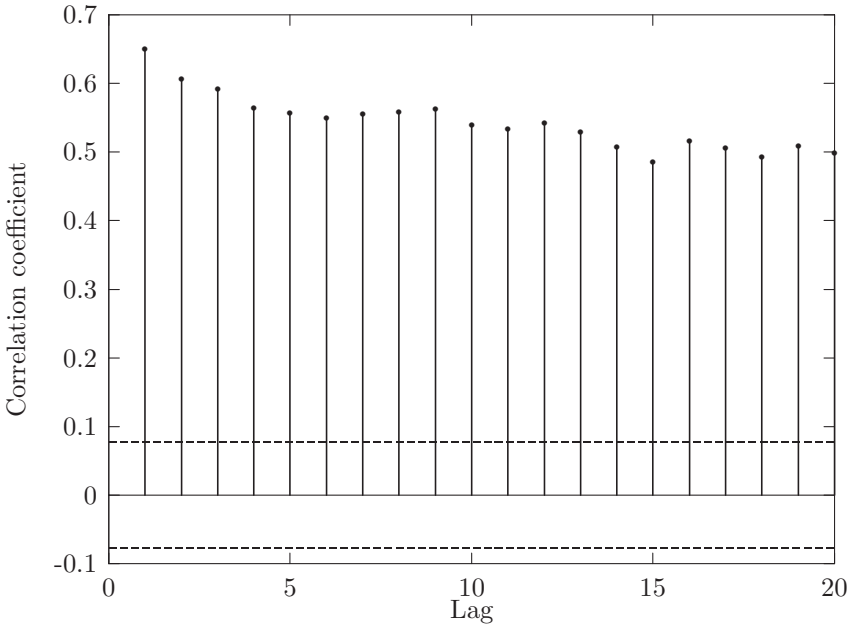


Figure 8.13: Autocorrelation function of mean composition values.

relations with expectation zero. Note that since we are plotting a number of autocorrelation coefficients at once, we expect to find, on average, one out of twenty outside these limits if the observations are uncorrelated. In addition, if the covariance between successive observations is nonzero, the sample autocorrelations are also correlated. These phenomena may seriously distort the interpretation of a sample autocorrelation function, and conclusions based upon this graph must be drawn with care.

Notwithstanding the above remarks, the conclusion from Figures 8.12 and 8.13 is that both layer thickness and layer composition are highly autocorrelated. Moreover, both autocorrelation functions suggest nonstationarity. When constructing control charts, we must take these findings into account. In the next section we will discuss why this is necessary.

8.4.2 Ignoring the serial correlation

First we ignore the serial correlation and set up a control chart for the mean of mean layer composition in the usual way; that is, assuming independence and normality of the observations.

Let us denote the observed mean composition value on time t by $\bar{y}_t$. Remember that means are taken over samples of size $n = 10$. Trial control limits with tail probabilities of $\frac{1}{2}\alpha$ are determined as (see Does and Schriever (1992))

$$\text{LCL} = \bar{\bar{y}} \; + \; \sqrt{\frac{N-1}{Nn}}\, t_{\frac{1}{2}\alpha}(N(n-1)) \; \tilde{s},$$

and

$$\text{UCL} = \bar{\bar{y}} \; + \; \sqrt{\frac{N-1}{Nn}}\, t_{1-\frac{1}{2}\alpha}(N(n-1)) \, \tilde{s},$$

where LCL and UCL stand for *Lower Control Limit* and *Upper Control Limit*, respectively, and $\bar{\bar{y}}$ is the overall mean of $N = 644$ observations of $\bar{y}_t$ $(t = 1, \cdots, 644)$. Furthermore, $t_{\frac{1}{2}\alpha}(N(n-1))$ is the $\frac{1}{2}\alpha$-th percentile of the t^2 distribution with $N(n-1)$ degrees of freedom, and $\tilde{s}$ is defined as

$$\tilde{s} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} s_i^2},$$

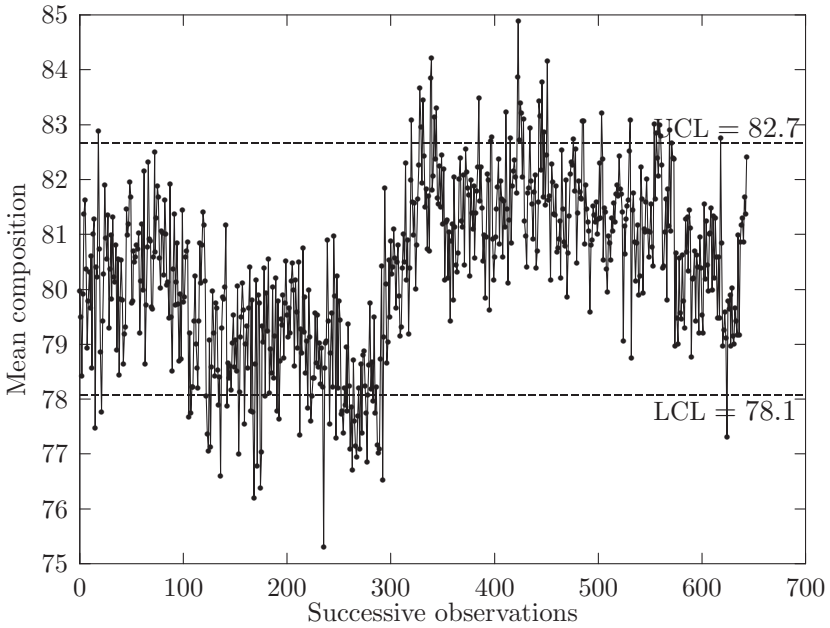


Figure 8.14: Trial control limits for monitoring the mean of composition.

where s_i^2 is the sample variance in sample i. The computed control limits for $\alpha = 0.002$ are depicted in Figure 8.14.

Figure 8.14 shows a lot of out-of-control signals. This can be explained as follows.

In general, a shift in the mean will increase the probability of observing an out-of-control signal. As we will see in the next subsection, an IMA(1,1) model is the most appropriate model for our data in the class of ARIMA(p, d, q) models. A feature of an IMA(1,1) process is that its level is changing constantly. It is therefore to be expected that, in testing for stability of the mean, a lot of out-of-control signals will be generated by such a data set.

Hence, the data must be monitored using a control chart that takes serial correlation into account. Using a modified control chart, where untransformed measurements are compared to control limits which are adjusted to allow for serial correlation, is not a good idea in this case. Therefore, we revert to residuals charts to monitor the process of tin-plating diodes.

In the next subsection, an appropriate time series model is selected and fitted to the data.

8.4.3 Finding an appropriate model for the data

The autocorrelation function of Figure 8.13 shows considerable autocorrelation which seems to decrease slowly and linearly with the length of the lags, an indication for the presence of a unit root. Box and Jenkins (1976) explain why this is so.

Let a sequence of random variables $\{Z_t\}$ follow a stationary ARMA(p,q) model. Assume that $\mathrm{E}(Z_t) = 0$. Then we have that Z_t can be written as

$$Z_t = \phi_1 Z_{t-1} + \cdots + \phi_p Z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_q \varepsilon_{t-q}, \tag{8.1}$$

where $\{\varepsilon_t\}$ is a *white noise* process. When B, the backward shift operator is used, (8.1) can be written as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) Z_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) \varepsilon_t$$

or

$$\phi(B) Z_t = \theta(B) \varepsilon_t,$$

where $\phi(\cdot)$ and $\theta(\cdot)$ are polynomials of degree p and q, respectively.

Multiplying both sides of Equation (8.1) by Z_{t-k} and taking expectations, we find that

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p} \qquad \text{for } k \geq q+1, \tag{8.2}$$

where γ_k is the covariance between Z_t and Z_{t-k}. For $k \geq q+1$ the covariance between Z_t and ε_{t-k} is zero. Dividing both sides of Equation (8.2) by γ_0, the variance of Z_t, this results in

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p} \qquad \text{for } k \geq q+1, \tag{8.3}$$

where ρ_k is the autocorrelation coefficient of Z_t and Z_{t-k}. And hence, the autocorrelation function satisfies the difference equation

$$\phi(B) \rho_k = 0 \qquad \text{for } k \geq q+1. \tag{8.4}$$

Furthermore, if it is assumed that $\phi(B) = \prod_{i=1}^{p} (1 - G_i B)$, with $G_1, \cdots, G_p$ distinct, so that $1/G_1, \cdots, 1/G_p$ are distinct roots of $\phi(\cdot)$, then ρ_k is of the form

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k \qquad \text{for } k \geq q+1.$$

If one or more of the roots of $\phi(\cdot)$ approaches one, say $G_1 = 1 - \nu$ with $\nu > 0$ small, then

$$\rho_k \approx A_1(1 - k\nu),$$

and the autocorrelation function will not die out quickly. Instead, it will be linearly decreasing in k.

On the other hand, for a model to be stationary, it is required that the roots of $\phi(\cdot)$ must lie outside the unit circle, so that $G_1, \cdots, G_p$ must lie inside the unit circle. Therefore, if none of the roots of $\phi(\cdot)$ is close to one, ρ_k will damp out quickly.

Hence, Figure 8.13 shows the typical behavior of a nonstationary process. The next step is to compute first differences of the mean composition values. The autocorrelation function of the new series will give insight in the remaining correlation structure. If taking first differences does not remove nonstationarity, we take first differences once more, and so on, until the series displays stationary behavior. The sample autocorrelation function of first differences of the mean composition values is depicted in Figure 8.15.

The autocorrelation function of Figure 8.15 shows a single spike at lag 1, indicating a *Moving Average* (MA) component of order 1 in the first differences. This can be seen from the theoretical autocorrelation function of a first-order MA process. For a sequence of random variables $\{Z_t'\}$, generated by an MA(1) model, i.e.

$$Z_t' = \varepsilon_t - \theta\varepsilon_{t-1},$$

with $\{\varepsilon_t\}$ *white noise*, we have for the first-order correlation coefficient

$$\rho(1) = \frac{\mathrm{E}(Z_t' Z_{t-1}')}{\mathrm{V}(Z_t')} = \frac{-\theta}{1+\theta^2},$$

whereas $\rho(2) = \rho(3) = \cdots = 0$. Furthermore, a first-order MA process can also be viewed as an infinite-order *AutoRegressive* (AR) process with coefficients that decrease exponentially in absolute value:

$$\begin{aligned} & Z_t' &=& \varepsilon_t - \theta\varepsilon_{t-1} = (1-\theta B)\varepsilon_t \\ \iff\ & (1-\theta B)^{-1} Z_t' &=& \varepsilon_t \\ \iff\ & Z_t' &=& \varepsilon_t - \theta Z_{t-1}' - \theta^2 Z_{t-2}' - \theta^3 Z_{t-3}' + \cdots, \end{aligned}$$

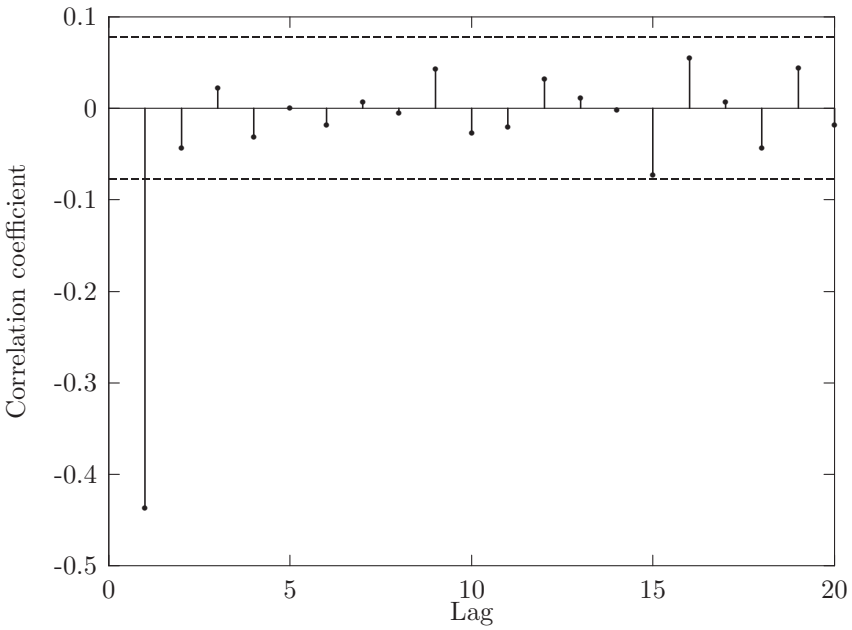


Figure 8.15: Autocorrelation function of differenced mean composition values.

so that the theoretical *partial autocorrelation function* (see for example Harvey (1993)) of a first-order MA process is exponentially declining in absolute value. So, if an MA(1) model is useful for modelling the first differences, we expect to see such behavior in the corresponding partial autocorrelation function. The observed partial autocorrelation function of first differences of the mean composition values of Figure 8.16 does indeed show such behavior.

Hence, based on the behavior of the (partial) autocorrelation functions of the data, we decide to select the IMA(1,1). The maximum likelihood estimate of the moving average coefficient was computed with the aid of SPLUS as $\hat{\theta} = 0.811$, with a standard error of 0.023. The model for the series of mean composition values then becomes

$$\bar{y}_t = \bar{y}_{t-1} + \varepsilon_t - \underset{(0.023)}{0.811}\ \varepsilon_{t-1}. \tag{8.5}$$

When model (8.5) is fitted to the data, the sample autocorrelation function of the residuals, as depicted in Figure 8.17, may be used to check whether there is still some remaining serial correlation present in the residuals.

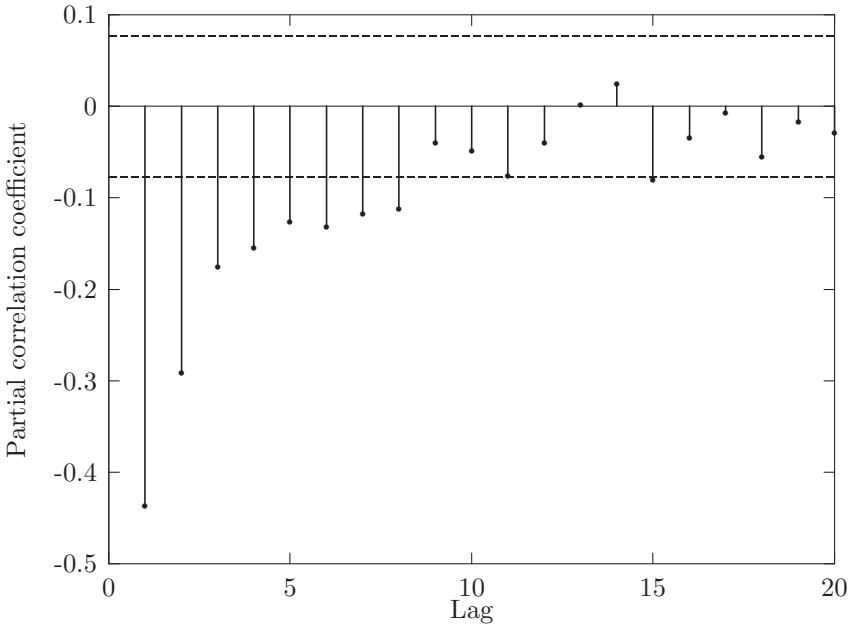


Figure 8.16: Partial autocorrelation function of differenced mean composition values.

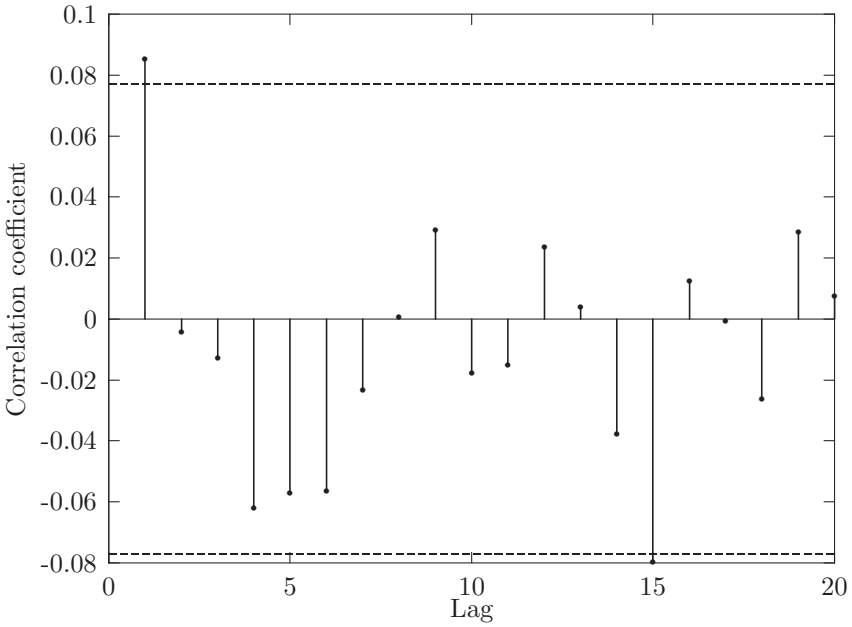


Figure 8.17: Autocorrelation function of residuals.

A warning regarding the interpretation of the residual sample autocorrelation function must be made at this point. Granger and Newbold (1986) cite references from which it follows that the asymptotic standard deviations of sample autocorrelation coefficients of *residuals* may be smaller than $1/\sqrt{N}$. Hence, the sample autocorrelation coefficients of Figure 8.17 need careful interpretation. Nevertheless, the lines drawn in this figure at $1.96 \times 1/\sqrt{N}$ can provide a crude check for model adequacy. These limits are only just exceeded at lag 1 and 15.

However, the autocorrelation coefficients are small, especially when compared to the autocorrelation function of Figure 8.13, so that the correlation is considerably reduced.

Also, evaluation of the well-known (modified) Portmanteau test statistic at various lags does not lead to rejection of on of the hypotheses of zero correlations. Hence, for this moment, we proceed as if the residuals are uncorrelated.

In Subsection 8.4.1, we deferred testing for normality because of the presence of serial correlation. Since the residuals of model (8.5) behave more like an uncorrelated sequence than the original observations, its normal probability plot is more reliable for judging normality than a normal probability plot of the original observations. In Figure 8.18 a normal probability plot of the residuals of model (8.5) is shown.
The normal probability plot does not indicate deviations of normality. Also more formal tests do not reject the hypothesis that the residuals are normally distributed. For example, a test based on observed kurtosis and skewness of the empirical distribution function as described by Harvey (1993), p. 45 results in a p-value of 0.330.

For the remainder of this chapter we assume that model (8.5) is the correct model for our data. The residuals of the fitted model behave more or less as a sequence of uncorrelated normally distributed variates. Hence, both conditions discussed on page 194 are fulfilled, and the standard control charts can be used to monitor the residuals. As we will see, not only the residuals contain information about changes in the process. It is wise to monitor the fitted values as well. In the next subsection we will discuss how the results of the data analysis can provide us with information about changes in the process.

8.4.4 The common cause chart

In the previous subsection, we modelled the mean composition observations using an IMA(1,1) model. The process generates observations with a

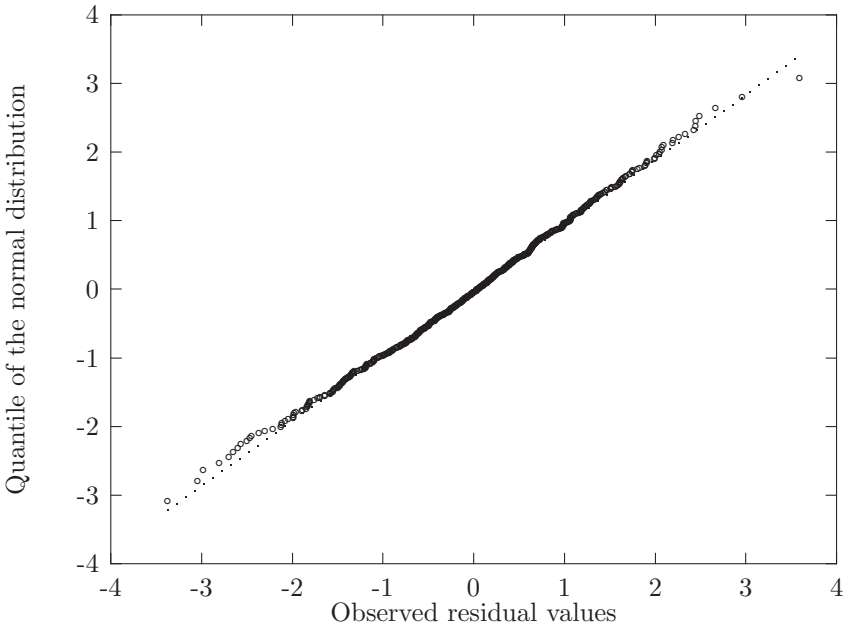


Figure 8.18: Normal probability of residuals.

wandering level, also in cases where only common causes of variation are affecting the process. A graph of the level of the process now contains nontrivial information, viz. how the process is affected by common causes. Alwan and Roberts (1988) call such a graph a *common cause* graph, since it shows the variation due to causes of variation that are inherent in the process. If the process is not allowed to wander too far from a certain target value, some kind of active control must be applied to assure that product measurements do not fall outside specification limits. Let us consider the sequence of fitted values.

For $\hat{\bar{y}}_t$, the optimal one-step ahead forecast of $\bar{y}_t$, we have

$$\begin{aligned}\hat{\bar{y}}_t &= \mathrm{E}(\bar{y}_t|\bar{y}_{t-1}, \bar{y}_{t-2}, \bar{y}_{t-3}, \cdots) \\ &= (1-\theta)\sum_{j=1}^{\infty}\theta^{j-1}\bar{y}_{t-j},\end{aligned}$$

so that the optimal forecast of $\bar{y}_t$ is an *Exponentially Weighted Moving Average* (EWMA) of previous observations. In Chapter 4, we explained that the EWMA can be easily updated in the following way

$$\hat{\bar{y}}_t = \theta \hat{\bar{y}}_{t-1} + (1-\theta)\bar{y}_{t-1}.$$

A plot of the EWMA of the mean composition values is presented in Figure 8.19.

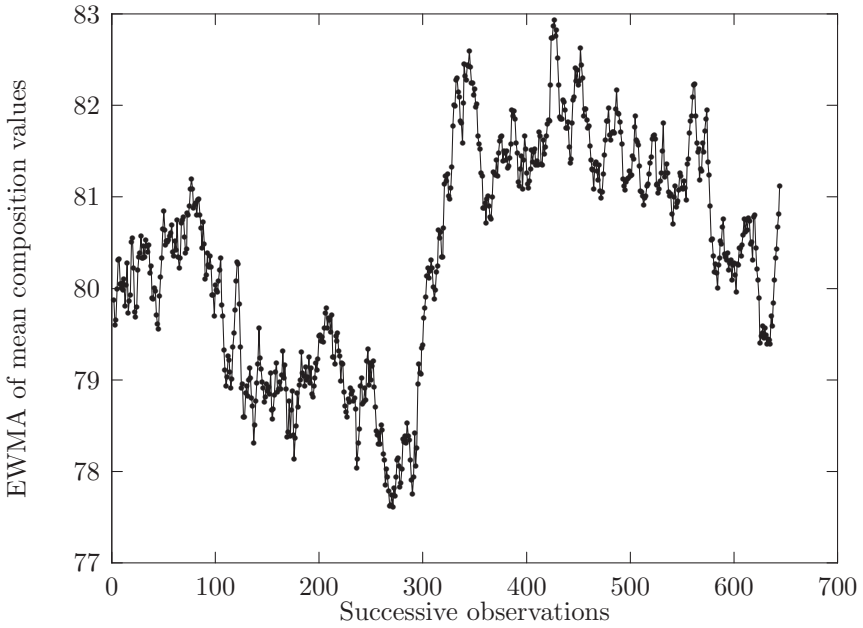


Figure 8.19: EWMA of mean composition values.

Figure 8.19 is a smoothed version of Figure 8.10. With the short-term variation removed, Figure 8.19 shows even more clearly the two phenomena already described in Section 8.4.1. As was remarked there, the first peculiarity that catches the eye is the sudden jump upwards around observation 300. The second one is the slow downward movement on the left and on the right of this jump.

The explanation for the sudden jump upwards was found in the log book that is kept by the operators. In spite of the agreement not to react to bath measurements, low levels of Sn^{2+} solution that were reported by bath analyses around observation 300 worried the operators. Combined with the low level of composition, it was decided to add 100 liters of new tin-plating liquid, 6.5 liters of formalin, and 6 liters of brightener. Since no immediate improvement was observed, four more liters of formalin, twenty more liters of Sn^{2+} solution and 1 more liter of brightener were added the same day. Eventually, there was some reaction which is clearly visible in Figure 8.19, but also can be observed in Figures 8.9 and 8.10.

The foregoing explanation may also give a hint for explaining the slow downward trend in the composition values. Due to secondary chemical reactions, it is possible that dissolved Sn^{2+} ions form Sn^{4+} ions and precipitate. This may cause a slowly decreasing Sn^{2+} concentration in the bath, with a similar effect on the composition observations on the products. However, for the process under consideration it is not possible to obtain accurate measurements of these bath parameters on a regular basis, so that this hypothesis cannot be verified. On the other hand, adding highly concentrated Sn^{2+} solution to the bath breaks the downward trend, and may therefore be considered as an indication for a too low Sn^{2+} concentration. For a complete understanding of the process mechanism, further research into this is needed.

In the database of process measurements, it was the first time that we were able to observe a long-term downward movement of the mean composition values. Previously, such movements were distorted by short-term movements, induced by making additions to the bath. Knowledge of the cause of this trend may prove to be very useful. If it is known how to influence the level of the process, knowledge of such a trend may form the basis of a control strategy. Box and Kramer (1992) have shown that control actions triggered by an EWMA crossing certain boundaries are a cost-efficient way to control IMA(1,1) processes. In the present case, this may, for example, lead to the adding of highly concentrated Sn^{2+} solution to the bath, to compensate for the downward trend, thereby stabilizing the production conditions, and reducing the variation in the product measurements.

8.4.5 The special cause control chart

In the present case, the data show nonstationary behavior. However, as in situations with uncorrelated observations, special causes of variation can influence the output of the process.

Usually, in the uncorrelated case, this is narrowed down to a persistent shock in the mean level, or a persistent shock in the dispersion of the process. Suppose that the mean of the process shifts from μ to $\mu + \delta$ on time $T+1$. The mean of deviations from the (estimated) model then shifts from 0 to δ, which will lead to a higher probability of crossing one of the control limits of a control chart for the mean.

In the case of IMA(1,1) observations, an impulse shock to the level of the process by an amount of δ results in a persistent change in the level of subsequent realizations by the same amount.

As a result, a persistent change in the level of an IMA(1,1) process is

only once encountered in the residuals. Detecting a shock in the level of an IMA(1,1) process by monitoring its residuals is therefore doomed from the start, unless the shock is so large that the control limits are reached in one observation.

This does not surprise us very much, since we chose the IMA(1,1) model for its ability to capture the nonstationarity of the mean of the observations. As a consequence, the model allows the mean of the process to wander. Detecting a change in the mean by looking at the residuals is then asking for trouble.

However, a trend in the observations of an IMA(1,1) process shows up as a persistent change in the level of the residuals. The errors, made in the EWMA forecasts of mean composition values, are depicted in Figure 8.20. The control limits are computed with equal tail probabilities of $\frac{1}{2}\alpha = 0.001$.

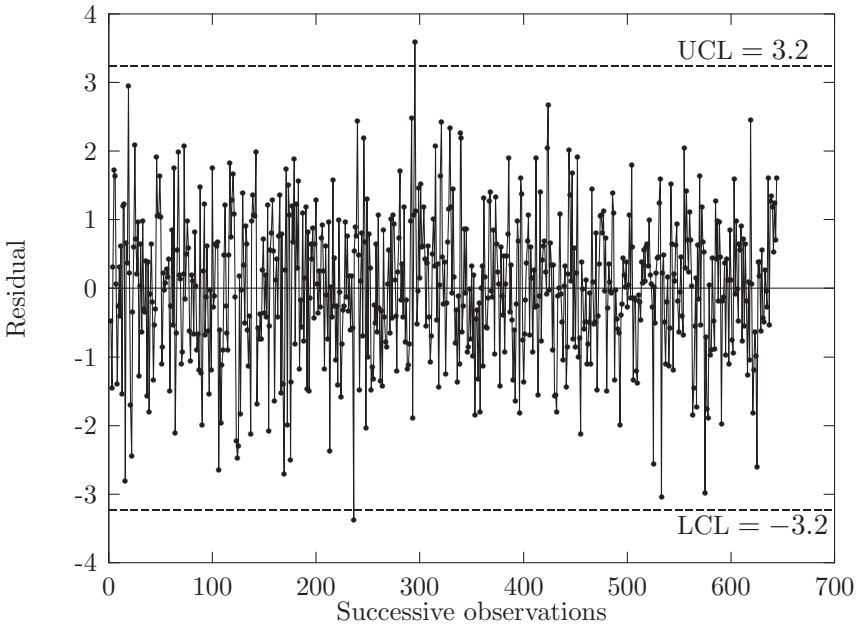


Figure 8.20: Residuals of one step ahead forecasts.

Since we observed two persistent downward trends and one upward trend in Figure 8.10, we would, in the light of the previous arguments, expect to see some changes in the mean level of the residuals of Figure 8.20.

And indeed, there is an out-of-control signal that could be linked to a positive shift in the mean level of the residuals due to the trend around observation 300. Also, an out-of-control signal indicating a possible negative shift in the mean level of the residuals is observed. However, Figure 8.20

does not show convincingly that there is something out of the ordinary happening to the process.

The latter can be explained as follows. If we make rough estimates of the trends from the raw data, we find that for the first downward trend the level drops approximately three units in 300 observations, resulting in a decline of about 0.01 units per observation. The upward trend raises the level about six units in 60 observations, resulting in an increase of about 0.1 units per observation. The last downwards trend can also be roughly estimated as a decline of 0.01 units per observation. Hence, in Figure 8.20, we expect to see a negative shift in the level of the residuals of about 0.01 for the two downwards trends, and a positive shift in the mean of 0.1 due to the upward trend. Since the sample standard deviation of the residuals equals $\hat{\sigma}_\varepsilon = 1.0479$, the change in the level of the residuals induced by the downwards trends is approximately $-0.01\sigma_\varepsilon$, and the change in the level induced by the upward trend equals approximately $0.1\sigma_\varepsilon$. Since these shifts are small compared to the standard error of the residuals, a regular Shewhart control chart is not the proper tool to detect such a shift.

The ARL equals 499.74 if $\delta = -0.01$ and $\alpha = 0.002$, so that, on average, 500 observations are needed before a shift of the size of the slow downward shifts is detected for the first time with a Shewhart chart such as Figure 8.20. The value of ARL(0.1) equals 475.15 if $\alpha = 0.002$ so that detecting a shift of the size of the upward shift takes, on average, about 475 observations. The ARL curve of a Shewhart chart with $\alpha = 0.002$ is depicted in Figure 8.21.

The computations above illustrate the well-known disadvantage of the Shewhart control chart that it is not very efficient in detecting small shifts in the mean. With additional run rules or warning limits its performance for detecting small shifts can be improved (see Does and Schriever (1992)). However, there are alternatives such as the EWMA and the CUSUM control chart. These control charts are more efficient in detecting small shifts in the mean.

The most common way to judge a CUSUM chart is using a *Decision Interval Scheme* (DIS) (see Section 5.1). Both the CUSUM for detecting positive shifts in the mean

$$S_{H_i} = \max\left[0, S_{H_{i-1}} + z_i - k\right]$$

(where z_i is the standardized observation at time i, and k is the reference value) and the CUSUM for detecting negative shifts

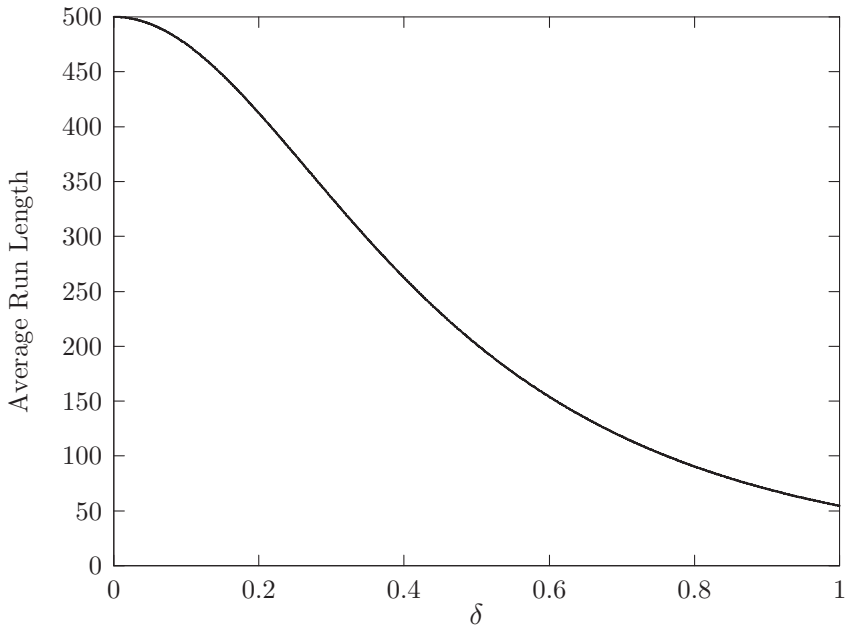


Figure 8.21: ARL curve of a Shewhart control chart ($\alpha = 0.002$).

$$S_{L_i} = \max\left[0, S_{L_{i-1}} - z_i - k\right]$$

are compared to some value of the decision interval h. One of the sums exceeding h is an indication for an out-of-control situation. The values of k and h can be chosen such that the two sided CUSUM is as efficient as possible (in terms of the ARL) in detecting a predescribed shift in the mean, while maintaining a certain value of the in-control ARL. In Figure 8.22 it is illustrated how k can be chosen to design the CUSUM for efficiently detecting a shift of size $0.1\sigma_\varepsilon$, while maintaining an in-control ARL of 500.

For each $k \in [0, 1]$ we computed what value of h is needed to maintain an in-control ARL of 500. Subsequently, for this choice of h and k the ARL for $\delta = 0.1\sigma_\varepsilon$ was computed. In Figure 8.22 these ARL(0.1) values are depicted against k.

The out-of-control ARL is minimized for $k = 0.055$. The corresponding value of h equals 19.025, and ARL(0.1) = 237.7. Hence, this CUSUM chart will, on average, detect a shift of $\delta = 0.1\sigma_\varepsilon$ twice as fast as a Shewhart chart. Note that this value of k agrees reasonably well with the general recommendation to set k equal to half the size of the shift we want to detect. The optimal value of k being slightly larger than half the size of the shift the chart is designed for is in accordance with the results obtained by

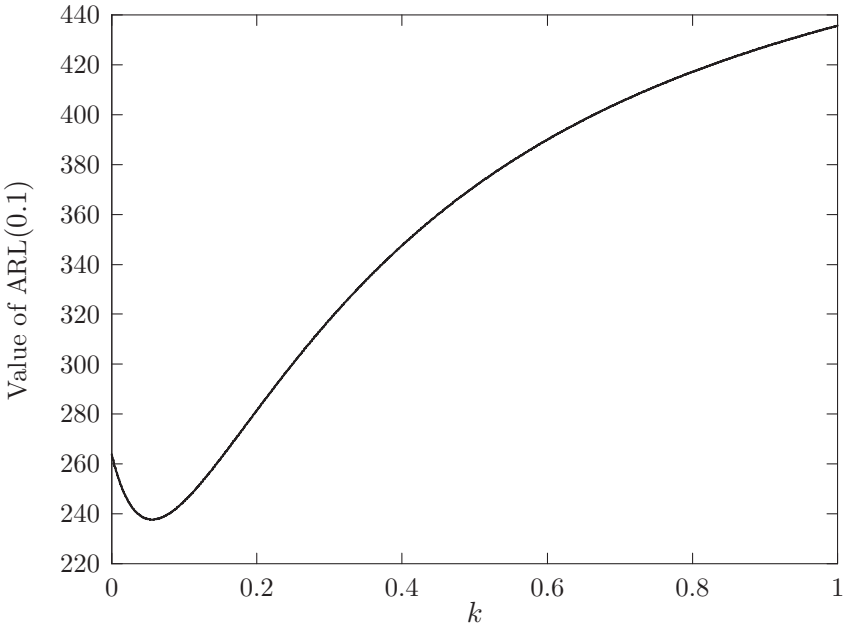


Figure 8.22: ARL(0.1) as a function of k, with h such that ARL(0)=500.

Gan (1991), who noted that this phenomenon occurs for shifts smaller than $0.2\sigma_\varepsilon$. In Figure 8.23, the ARL curve of a CUSUM chart with $k = 0.055$ and $h = 19.025$ is depicted.

The ARL values of the CUSUM are evaluated using the Fredholm-integral approach that is discussed in Chapter 5 and in Appendix B.

Comparing Figure 8.23 with Figure 8.21 shows that the ARL of the CUSUM is also smaller than the ARL of the Shewhart chart for other small δ. However, numerical calculations show that the ARL curves of the two control charts intersect at $\delta = 1.7\sigma_\varepsilon$. The corresponding ARL value is 12.26. For shifts in the mean larger than $1.7\sigma_\varepsilon$ the ARL of the Shewhart chart is smaller than the ARL of the CUSUM chart. The value of $\text{ARL}_{\text{CUSUM}}$-$\text{ARL}_{\text{Shewhart}}$ is maximized for $\delta = 2.8\sigma_\varepsilon$. The ARL of the Shewhart chart then equals 2.6, whereas the ARL of the CUSUM equals 7.5. For larger values of δ this difference reduces to zero since both ARL curves converge to one.

Hence, designing the CUSUM chart in such a way that it is as sensitive as possible for detecting small shifts of size $0.1\sigma_\varepsilon$ in the mean results in smaller sensitivity for larger shifts as compared to the Shewhart chart.

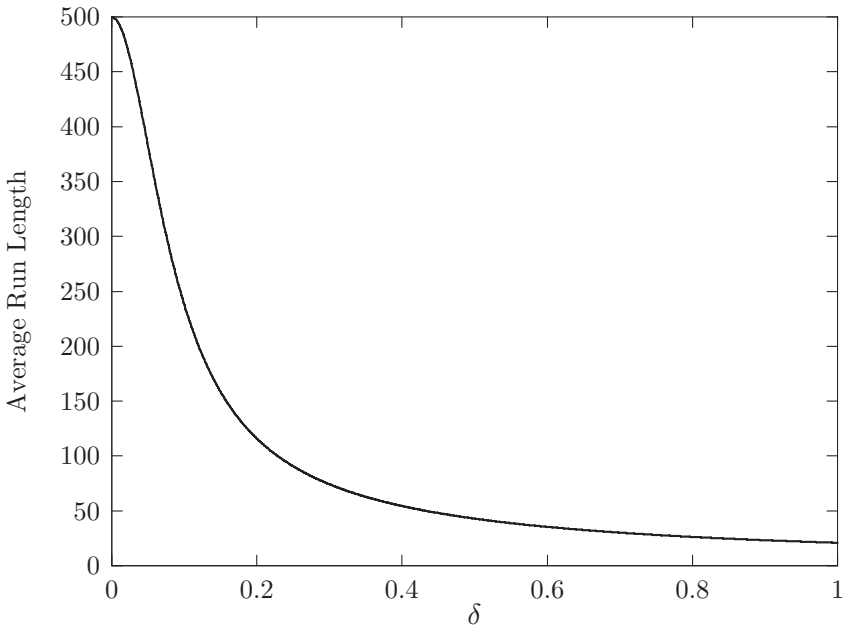


Figure 8.23: ARL curve of a CUSUM control chart ($k = 0.055$ and $h = 19.025$).

The CUSUM chart for the mean composition values is depicted in Figure 8.24.

An out-of-control signal on the high side is observed at observation 331, indicating the upward trend mentioned earlier. This control chart shows more convincingly than the Shewhart chart that the trend in the data distorts the behavior of the one step ahead forecasts.

The previous analysis raises the question of how the model identification process was influenced by the presence of a deterministic trend. To investigate this, the data were detrended and the sample autocorrelation function of the resulting series was studied. The nonstationary behavior still appeared to be present. The estimate of the MA parameter changed from 0.811 to 0.878.

In order to estimate θ and the two trend parameters simultaneously, two dummy variables were added to model (8.5), one for the slow (linear) downward trends, and one for the steeper upward (linear) trend. The dummies both start from zero, and are increased by one for each observation where the trend is supposed to be active.

This resulted in the following fitted model

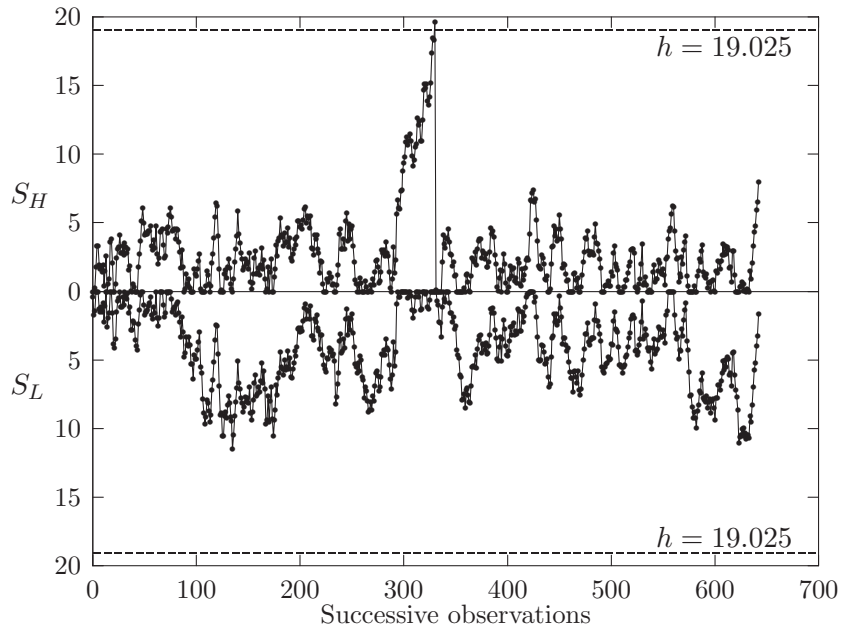


Figure 8.24: CUSUM chart of residuals of one step ahead forecasts.

$$\bar{y}_t = \bar{y}_{t-1} + \varepsilon_t - \underset{(0.021)}{0.854}\ \varepsilon_{t-1} - \underset{(0.007)}{0.010}\ u_t + \underset{(0.024)}{0.079}\ v_t, \tag{8.6}$$

where u_t and v_t are dummies associated with the slow downward trends, and the upward trend, respectively. The value of u_t is increased by one for $t = 1, \cdots, 262$ and for $t = 425, \cdots, 624$. The value of v_t is increased by one for $t = 293, \cdots, 339$.

8.4.6 Implementation and results

During the four weeks of the experiment, the mean thickness and composition observations were closely monitored. In the previous subsection we discussed in detail how this was done for the composition data. The thickness data were monitored roughly in the same way, after making a log transform in order to be able to assume normality. The variability in the measurements was monitored using regular S-charts, since there appeared to be no correlation in subsequent sample standard deviations.

Studying the data as described in this chapter provided valuable information that helped to achieve a better understanding of the process. In practice, however, the procedure is too laborious to be implemented,

and requires knowledge of statistical tools that did not come up in the education of the operators. A Shewhart chart of (means of) observations is much easier to work with than a CUSUM chart of residuals of one-step ahead predictions. Therefore, the following procedure is now used at Philips Stadskanaal.

From the experiment it became clear that the output of the process will slowly deteriorate if the bath is not replenished regularly. Not replenishing for a period of two weeks will bring output measurements outside specification limits. Since the bath is filtered once a week, this is a natural moment to take a sample and have it analyzed accurately by the laboratory department. Based on these measurements, the bath is replenished in such a way that all concentrations fall amply within their limits. For the rest of the week, apart from daily additions of formalin and brightener, additions are only allowed if this follows from the *Out-of-Control Action Plan* (OCAP), which is integrated in the automated SPC software. In the OCAP the process knowledge of the operators and chemical engineers as well as process supervisors is combined into a set of questions which initiate a systematic search for the cause of the out-of-control observations, and advise a remedy that in most cases can be applied by the operator to remove the cause that is responsible for the out-of-control signal.

The OCAP is triggered by out-of-control signals on regular Shewhart charts with widened control limits to allow for a slight wandering of the mean. This type of process monitoring is applicable since only relatively short series (one week's data) are considered so that in 'normal' situations only a small deviation *due to common causes* in the mean will occur. We acknowledge that this approach could be improved upon, given the serial correlation present. However, in practice a balance must be struck between what is possible and what is optimal.

The result of the new replenishing strategy is illustrated by Figures 8.25 and 8.26. Figure 8.25 shows layer thickness measurements of one weeks production before changing the replenishment strategy. The production data of the same week one year later, after changing the replenishment strategy, is depicted in Figure 8.26.

From these Figures, it can be observed that the variation in the individual observations decreased considerably, and that there is a substantial decline in the number of outliers. The outcomes of the process are much more predictable. In addition, before the experiment, the mean level of layer thickness was raised to ensure that no individual measurement was found below 1μm. In the new situation, it was possible to reduce the mean level of layer thickness, without observing layer thicknesses below

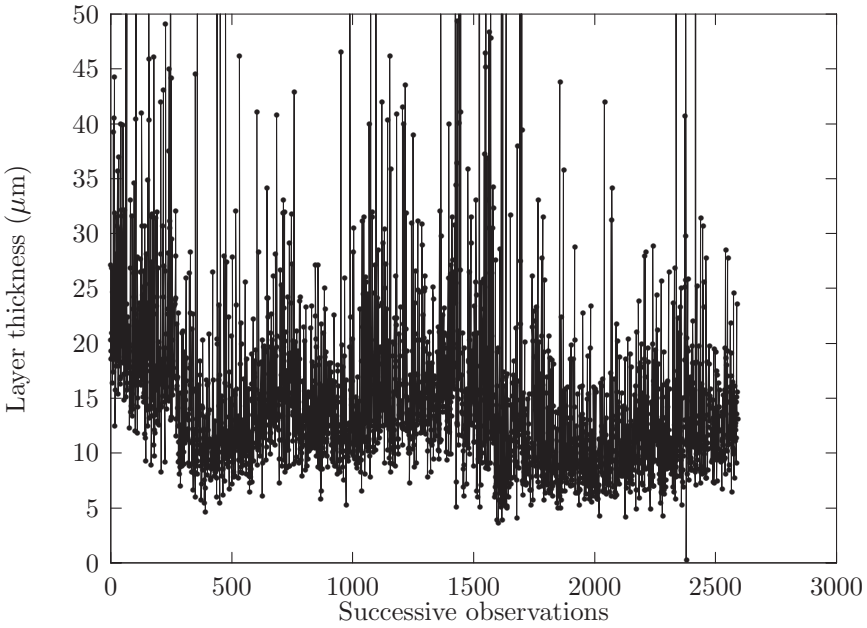


Figure 8.25: Layer thickness measurements before experiment.

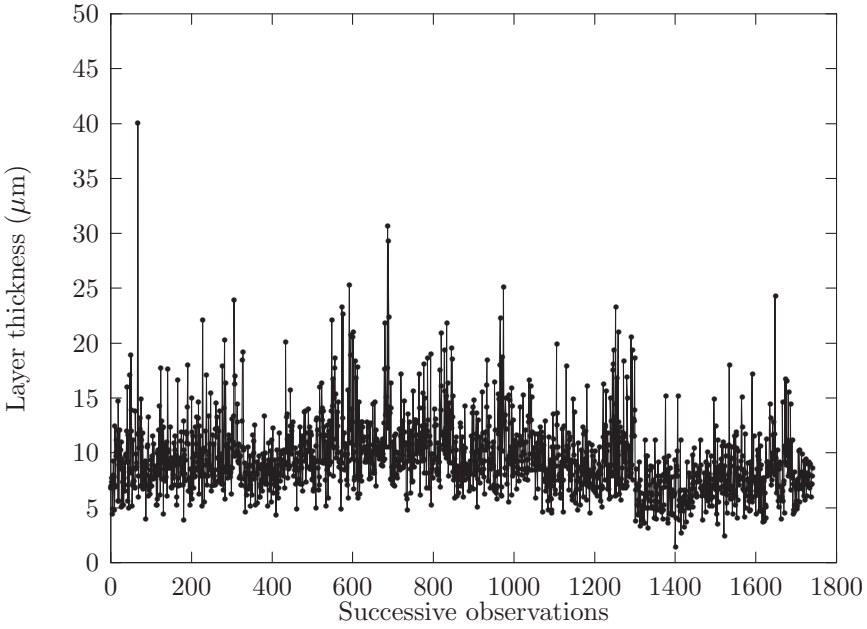


Figure 8.26: Layer thickness measurements after experiment.

1μm. This can be illustrated by comparing the average of the observations of Figure 8.25 (15.2350μm) to the average of Figure 8.26, which equals 9.3990μm.

Recall from Section 8.1 that the assignment of the PAT was to improve the solderability of the diodes. Solderability is to a large extent determined by the quality of the tin/lead layer that is applied to the diodes. Therefore, the efforts of the PAT were aimed at improving the quality of the tin/lead layer, by controlling the quality characteristics layer thickness and layer composition. The PAT succeeded in reducing the variation in these quality characteristics of the tin/lead layer, while reducing the cost of running the process of tin-plating. As a result of their achievements, the number of customer complaints regarding bad solderability of the diodes decreased to zero.

Summary and conclusions

Statistical Process Control (SPC) aims at quality improvement through reduction of variation. The best known tool of SPC is the control chart. Over the years, the control chart has proved to be a successful practical technique for monitoring process measurements. However, its usefulness in practice is limited to those situations where it can be assumed that successive measurements are independently distributed, whereas most data sets encountered in practice exhibit some form of serial correlation. In Chapter 1, several 'real-life' examples are discussed in which the independence assumption is violated. The examples show that in some cases, a control chart signals too frequently when the process is actually in control. In other cases, a control chart does not signal when it should. In either case, it is obvious that such a control chart is not the proper tool to monitor serially correlated process data.

The question that is considered in this thesis is what control chart methods should be used to monitor serially correlated data, and how the signals on such charts should be interpreted.

In Chapter 2, the basic principles of SPC are discussed. An attempt is made to define the term 'quality'. Furthermore, it is argued that there is a close relationship between quality and the reduction of variation. A distinction between two types of variation is made: variation due to common causes and variation due to special causes. It is explained that the purpose of a control chart is to detect special causes of variation, so that they can be removed, thereby improving the quality of the process. However, this technique is developed for observations that are independently distributed. At the end of the chapter, ARIMA time series models are briefly discussed. These models can be employed to capture a wide range of serial correlation structures in the data. In this thesis, we mainly concentrated on a special case: the AR(1) model.

In Chapter 3, Shewhart-type control charts for the mean of individual AR(1) observations are discussed. Shewhart-type control charts for

individual observations utilize only the last measurement to monitor the process for special causes of variation. It is explained how the Shewhart chart works, and how its signals are interpreted in case of independent data. Next, it is investigated how well a classical Shewhart chart (that was designed to monitor the mean of independent data) performs when it is unknowingly applied to AR(1) data. It turns out that, in case of negative autocorrelation, the classical Shewhart chart is less sensitive to shifts in the mean than intended, whereas the chart produces too many false signals in case of positive autocorrelation. This is partly caused by the fact that the most commonly used estimators for the variance are biased if the data are serially correlated.

Next it is investigated how Shewhart-type control charts for the mean perform in the ideal situation where the model and all of its parameters are known. In the literature on SPC we encountered two Shewhart-type control charts that take serial correlation into account. The first chart is the modified Shewhart chart. The points that are plotted on the control chart are simply the correlated observations. However, the control limits are adjusted to allow for serial correlation in the data. Secondly, the residuals chart is suggested in the literature on SPC. In this control chart, the residuals of fitting a time series model to the data are monitored for the presence of a special cause of variation.

For each of these two charts *Average Run Length* (ARL) considerations are presented. Based on a comparison of the ARL curves, it is recommended to use a residuals chart in case of negative autocorrelation, and to use the modified Shewhart chart in case of positive autocorrelation. It is explained why the residuals chart has a very bad ARL performance for positive autocorrelation. This is an important drawback of the residuals chart, since positive autocorrelation is more commonly encountered in practice than negative autocorrelation. The ARL performance of the modified Shewhart chart is better, but we are a little uncomfortable with the fact that the information on the time series structure is not used at all.

To overcome both of these drawbacks, a third Shewhart-type control chart is suggested: the modified residuals chart. This chart turns out to have the best ARL performance for the case of positive autocorrelation. For the case of negative autocorrelation, the ARL performance of the modified residuals chart is almost as good as the ARL performance of the residuals chart. Besides a good overall ARL behavior, the modified residuals chart explicitly takes the correlation structure into account. On these grounds, it is recommended to use the modified residuals chart to monitor the mean of serially correlated data.

In Chapter 4, *Exponentially Weighted Moving Average* (EWMA) charts for the mean of individual AR(1) measurements are discussed. The setup of this chapter is similar to that of Chapter 3. Firstly, it is explained how the EWMA chart for independent observations works and how signals of this control chart are to be interpreted. Secondly, the effect of unknowingly applying an EWMA chart that was designed to monitor independent observations to AR(1) data is investigated. The effect on the ARL behavior is comparable to the effect that was observed in Chapter 3, but much stronger. In the remainder of this chapter, three EWMA-type control charts are discussed and compared on the basis of the corresponding ARL curves. Successively, the modified EWMA chart, the EWMA chart of residuals, and the EWMA chart of modified residuals are discussed. Not surprisingly, the overall ARL performance of EWMA-type control charts turns out to be much better compared to the ARL performance of Shewhart-type control charts. The differences between the three EWMA control charts exhibit the same pattern as that observed in Chapter 3. The EWMA of modified residuals turned out to be the best choice.

In Chapter 5, *CUmulative SUM* (CUSUM) charts for the mean of individual AR(1) measurements are discussed. This chapter is set up similarly to the previous two chapters. After an introduction to the CUSUM for the case of independent observations, it is established that the CUSUM chart is also very sensitive to serial correlation. If AR(1) dependence is unknowingly ignored, this leads to misplacement of the control limits and, consequently, to misinterpretation of the signals of the CUSUM chart. Subsequently, three CUSUM-type control charts are discussed that take serial correlation into account. Successively, the modified CUSUM, the CUSUM of residuals, and the CUSUM of modified residuals are brought to the attention of the reader. At the end of the chapter, the ARL performance of these three charts is evaluated for different types of autocorrelation. The overall performance of the charts turns out to be similar to that of the EWMA-type control charts of Chapter 4. However, the differences between the ARL behavior of the three CUSUM-type control charts are smaller. Notably, the CUSUM chart of residuals performs better than the EWMA chart of residuals for strong positive autocorrelation. For the case of negative autocorrelation, the modified CUSUM turns out to be very inefficient in signalling smaller shifts than the chart was designed for to detect. If a CUSUM-type control chart is to be used for monitoring the mean of serially correlated data, it is recommended to use either the CUSUM of residuals or the CUSUM of modified residuals.

In Chapter 6, the use of the charts of the previous three chapters is

illustrated by means of two examples. The first example is taken from the classical work of Shewhart (1931). The data of this example are shown to exhibit AR(1) dependence. Shewhart, who ignored the serial correlation in the data, arrived at other conclusions than we do. In the second example, a simulated sequence of AR(1) observations is monitored for a shift in the mean. In this example, the ability of the modified residuals chart to detect shifts in the mean earlier than the modified Shewhart chart and the Shewhart chart of residuals is demonstrated. In addition, the use of an EWMA-type control chart is illustrated.

In Chapter 7, control charts for the spread of AR(1) data are discussed. There has been an ongoing debate in the literature on SPC over the question of whether or not to complement a control chart for the mean of individual independent observations with a moving range control chart for the spread. We support the view that the moving range chart adds little power to a control chart for the mean of individual observations. However, if one decides to add a moving range chart, we recommend to use only an upper limit on the *MR* chart. Furthermore, in case of serially correlated individual measurements, we argue that a control chart for the spread should be based on residuals, and not on the correlated measurements. Next, control charts for subgrouped serially correlated data are discussed. Successive means of subgroups of AR(1) data are shown to exhibit ARMA(1,1) correlation.

Successively, four control charts for the spread of subgrouped AR(1) data are discussed, viz. the $\overline{MR}$-chart, the S^2-chart, the R-chart, and the R-chart of residuals. The chapter ends with an ARL comparison of these four charts. It turns out that, of the four charts considered, the R-chart based on residuals provides the best results. Therefore, it is our recommendation that a control chart for the spread of subgrouped serially correlated data should be based on residuals.

Chapter 8 discusses a case-study in which the author got involved. At Philips Semiconductors Stadskanaal, diodes are produced that are to be soldered on printed circuit boards. Customer complaints regarding the solderability of the diodes were the reason that a so-called *Process Action Team* (PAT) was started. The objective of this PAT was to improve solderability of the diodes by improving the quality of the tin/lead layer that is applied to the connection points of the diodes. In this chapter, the experiences the author has gained when he had the opportunity to assist this PAT, are written down. Different aspects of the quality improvement project are discussed, ranging from *Linear Programming* to monitoring serially correlated data. The PAT eventually succeeded in improving the quality of the tin/lead layer, and the number of customer complaints decreased to zero.

References

Adams, B. M., W. H. Woodall, and C. Lowry (1992), "The use (and misuse) of false alarm probabilities in control chart design", *Frontiers in Statistical Quality Control*, **4**, 155–168.

Adke, S. R. and X. Hong (1997), "A supplementary test based on the control chart for individuals", *Journal of Quality Technology*, **29**(1), 16–20.

Alwan, L. C. (1989), *Time Series Modeling for Statistical Process Control*, PhD thesis, Graduate School of Business, Chicago.

Alwan, L. C. (1992), "Effects of autocorrelation on control chart performance", *Communications in Statistics: Theory and Methods*, **21**(4), 1025–1049.

Alwan, L. C. and M. G. Bissell (1988), "Time series modeling for quality control in clinical chemistry", *Clinical Chemistry*, **34**(7), 1396–1406.

Alwan, L. C. and D. Radson (1992a), "Investigation of the subsample range from positive autocorrelated processes", in *Modeling and Simulation: Control, Digital Signal Processing, Robotics, Systems, Power*, University of Pittsburg, 1791–1798.

Alwan, L. C. and D. Radson (1992b), "Time series investigation of subsample mean charts", *IIE Transactions*, **24**(5), 66–80.

Alwan, L. C. and H. V. Roberts (1988), "Time-series modeling for statistical process control", *Journal of Business & Economic Statistics*, **6**(1), 87–95.

Alwan, L. C. and H. V. Roberts (1995), "The problem of misplaced control limits", *Journal of the Royal Statistical Society, Series C*, **44**(3), 269–306 [With Discussion and Reply].

Amin, R. W. and R. A. Ethridge (1998), "A note on individual and moving range control charts", *Journal of Quality Technology*, **30**(1), 70–74.

Anderson, O. D. (1976), *Time Series Analysis and Forecasting*, Butterworths, London.

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, John Wiley & Sons, New York.

Bagshaw, M. and R. A. Johnson (1975), "The effect of serial correlation on the performance of CUSUM tests II", *Technometrics*, **17**(1), 73–80.

Banens, P. J. A., R. J. M. M. Does, G. B. W. van Dongen, J. Engel, M. M. A. Hasselaar, R. A. J. M. van Lieshout, J. Praagman, B. F. Schriever, A. Trip, and H. van der Veen (1994), *Industriële Statistiek en Kwaliteit*, Kluwer Bedrijfswetenschappen, Deventer, The Netherlands [In Dutch].

Barnard, G. A. (1959), "Control charts and stochastic processes", *Journal of the Royal Statistical Society, Series B*, **21**(2), 239–271 [With Discussion and Reply].

Berthouex, P. M., W. G. Hunter, and L. Pallesen (1978), "Monitoring sewage treatment plants: Some quality control aspects", *Journal of Quality Technology*, **10**(4), 139–149.

Bischak, D. P., W. D. Kelton, and S. M. Pollock (1993), "Weighted batch means for confidence intervals in steady-state simulations", *Management Science*, **39**(8), 1002–1019.

Box, G. and T. Kramer (1992), "Statistical process monitoring and feedback adjustment—a discussion", *Technometrics*, **34**(2), 251–285.

Box, G. and A. Luceño (1997a), "Discrete proportional-integral adjustment and statistical process control", *Journal of Quality Technology*, **29**(3), 248–260.

Box, G. and A. Luceño (1997b), *Statistical Control by Monitoring and Feedback Adjustment*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York.

Box, G. E. P., D. E. Coleman, and R. V. Baxley, Jr. (1997), "A comparison of statistical process control and engineering process control", *Journal of Quality Technology*, **29**(2), 128–130.

Box, G. E. P. and G. M. Jenkins (1963), "Further contributions to adaptive quality control: Simultaneous estimation of dynamics: Non-zero costs", *Bulletin International Statistical Institute*, 943–973 [34th Session].

Box, G. E. P. and G. M. Jenkins (1976), *Time Series Analysis, Forecasting and Control*, revised edition, Holden-Day, San Francisco.

Brockwell, P. J. and R. A. Davis (1991), *Time Series: Theory and Methods*, Springer Series in Statistics, second edition, Springer-Verlag, New York.

Brook, D. and D. A. Evans (1972), "An approach to the probability distribution of CUSUM run length", *Biometrika*, **59**(3), 539–549.

Burr, I. W. (1967), "The effect of non-normality on constants for $\bar{X}$ and R charts", *Industrial Quality Control*, **23**(11), 563–569.

Champ, C. W. and S. E. Rigdon (1991), "A comparison of the Markov chain and the integral equation approaches for evaluating the run length distribution of quality control charts", *Communications in Statistics: Simulation and Computation*, **20**(1), 191–204.

Champ, C. W. and W. H. Woodall (1987), "Exact results for Shewhart control charts with supplementary runs rules", *Technometrics*, **29**(4), 393–399.

Chou, Y.-M., A. M. Polansky, and R. L. Mason (1998), "Transforming non-normal data to normality in statistical process control", *Journal of Quality Technology*, **30**(2), 133–141.

Crosby, P. B. (1979), *Quality is Free: the Art of Making Quality Certain*, McGraw-Hill, New York.

Crowder, S. V. (1987), "A simple method for studying run-length distributions of exponentially weighted moving average charts", *Technometrics*, **29**(4), 401–407.

Crowder, S. V., D. M. Hawkins, M. R. Reynolds, Jr., and E. Yashchin (1997), "Process control and statistical inference", *Journal of Quality Technology*, **29**(2), 134–139.

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

Deming, W. E. (1982), *Out of the Crisis*, Cambridge University Press, Cambridge.

Dobben de Bruyn, C. S. van (1968), *Cumulative Sum Tests: Theory and Practice*, Griffin's Statistical Monographs and Courses, Griffin, London.

Does, R. J. M. M., M. L. van Oord, and A. Trip (1994), "Een succesvolle implementatie van statistische procesbeheersing (SPC)", *Sigma*, **40**(6), 10–13 [In Dutch].

Does, R. J. M. M., K. C. B. Roes, and A. Trip (1999), *Statistical Process Control in Industry: Implementing and Assuring SPC*, Kluwer Academic, Dordrecht, The Netherlands.

Does, R. J. M. M. and B. F. Schriever (1992), "Variables control charts limits and tests for special causes", *Statistica Neerlandica*, **46**(1), 229–245.

Domangue, R. and S. C. Patch (1991), "Some omnibus exponentially weighted moving average statistical process monitoring schemes", *Technometrics*, **33**(3), 299–313.

Duncan, A. J. (1986), *Quality Control and Industrial Statistics*, fifth edition, Irwin, Homewood.

Ermer, D. S. (1980), "A control chart for dependent data", in *ASQC Technical Conference Transactions*, ASQC, 121–128.

Ermer, D. S., M. C. Chow, and S. M. Wu (1979), "A time series control chart for a nuclear reactor", in *Proceedings 1979 Annual Reliability and Maintainability Symposium*, 92–98.

Ewan, W. D. and K. W. Kemp (1960), "Sampling inspection of continuous processes with no autocorrelation between successive results", *Biometrika*, **47**, 363–380.

Faltin, F. W., C. M. Mastrangelo, G. C. Runger, and T. P. Ryan (1997), "Considerations in the monitoring of autocorrelated and independent data", *Journal of Quality Technology*, **29**(2), 131–133.

Gan, F. F. (1991), "An optimal design of CUSUM quality control charts", *Journal of Quality Technology*, **23**(4), 279–286.

Garvin, D. A. (1987), "Competing on the eight dimensions of quality", *Harvard Business Review*, **87**(6), 101–109.

Gilbert, K. C., K. Kirby, and C. R. Hild (1997), "Charting autocorrelated data: Guidelines for practitioners", *Quality Engineering*, **9**(3), 367–382.

Goldsmith, P. L. and H. Whitfield (1961), "Average run length in cumulative chart quality control schemes", *Technometrics*, **3**(1), 11–20.

Granger, C. W. J. and P. Newbold (1986), *Forecasting Economic Time Series*, second edition, Academic Press, San Diego.

Haridy, A. M. A. and A. Z. El-Shabrawy (1996), "The economic design of cumulative sum charts used to maintain current control of non-normal process means", *Computers & Industrial Engineering*, **31**, 783–790.

Harris, T. J. and W. H. Ross (1991), "Statistical process control procedures for correlated observations", *Canadian Journal of Chemical Engineering*, **69**, 48–57.

Harvey, A. C. (1993), *Time Series Models*, second edition, Harvester Wheatsheaf, New York.

Hawkins, D. M. (1981), "A CUSUM for a scale parameter", *Journal of Quality Technology*, **13**(4), 228–231.

Healy, J. D. (1987), "A note on multivariate CUSUM procedures", *Technometrics*, **29**, 409–412.

Hunter, S. J. (1986), "The exponentially weighted moving average", *Journal of Quality Technology*, **18**(4), 203–210.

Janakiram, M. and J. B. Keats (1998), "Combining SPC and EPC in a hybrid industry", *Journal of Quality Technology*, **30**(3), 189–200.

Johnson, N. L. (1961), "A simple theoretical approach to cumulative sum control charts", *Journal of the American Statistical Association*, **56**, 835–840.

Johnson, N. L. and S. Kotz (1972), *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley & Sons, New York.

Johnson, R. A. and M. Bagshaw (1974), "The effect of serial correlation on the performance of CUSUM tests", *Technometrics*, **16**(1), 103–112.

Juran, J. M. and F. M. Gryna, editors (1988), *Juran's Quality Control Handbook*, fourth edition, McGraw-Hill, New York.

Kang, K. and B. Schmeiser (1987), "Properties of batch means from stationary ARMA time series", *Operations Research Letters*, **6**(1), 19–24.

Keats, J. B. and N. F. Hubele, editors (1989), *Statistical Process Control in Automated Manufacturing*, Quality and Reliability, Marcel Dekker, Inc, New York.

Kemp, K. W. (1961), "The average run length of the cumulative sum chart when a V-mask is used", *Journal of the Royal Statistical Society, Series B*, **23**, 149–153.

Knypstra, S. (1997), "Kansverdelingen en objecten", *Kwantitatieve Methoden*, **18**(54), 91–95 [In Dutch].

Kramer, H. and W. Schmid (1996), "Control charts for time series", Arbeitsbericht 59, Europa-Universität Viadrina Frankfurt (Oder), Fakultät für Wirtschaftswissenschaften.

Kramer, H. and W. Schmid (1996a), "The influence of parameter estimation on the ARL of Shewhart type charts for time series", Arbeitsbericht 60, Europa-Universität Viadrina Frankfurt (Oder), Fakultät für Wirtschaftswissenschaften.

Kress, R. (1989), *Linear Integral Equations*, Springer-Verlag, Berlin.

Lin, W. S. W. and B. M. Adams (1996), "Combined control charts for forecast-based monitoring schemes", *Journal of Quality Technology*, **28**(3), 289–301.

Longnecker, M. T. and T. P. Ryan (1992), "Charting correlated process data", Technical Report 166, Texas A&M University, Department of Statistics.

Lu, C.-W. and M. R. Reynolds, Jr. (1997), "Control charts for monitoring the mean and variance of autocorrelated processes", Technical report, Virginia Polytechnic Institute and State University [To appear in Journal of Quality Technology].

Lucas, J. M. (1973), "A modified "V" mask control scheme", *Technometrics*, **15**(4), 833–847.

Lucas, J. M. and R. B. Crosier (1982), "Fast initial response for CUSUM schemes: Give your CUSUM a head start", *Technometrics*, **24**(3), 199–205.

Lucas, J. M. and M. S. Saccucci (1990), "Exponentially weighted moving average control schemes: Properties and enhancements", *Technometrics*, **32**(1), 1–29.

MacGregor, J. F. (1988), "On-line statistical process control", *Chemical Engineering Progress*, **84**(10), 21–31.

MacGregor, J. F. and T. J. Harris (1993), "The exponentially weighted moving variance", *Journal of Quality Technology*, **25**(2), 106–118.

Magnus, J. R. and H. Neudecker (1979), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, New York.

Maragah, H. D. and W. H. Woodall (1992), "The effect of autocorrelation of the retrospective $\bar{X}$-chart", *Journal of Statistical Computation and Simulation*, **40**, 29–42.

McCoun, V. E. (1974), "The case of the perjured control chart, when actual piece part results don't jibe with the chart, what's the answer?", *Quality Progress*, **7**(10), 17–19 [Originally published in *Industrial Quality Control*, May 1949].

Montgomery, D. C. (1996), *Introduction to Statistical Quality Control*, third edition, John Wiley & Sons, New York.

Montgomery, D. C., L. A. Johnson, and J. S. Gardiner (1990), *Forecasting & Time Series Analysis*, McGraw-Hill, New York.

Montgomery, D. C., J. B. Keats, G. C. Runger, and W. S. Messina (1994), "Integrating statistical process control and engineering process control", *Journal of Quality Technology*, **26**(2), 79–87.

Montgomery, D. C. and C. M. Mastrangelo (1991), "Some statistical process control methods for autocorrelated data", *Journal of Quality Technology*, **23**(3), 179–204.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974), *Introduction to the Theory of Statistics*, third edition, McGraw-Hill, Auckland.

Nelson, L. S. (1982), "Control charts for individual measurements", *Journal of Quality Technology*, **14**(3), 172–173.

Nelson, L. S. (1988), "Control charts: Rational subgroups and effective applications", *Journal of Quality Technology*, **20**(1), 73–75.

Nelson, L. S. (1990), "Monitoring reduction in variation with a range chart", *Journal of Quality Technology*, **22**(2), 163–165.

Nolan, T. W. and L. P. Provost (1990), "Understanding variation", *Quality Progress*, **23**(5), 70–78.

Notohardjono, B. D. and D. S. Ermer (1986), "Time series control charts for correlated and contaminated data", *Journal of Engineering for Industry*, **108**, 219–226.

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

Pandit, S. M. and S. Wu (1983), *Time Series and System Analysis with Applications*, John Wiley & Sons, New York.

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

Robinson, P. B. and T. Y. Ho (1978), "Average run lengths of geometric moving average charts by numerical methods", *Technometrics*, **20**(1), 85–93.

Roes, K. C. B., R. J. M. M. Does, and Y. Schurink (1993), "Shewhart-type control charts for individual observations", *Journal of Quality Technology*, **25**(3), 188–198.

Runger, G. C. and T. R. Willemain (1995), "Model-based and model-free control of autocorrelated processes", *Journal of Quality Technology*, **27**(4), 283–292.

Ryan, T. P. (1991), "Discussion of: Some statistical process control methods for autocorrelated data", *Journal of Quality Technology*, **23**(3), 200–202.

Schilling, E. G. and P. R. Nelson (1976), "The effect of non-normality on the control limits of $\bar{X}$ charts", *Journal of Quality Technology*, **8**(4), 183–188.

Schmid, W. (1995a), "The effects of autocorrelation on the R-chart and the S^2-chart", Arbeitsbericht 10/95, Europa-Universität Viadrina Frankfurt (Oder), Fakultät für Wirtschaftswissenschaften.

Schmid, W. (1995b), "On the run length of a Shewhart chart for correlated data", *Statistical Papers*, **36**, 111–130.

Schmid, W. (1997a), "CUSUM control schemes for gaussian processes", *Statistical Papers*, **38**, 191–217.

Schmid, W. (1997b), "On EWMA charts for time series", *Frontiers in Statistical Quality Control*, **5**, 114–137.

Schmid, W. and A. Schöne (1997), "Some properties of the EWMA control chart in the presence of autocorrelation", *Annals of Statistics*, **25**(3), 1277–1283.

Shewhart, W. A. (1931), *Economic Control of Quality of Manufactured Product*, D. van Nostrand Company, Inc., Toronto.

Shore, H. (1998), "A new approach to analysing non-normal quality data with application to process capability analysis", *International Journal of Production Research*, **36**(7), 1917–1934.

Sierksma, G. (1996), *Linear and Integer Programming: Theory and Practice*, Marcel Dekker Inc., New York.

Snee, R. D. (1990), "Statistical thinking and its contribution to total quality", *American Statistician*, **44**(2), 116–121.

Sullivan, L. P. (1984), "Reducing variability: A new approach to quality", *Quality Progress*, **17**(7), 15–21.

Tseng, S. and B. M. Adams (1994), "Monitoring autocorrelated processes with an exponentially weighted moving average forecast", *Journal of Statistical Computation and Simulation*, **50**, 187–195.

VanBrackle, L. N. and M. R. Reynolds, Jr. (1997), "EWMA and CUSUM control charts in the presence of correlation", *Communications in Statistics: Simulation and Computation*, **26**(3), 979–1008.

Vasilopoulos, A. V. and A. P. Stamboulis (1978), "Modification of control chart limits in the presence of data correlation", *Journal of Quality Technology*, **10**(1), 20–30.

Wardell, D. G., H. Moskowitz, and R. D. Plante (1992), "Control charts in the presence of data correlation", *Management Science*, **38**(8), 1084–1105.

Wardell, D. G., H. Moskowitz, and R. D. Plante (1994), "Run length distributions of residual control charts for autocorrelated processes", *Journal of Quality Technology*, **26**(4), 308–317.

Wardell, D. G., H. Moskowitz, and R. D. Plante (1994), "Run-length distributions of special-cause control charts for correlated processes", *Technometrics*, **36**(1), 3–27.

Wetherill, G. B. and D. W. Brown (1991), *Statistical Process Control, Theory and Practice*, Chapman and Hall, London.

Wheeler, D. J. (1989), "Shewhart's control charts: Foundations & myths", *Quality Digest*, 32–39.

Wheeler, D. J. and D. S. Chambers (1990), *Understanding Statistical Process Control*, Addison-Wesley Publishing Company, Wokingham.

Wieringa, J. E. (1997), "The case of tin-plating of surface mounted glass diodes", Research Report 97A18, Research Institute Systems, Organisations and Management, University of Groningen, Groningen.

Wieringa, J. E. (1998), "Control charts for monitoring the mean of AR(1) data", in *54th Annual Quality Conference 1998 Transactions*, American Society for Quality Rochester Section, 67–103.

Woodall, W. H. and B. M. Adams (1993), "The statistical design of CUSUM charts", *Quality Engineering*, **5**(4), 559–570.

Yashchin, E. (1993), "Performance of CUSUM control schemes for serially correlated observations", *Technometrics*, **35**(1), 37–52.

Zhang, N. F. (1998), "A statistical control chart for stationary process data", *Technometrics*, **40**(1), 24–38.

Author index

Subject index

Nederlandse samenvatting

Statistical Process Control (SPC) wordt in het Nederlands vertaald met Statistische Procesbeheersing. Het doel van SPC is het realiseren van kwaliteitsverbetering door het beheersen van variatie in procesuitkomsten. Een van de belangrijkste SPC-instrumenten die hierbij gebruikt kunnen worden is de regelkaart. De regelkaart werd in de twintiger jaren ontwikkeld door dr. Walter A. Shewhart. In de loop der jaren is gebleken dat deze techniek in de praktijk met succes kan worden ingezet voor het bewaken van processen. Echter, de toepasbaarheid van de regelkaart is beperkt tot situaties waarin kan worden aangenomen dat opeenvolgende procesmetingen statistisch onafhankelijk zijn. Deze veronderstelling kan in de praktijk vaak niet gemaakt worden: veelal vertonen procesmetingen een of andere samenhang in de tijd (seriële correlatie). In hoofdstuk 1 worden verschillende 'echte' praktijkvoorbeelden besproken waarbij de onafhankelijkheidsaanname niet kan worden gemaakt. De voorbeelden geven aan dat de standaardregelkaart in sommige van deze gevallen te vaak een signaal geeft terwijl het proces in werkelijkheid beheerst is. In andere gevallen is het effect van seriële correlatie dat een signaal van de regelkaart te lang op zich laat wachten wanneer het proces niet beheerst is. In beide gevallen is het duidelijk dat de regelkaart niet zonder meer gebruikt kan worden voor het bewaken van serieel gecorreleerde data.

De vraag die in dit proefschrift aan de orde gesteld wordt is welke regelkaarten gebruikt kunnen worden voor het bewaken van serieel gecorreleerde data, en hoe signalen van deze regelkaarten geïnterpreteerd dienen te worden.

In hoofdstuk 2 worden de basisbeginselen van SPC besproken. Allereerst wordt getracht een definitie te geven van de term 'kwaliteit'. Vervolgens komt de relatie tussen kwaliteit en variatie aan de orde. Er wordt een tweedeling in de oorzaken van variatie aangebracht: gewone oorzaken van variatie en bijzondere oorzaken van variatie. De regelkaart is ontworpen om de aanwezigheid van speciale oorzaken van variatie te signaleren,

zodat deze verwijderd kunnen worden. De variatie in de uitkomsten van een proces wordt hierdoor gereduceerd, waarmee de kwaliteit van het proces verbetert. Een proces waarvan de uitkomsten slechts beïnvloed worden door gewone oorzaken van variatie wordt 'statistisch beheerst' genoemd. Voor de juiste interpretatie van de signalen van een regelkaart wordt in het grootste deel van de SPC-literatuur de aanname gemaakt dat de procesuitkomsten onderling onafhankelijk zijn. Het hoofdstuk besluit met een korte bespreking van ARIMA tijdreeksmodellen, die kunnen worden gebruikt voor het modelleren van seriële correlatie. In dit proefschrift komt met name een belangrijk speciaal geval, het AR(1) model, aan de orde.

In hoofdstuk 3 worden Shewhart-regelkaarten voor het gemiddelde van individuele AR(1) waarnemingen besproken. Een Shewhart-regelkaart voor individuele observaties gebruikt alleen de laatste waarneming voor het detecteren van bijzondere oorzaken van variatie. Er wordt ingegaan op de werking en op de interpretatie van signalen van de Shewhart-regelkaart in geval van onafhankelijke waarnemingen. Vervolgens wordt vastgesteld hoe een dergelijke regelkaart, die ontworpen is voor gebruik met onafhankelijke waarnemingen, presteert wanneer deze wordt toegepast op AR(1) waarnemingen. In geval van negatieve autocorrelatie blijkt de Shewhart-regelkaart minder gevoelig voor verschuivingen in het gemiddelde dan de bedoeling was. Wanneer er sprake is van positieve autocorrelatie zal de regelkaart veel valse signalen genereren. Een van de oorzaken voor deze fenomenen is het feit dat de gebruikelijke schatters voor de variantie onzuiver zijn wanneer de observaties seriële correlatie vertonen.

Vervolgens wordt onderzocht hoe Shewhart-regelkaarten die geschikt zijn voor gebruik met serieel gecorreleerde data presteren wanneer het onderliggende tijdreeksmodel en de bijbehorende parameters bekend zijn. In de SPC-literatuur kunnen twee soorten regelkaarten onderscheiden worden die rekening houden met seriële correlatie. De eerste soort is de groep van zogenaamde 'modified Shewhart charts' (aangepaste Shewhart-kaarten). De datapunten die in een dergelijke regelkaart getekend worden zijn eenvoudigweg de gecorreleerde waarnemingen. De breedte van de regelgrenzen is zodanig aangepast dat signalen op de gebruikelijke manier geïnterpreteerd kunnen worden. De tweede groep van regelkaarten bestaat uit 'residuals charts' (residuenkaarten). Deze regelkaart volgt residuen van een geschat tijdreeksmodel om bijzondere oorzaken van variatie te detecteren.

Voor beide kaarten wordt de zogenaamde *Average Run Length* (ARL) curve besproken. Een vergelijking van deze curves leidt tot het advies om een residuenregelkaart slechts te gebruiken in geval van negatieve seriële correlatie. Een aangepaste Shewhart-kaart verdient de voorkeur wanneer de

observaties positief serieel gecorreleerd zijn. Vervolgens wordt uiteengezet waarom voor positieve autocorrelatie het ARL-gedrag van een residuenkaart beduidend slechter is dan dat van een aangepaste Shewhart-kaart. Dit vormt een belangrijk praktisch nadeel van de residuenkaarten, daar positieve autocorrelatie in de praktijk veel vaker wordt aangetroffen dan negatieve autocorrelatie. Op het gebruik van aangepaste Shewhart-kaarten valt aan te merken dat de extra informatie die de data bevat ten aanzien van de afhankelijkheidsstructuur in feite niet wordt gebruikt.

In hoofdstuk 3 wordt daarom een derde soort Shewhart-regelkaarten voorgesteld, die beide nadelen niet kent: de 'modified residuals charts' (aangepaste residuenkaarten). Het ARL-gedrag van een dergelijke kaart blijkt beter te zijn dan dat van een aangepaste Shewhart-kaart en dan dat van een residuenkaart in geval van positieve autocorrelatie. Wanneer de observaties negatief gecorreleerd zijn, is het ARL-gedrag bijna net zo goed als dat van een residuenkaart, en beduidend beter dan dat van een aangepaste Shewhart-kaart. Naast een goed ARL-gedrag voor zowel positieve als negatieve autocorrelatie maakt de aangepaste residuenkaart expliciet gebruik van de afhankelijkheid die is aangetroffen in de data. Op grond van deze argumenten wordt geadviseerd om de aangepaste residuenkaart te gebruiken voor het bewaken van het gemiddelde van individuele AR(1) metingen.

In hoofdstuk 4 is de aandacht gericht op *Exponentially Weighted Moving Average* (EWMA) regelkaarten voor het gemiddelde van individuele AR(1) observaties. De opzet van dit hoofdstuk is gelijk aan die van hoofdstuk 3. Allereerst wordt ingegaan op de werking van een EWMA-kaart. Vervolgens wordt uiteengezet hoe een signaal op een EWMA-regelkaart geïnterpreteerd dient te worden wanneer de observaties onderling onafhankelijk mogen worden verondersteld. Daarnaast wordt vastgesteld wat het effect is van het toepassen van een EWMA-kaart voor het gemiddelde van onafhankelijke waarnemingen op AR(1) data. Het effect op het ARL-gedrag van de EWMA-kaart blijkt nog veel groter te zijn dan dat wat in hoofdstuk 3 is waargenomen voor Shewhart-kaarten. Vervolgens worden drie soorten EWMA-kaarten besproken, en vergeleken op basis van hun ARL-gedrag. Achtereenvolgens komen de 'modified EWMA chart' (aangepaste EWMA-kaart), de 'EWMA chart of residuals' (EWMA-kaart van residuen) en de 'EWMA chart of modified residuals' (EWMA-kaart van aangepaste residuen) aan de orde. Over het algemeen blijkt het ARL-gedrag van EWMA-kaarten beter te zijn dan dat van Shewhart-kaarten. De verschillen tussen de drie soorten EWMA-kaarten vertonen eenzelfde patroon als de verschillen tussen de drie soorten Shewhart-kaarten. De EWMA-kaart van

aangepaste residuen blijkt de beste keuze te zijn wanneer het gemiddelde van AR(1) data met een EWMA-kaart bewaakt dient te worden.

In hoofdstuk 5 worden *CUmulative SUM* (CUSUM) kaarten voor het gemiddelde van individuele AR(1) metingen besproken. De opzet van dit hoofdstuk is analoog aan de opzet van de twee voorgaande hoofdstukken. Na een inleiding waarin de werking en de interpretatie van een CUSUM-kaart voor onafhankelijke data uiteengezet wordt, wordt vastgesteld dat de CUSUM-kaart eveneens zeer gevoelig is voor de aanwezigheid van seriële correlatie. Wanneer de observaties seriële correlatie vertonen, is de plaatsing van de regelgrenzen incorrect. Een signaal op de CUSUM-kaart (of het ontbreken daarvan) wordt dientengevolge onjuist geïnterpreteerd. Vervolgens worden drie CUSUM-regelkaarten besproken die geschikt zijn voor het bewaken van het gemiddelde van AR(1) data. Achtereenvolgens worden 'modified CUSUM charts' (aangepaste CUSUM-kaarten), 'CUSUM charts of residuals' (CUSUM-kaarten van residuen) en 'CUSUM charts of modified residuals' (CUSUM-kaarten van aangepaste residuen) besproken. Tot slot van het hoofdstuk worden de ARL-curves van deze drie soorten kaarten besproken voor het geval van AR(1) afhankelijkheid. In het algemeen is het ARL-gedrag van CUSUM-kaarten vergelijkbaar met dat van EWMA-kaarten, en dus beter dan dat van Shewhart-kaarten. Echter, de onderlinge verschillen tussen de ARL-curves van de drie CUSUM-kaarten zijn kleiner. In het bijzonder valt hierbij op te merken dat een CUSUM-kaart van residuen beduidend beter presteert dan een EWMA-kaart van residuen voor sterke positieve autocorrelatie. Bovendien blijkt de aangepaste CUSUM-kaart erg inefficiënt te zijn in het detecteren van verschuivingen die kleiner zijn dan een verschuiving waarvoor de kaart is ingericht. Wanneer een CUSUM-kaart voor het gemiddelde van AR(1) data gebruikt dient te worden, heeft òf een CUSUM-kaart van residuen, òf een CUSUM-kaart van aangepaste residuen de voorkeur.

In hoofdstuk 6 wordt het gebruik van de regelkaarten uit de vorige hoofdstukken geïllustreerd aan de hand van twee voorbeelden. De dataset van het eerste voorbeeld is afkomstig uit het klassieke werk van Shewhart (1931). De metingen laten zich goed modelleren als een AR(1) tijdreeks. Shewhart stelt een regelkaart voor het gemiddelde van deze data op, zonder daarbij rekening te houden met de seriële correlatie. De conclusies die getrokken worden op basis van deze kaart zijn andere dan de onze, wanneer we regelkaarten voor het gemiddelde van AR(1) data gebruiken. In het tweede voorbeeld wordt gedemonstreerd dat een Shewhart-regelkaart van aangepaste residuen sneller een verschuiving in het gemiddelde van gesimuleerde AR(1) data opspoort dan een aangepaste Shewhart-kaart of een

Shewhart-residuenkaart.

In hoofdstuk 7 worden regelkaarten voor de spreiding van AR(1) data besproken. In de SPC-literatuur bestaat er in geval van onafhankelijke individuele waarnemingen controverse ten aanzien van het gebruik van een zogenaamde 'moving range' regelkaart voor de spreiding als toevoeging op een regelkaart voor het gemiddelde. Ons standpunt is dat de moving range kaart weinig toevoegt aan een regelkaart voor het gemiddelde van individuele onafhankelijke waarnemingen. Echter, wanneer toch besloten wordt tot het gebruik van een moving range kaart, verdient het aanbeveling om de data te beoordelen slechts met behulp van de bovenste regelgrens van deze kaart. De onderste regelgrens dient buiten beschouwing gelaten te worden. In geval van serieel gecorreleerde individuele waarnemingen is het advies de regelkaart voor de spreiding te baseren op residuen, en niet op de gecorreleerde waarnemingen.

Vervolgens wordt ingegaan op regelkaarten voor de spreiding, gebaseerd op subgroepen van AR(1) data. Er wordt aangetoond dat de gemiddelden van deze groepen ARMA(1,1) correlatie vertonen. Vier regelkaarten worden besproken waarmee de spreiding in de subgroepen bewaakt kan worden. Achtereenvolgens komen aan de orde de $\overline{MR}$-kaart, de S^2-kaart, de R-kaart, en de R-kaart van residuen. Het hoofdstuk besluit met een vergelijking van de ARL van deze vier kaarten. De resultaten van de R-kaart van residuen blijken beter te zijn dan die van de andere drie regelkaarten. Ook voor een regelkaart voor de spreiding van subgroepen van serieel gecorreleerde data luidt het advies deze te baseren op residuen.

Het laatste hoofdstuk 8 behandelt een case-study die door de auteur werd uitgevoerd. Bij Philips Semiconductors Stadskanaal worden diodes geproduceerd die door de klant op printplaten gesoldeerd worden. Klantenklachten ten aanzien van de soldeerbaarheid van de diodes waren de aanleiding voor de instelling van een zogenaamd 'Process Action Team' (PAT). Dit PAT had zich ten doel gesteld de soldeerbaarheid van de diodes te verbeteren door de beschermende tin/lood-laag die op de aansluitpunten wordt aangebracht te verbeteren. In het hoofdstuk worden de ervaringen besproken die de auteur heeft opgedaan toen hij in de mogelijkheid werd gesteld om dit PAT te assisteren. Verschillende aspecten van dit kwaliteitsverbeteringsproject komen aan de orde, zoals *Lineair Programmeren*, maar ook het bewaken van serieel gecorreleerde data met behulp van een regelkaart. Het PAT slaagde erin de kwaliteit van de tin/lood-laag te verbeteren, en het aantal klantenklachten terug te dringen naar nul.