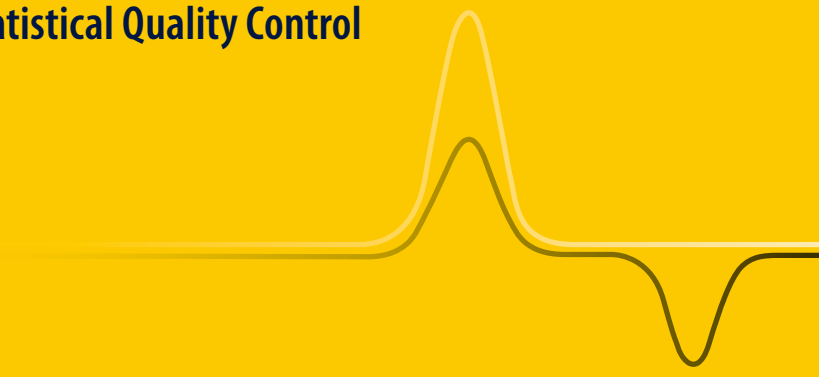
Frontiers in Statistical Quality Control



Sven Knoth · Wolfgang Schmid *Editors*

Frontiers in Statistical Quality Control 12

Springer

Frontiers in Statistical Quality Control

More information about this series at http://www.springer.com/series/10821

Sven Knoth • Wolfgang Schmid
Editors

Frontiers in Statistical Quality Control 12

Springer

Editors
Sven Knoth
Department of Mathematics and Statistics
Helmut Schmidt University
Hamburg, Germany

Wolfgang Schmid
Department of Statistics
European University Viadrina
Frankfurt (Oder), Germany

Frontiers in Statistical Quality Control
ISBN 978-3-319-75294-5 ISBN 978-3-319-75295-2 (eBook)
https://doi.org/10.1007/978-3-319-75295-2

Library of Congress Control Number: 2018940538

Preface

The XIIth International Workshop on *Intelligent Statistical Quality Control* took place in Hamburg, Germany, from August 16 to 18, 2016. The invitational workshop was jointly organized by Professor S. Knoth from Helmut Schmidt University in Hamburg, Germany, and Professor W. Schmid from the European University Viadrina in Frankfurt (Oder), Germany. The former was also the local organizer of the workshop.

This book consists of 20 chapters that were carefully selected and reviewed by the scientific program committee. The focus of the book is on major areas of statistical quality control (SQC). The majority of the chapters address statistical process control (SPC), which is now often called statistical process monitoring (SPM). Important fields such as design of experiments (DOE) and acceptance sampling are also treated.

The book is divided into three parts. The subject of Part I is SPC. Part II is devoted to DOE, and in Part III, related fields are considered.

Part I: Statistical Process Control

In recent years, control charts have not been exclusively applied within Phase II analyses to monitor data. It has been shown that they can be successfully applied in Phase I analyses to identify a stable in-control process. Capizzi and Masarotto present the R package dfphase 1 which provides implementations of many recently proposed distribution-free methods for Phase I analysis. The application of the package is illustrated using data from an oil refinery.

In practice, there are many possible types of changes that may affect an in-control process. M. Testik, Weiß, Koca, and O. Testik analyze the use of a Shewhart chart for Phase I analysis. In their study, they assume independent and normally distributed observations. The behavior of the chart is analyzed for several mean shifts and contamination rates. The authors vary the width of the Shewhart control limits to assess the performance in Phase I implementations. Various performance metrics

such as the true and false alarm percentages, the number of iterations to complete estimation, and the mean squared error of the estimates are investigated.

Knoth addresses a combined CUSUM–Shewhart mean control chart. He compares several numerical techniques to calculate the average run length (ARL) of this chart for both the one-sided and the two-sided case. To obtain the ARL, an integral equation is numerically solved. Assuming normally distributed samples, he employs new numerical techniques such as collocation to determine the ARL. These procedures enable an accurate calculation of this quantity.

In the chapter of Polunchenko, a numerical study is provided to examine the effect of a headstart on the performance of a Shiryaev–Roberts (SR) chart for the mean of an independent normal process. The main result of the author consists in the observation that a fast initial response SR with a carefully designed optimal headstart is not just quicker to react to an initial out-of-control situation but is also nearly the fastest uniformly over the potential change point positions.

Morais and Knoth consider the problem of monitoring the traffic intensity of a queuing system. Their aim is to detect an increase or decrease in the traffic intensity. In their contribution, they focus on a single server queue and discuss the M/G/1, GI/M/1, and GI/G/1 systems in more detail. Three control statistics are proposed, all having a Markovian structure. The intention of the authors is to obtain ARL-unbiased charts, i.e., charts in which the out-of-control ARL is always smaller than or equal to the in-control ARL. They derive the ARLs of the proposed charts and use the Markov chain approach to calculate these quantities.

Tang and Gan investigate an application of SPC in public health. In this chapter, a risk-adjusted EWMA charting procedure for monitoring surgical procedures is developed. It is based on two or more outcomes. The monitoring statistic is obtained by combining log-likelihood ratio statistics for detecting improvement and deterioration. The properties of the procedure are determined, and the procedure is compared with the risk-adjusted CUSUM chart using a surgical data set. The risk-adjusted EWMA procedure turns out to be an attractive alternative because of its good performance and ease of interpretation.

Saniga, Davis, and Lucas compare visitor data for two websites generated by a variety of commercial analytics packages and discuss the issues of data accuracy, consistency, and unavailability of important measures. How control charts can be used for Phase I and Phase II analyses to monitor the website effectiveness is described. Since the number of visitors is a count variable, the c chart and the CUSUM chart for counts are applied. Another interesting quantity of website effectiveness is the bounce rate which can be monitored using a p chart or a binomial CUSUM chart.

Epprecht, Aparisi, and Ruiz discuss a problem from multivariate statistical process control. They consider the case in which some quality characteristics are more expensive and/or more difficult to measure than others. The authors make use of the recently introduced variable dimension approach. This means that the "non-expensive" variables are monitored, whereas only if there is a hint of an out-of-control signal are the "expensive" quantities measured. They review and compare

several variants of the approach that may lead to significant savings in terms of production costs.

Sparks and Chakraborti describe a sewerage treatment plant where it is of relevance to monitor the dispersions of certain environmental quantities that have skewed distributions. In their chapter they focus on bivariate data. Four different monitoring strategies are discussed. One control procedure is based on the Box-Cox transformation, two approaches are based on robust regression, and one attempt uses the concept of data depth. The introduced schemes are compared within a simulation study for many different skewed distributions.

The aim of the chapter of Nishina, Kawamura, Okamoto, and Takahashi is to monitor causal relationships among variables. Such results are important to protect a system against cyberattacks, for example. The authors propose a method of diagnosis for isolating an unusual causal relationship in a process causal model. The nearest unusual model is identified by utilizing the Mahalanobis distance between some supposed unusual models and the data to indicate the out-of-control region in Q charts. The proposed method is analyzed within a simulation study for two examples of causal models.

In many complex processes, a large number of variables are monitored simultaneously. The chapter of Yashchin addresses multistage data where very large amounts of data are collected at various process stages. In practice, it is not only necessary to detect a change in the production process as early as possible; rather, a methodology to diagnose the stage that is the most likely culprit is also needed. He describes the quality early warning system for variables (QEWSV) data and discusses an example related to monitoring the characteristics of tape storage devices. The detection algorithm is based on the CUSUM–Shewhart methodology.

Weiß considers monitoring of categorical time series. A brief survey of approaches for modeling and analyzing serially dependent categorical processes is given. Two scenarios of monitoring are discussed: a sample-based approach, in which the dependence within the samples has to be considered, and a continuous monitoring approach, in which the dependence between successive observations has to be taken into account for chart design. For both cases, appropriate control charts are proposed and their performance is investigated through simulations.

Hryniewicz and Kaczmarek-Majer address monitoring of short time series. In their chapter, the case in which the information from the available data is insufficient for good estimation of the model is considered. A new method for the construction of Shewhart control charts for residuals is proposed. The inspiration for the introduced methodology comes from the concept of Bayesian model averaging. The novelty of the proposed XWAM (X-weighted average model) control chart is the usage of computational intelligence methodology for the construction of alternative models and the calculation of their weights.

In Lazariv and Schmid's contribution, different approaches for monitoring nonstationary multivariate time series are discussed. Control charts for a very general family of time series are introduced. It is assumed that the in-control process is a multivariate state-space process. The out-of-control process is modeled by a general change point model which includes shifts and drifts. Using the likelihood

ratio, the sequential probability ratio and the Shiryaev–Roberts approach control charts with a reference value are derived. Moreover, generalized control schemes without reference values are also obtained. Using various performance criteria, the introduced control charts are compared via a simulation study.

Part II: Design of Experiments

Montgomery and Silvestrini review several important new developments regarding the field of design of experiments. The role of designed experiments in innovation is examined, and new developments and applications of the methods are discussed. Design of experiments provides a structured methodology for experimentation, and this can greatly aid creative thinking. The authors emphasize that statistical methodology is an important aid in the innovative process and should be employed to obtain improved results.

The precision of measurement results can be quantified by using variance components of random effect models. In Yasui and Ojima's chapter, the measurement results are statistically modeled using a nested design. Although balanced nested designs are widely used, staggered nested designs, which are one type of unbalanced nested design, have the statistical advantage that the degrees of freedom in all stages except for the top stage are equal. The authors identify D-optimal three-stage unbalanced nested designs for the determination of measurement precision.

Part III: Related Areas

Wilrich addresses Type I censored sampling plans for inspection by variables which have the advantage that the test time is fixed in advance. The lifetime of the product is assumed to obey a Weibull distribution with unknown parameters. The considered sampling plan is based on the logarithm of the lifetime. These quantities follow a Gumbel distribution. The sampling plan uses maximum likelihood estimators of the parameters of a Gumbel distribution. In a simulation study, the operating characteristic function of the sampling plan is analyzed under various conditions.

Yamamoto and Jin note that an important problem in assessing the risk of failure events of a system is the choice of the timescale. Although there should be genuine timescales for each failure phenomenon, the field data may not be sufficient to provide evidence for them. There are many uncontrollable factors in the field. Their chapter attempts to build a bridge between two useful approaches: alternative timescales and cumulative-exposure models by assuming stationarity of the increments of these measurements within a system.

Bayesian approaches are increasingly popular within the statistics community. However, they do not seem to be widely applied within the field of industrial statistics. Vining examines some of the basic reasons for this lack of applications.

He reviews Box's perspective on the scientific method and discovery and Deming's concepts of analytic versus enumerative studies. The chapter addresses applications of Bayesian methods to process monitoring and experimental design and analysis. The author examines the use of Bayesian approaches and concludes that in some cases, the appropriate tools are Bayesian if they are used with care.

Collani critically discusses the quality of the current practice of performing medical statistics. He analyzes the standards of laboratories and demonstrates the requirements using a clinical trial. Based on several examples, he tries to illustrate that the quality of medical statistics is not good and that this is also due to the statistical methodology and statistical methods themselves. He casts doubt on the practical usage of significance tests in medicine.

The level of a workshop on *Intelligent Statistical Quality Control* is determined by the quality of its chapters. We believe that this volume truly represents the frontiers of statistical quality control. The editors would like to express their deep gratitude to the members of the scientific program committee, who carefully invited researchers from around the world and the reviewers of all submitted chapters:

Sven Knoth, Germany
Fadel Megahed, USA
Yoshikazu Ojima, Japan
Wolfgang Schmid, Germany
Peter-Th. Wilrich, Germany
William H. Woodall, USA
Kwok L. Tsui, Hong Kong
Emmanuel Yashchin, USA

Moreover, we thank Springer, Heidelberg, for the continuing collaboration.

Hamburg, Germany — Sven Knoth
Frankfurt (Oder), Germany — Wolfgang Schmid
September 2017

To the Memory of Elart von Collani

Elart von Collani passed away on February 25, 2017. He struggled with a serious disease for several years. Elart participated in the first Workshop on Intelligent Statistical Quality Control in Berlin, 1980. He regularly participated in the workshops and was the organizer of the sixth Workshop in Würzburg, 1998. After missing the Sydney workshop in 2013, we were very glad to welcome him in Hamburg in 2016.

Elart was a highly respected colleague. He was very creative, had many new ideas, critically reflected upon existing statistical methodologies, and was able to find new paths for statistics.

We will miss him both as a researcher and as a very friendly and kind colleague.

Contents

Contributors

Francisco Aparisi Departamento de Estadística e I.O. Aplicadas y Calidad, Universidad Politécnica de Valencia, Valencia, Spain

Giovanna Capizzi Department of Statistical Sciences, University of Padua, Padua, Italy

Subha Chakraborti Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa, AL, USA

Darwin Davis Department of Business Administration, Alfred Lerner College of Business and Economics, University of Delaware, Newark, DE, USA

Eugenio K. Epprecht Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil

Fah Fatt Gan Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore

Olgierd Hryniewicz Systems Research Institute, Polish Academy of Sciences, Warszawa, Poland

Lu Jin University of Electro-Communications, Tokyo, Japan

Katarzyna Kaczmarek-Majer Systems Research Institute, Polish Academy of Sciences, Warszawa, Poland

Hironobu Kawamura Nagoya Institute of Technology, Showa-ku, Nagoya, Japan

Sven Knoth Institute of Mathematics and Statistics, Department of Economics and Social Sciences, Helmut Schmidt University Hamburg, Hamburg, Germany

Yesim Koca Hacettepe University, Department of Industrial Engineering, Beytepe-Ankara, Turkey

Taras Lazariv Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany

James M. Lucas James Lucas and Associates, Newark, NJ, USA

Guido Masarotto Department of Statistical Sciences, University of Padua, Padua, Italy

Douglas C. Montgomery School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

Manuel Cabral Morais CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

Ken Nishina Nagoya Institute of Technology, Showa-ku, Nagoya, Japan

Yoshikazu Ojima Tokyo University of Science, Yamazaki, Noda, Chiba, Japan

Kosuke Okamoto Nagoya Institute of Technology, Showa-ku, Nagoya, Japan

Aleksey S. Polunchenko Department of Mathematical Sciences, State University of New York at Binghamton, Binghamton, NY, USA

Omar Ruiz ESPOL, Polytechnic University, Escuela Superior Politécnica del Litoral, Facultad de Ciencias de la Vida, Guayaquil, Ecuador

Erwin Saniga Department of Business Administration, Alfred Lerner College of Business and Economics, University of Delaware, Newark, DE, USA

Wolfgang Schmid Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany

Rachel T. Silvestrini Department of Industrial and Systems Engineering, Rochester Institute of Technology, Rochester, NY, USA

Ross Sparks CSIRO Australia, Data61, Sydney, Australia

Tatsuya Takahashi Nagoya Institute of Technology, Showa-ku, Nagoya, Japan

Xu Tang Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore

Murat Caner Testik Hacettepe University, Department of Industrial Engineering, Beytepe-Ankara, Turkey

Özlem Müge Testik Hacettepe University, Department of Industrial Engineering, Beytepe-Ankara, Turkey

G. Geoffrey Vining Department of Statistics, Virginia Tech, Blacksburg, VA, USA

Elart von Collani University Würzburg, Würzburg, Germany

Christian H. Weiß Helmut Schmidt University, Department of Mathematics and Statistics, Hamburg, Germany

Peter-Th. Wilrich Institut für Statistik und Ökonometrie, Freie Universität Berlin, Berlin, Germany

Watalu Yamamoto University of Electro-Communications, Tokyo, Japan

Emmanuel Yashchin IBM, Thomas J. Watson Research Center, Yorktown Heights, NY, USA

Seiichi Yasui Tokyo University of Science, Yamazaki, Noda, Chiba, Japan

Part I
Statistical Process Control

Phase I Distribution-Free Analysis with the R Package `dfphase1`

Giovanna Capizzi and Guido Masarotto

Abstract Phase I distribution-free methods have received an increasing attention in the recent statistical process monitoring literature. Indeed, violations of distributional assumptions may largely degrade the performance and sensitivity of parametric Phase I methods. For example, the real false alarm probability, i.e., the probability to declare unstable a process that is actually stable, may be substantially larger than the desired value. Thus, several researchers recommend to test the shape of the underlying IC distribution *only after* process stability has been established using a distribution-free control chart. In the chapter, we describe the R package `dfphase1` which provides an implementation of many of recently suggested Phase I distribution-free methods. Indeed, because of the relatively high computational complexity of some of these methods, we believe that their diffusion can be helpfully encouraged supporting practitioners with an easy-to-use dedicated software. The use of the package is illustrated with real data from an oil refinery.

Keywords Change point · Control charts · Nonparametric · Statistical process monitoring

1 Introduction

Control charts are well known techniques used in statistical process monitoring (SPM) to establish whether a process is "in-control" (IC) or "out-of-control" (OC), i.e. whether it is operating under random or assignable causes of variations that need to be detected as soon as possible (Montgomery 2009; Qiu 2013). Control charts are conceived and designed differently according to the full or partial knowledge on the underlying IC process distribution. When a full knowledge on process distribution, and on all its parameters, is available, data are prospectively charted in Phase II

G. Capizzi · G. Masarotto (✉)
Department of Statistical Sciences, University of Padua, Padua, Italy
e-mail: giovanna.capizzi@unipd.it; guido.masarotto@unipd.it

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_1

for promptly detecting an OC situation. However, whether either the underlying IC distribution or some parameters of that distribution are unknown, a Phase I analysis is conducted to characterize process variation under stable conditions and estimate a set of accurate control limits for on-line monitoring in Phase II.

Phase I control charts aim to test retrospectively whether observations on a univariate (or multivariate) quality characteristic X, collected in m subgroups each of size $n \geq 1$, all come from a common IC distribution or from a distribution whose parameters have changed. In recent years, attention and emphasis for Phase I analysis have progressively grown among researchers and users because of some critical aspects and issues of SPM that, when not appropriately faced and addressed in Phase I, can seriously degrade the performance of Phase II control charts (see, for example, Chakraborti et al. 2009; Jones-Farmer et al. 2014; Capizzi 2015). One of the most challenging tasks in Phase I is evaluating process stability with respect to a specified parametric model. Indeed, the uncertainty on the correct specification of the underlying IC model makes parametric control charts quite unpredictable in terms of their ability to distinguish true OC points from IC points coming from a misspecified IC process distribution. Hence, when the specification of a correct IC statistical model is a point of concern, the identification of OC conditions without any a priori selection of a model can be more useful to practitioners. For all these reasons, researchers have recently stressed the importance of using distribution-free control charts in Phase I (see for example Jones-Farmer et al. 2009; Jones-Farmer and Champ 2010; Graham et al. 2010; Human et al. 2010; Bell et al. 2014; Capizzi and Masarotto 2013b; Cheng and Shiau 2015; Capizzi 2015; Woodall 2017; Capizzi and Masarotto 2017).

Despite of their documented effectiveness in Phase I, there is still some reluctance to practically apply distribution-free procedures, because they are based on control statistics not very familiar to users and because their practical design and implementation can show some mathematical and/or computational complexity. The availability of an easy-to-use software implementing recent nonparametric Phase I proposals can make their usage much more appealing to practitioners.

Thus, in this chapter we illustrate the R package `dfphase1` implemented to perform the Phase I analysis of either univariate or multivariate data. The package complements the functionalities offered by other `R` packages such as `qcc` (Scrucca 2004), `changepoint` (Killick and Eckley 2014), `cpm` (Ross 2015), and `spc` (Knoth 2016). The `dfphase1` package covers the design and use of different distribution-free procedures recently proposed for testing the stability of process location and variation. It also implements the combination of some distribution-free Phase I methods, originally conceived to test for the stability of only one of these two process parameters. The `R` package also allows the distribution-free design of some univariate and multivariate methods developed for the parametric framework. All methods implemented in the package attain a desired false alarm probability (FAP) with no assumption on the underlying probability distribution of quality characteristics. Coherently to the "standard framework" handled in the SPM literature, in this chapter we assume that (1) the number of data points is larger than number of the variables, and (2) when the process is IC, the observation vectors are independent and identically distributed. Extensions to the high-dimensional

and/or time-dependent framework will require further research. Notwithstanding these limitations, the implemented methods provide a distribution-free design of several Phase I procedures frequently used in many practical situations.

The chapter is organized as follows. In Sect. 2, we briefly argue why the SPM literature has recently been paying increasing attention to a distribution-free approach to Phase I analysis. Then, in Sect. 3, the main approaches to the distribution-free Phase I analysis of univariate and multivariate data are shortly reviewed, also outlining some possible drawbacks in their design and implementation, above all in the multivariate framework. Some details on the `dfphase1` package are given in Sect. 4. In Sect. 5, an example is discussed. Some concluding remarks are given in Sect. 6.

2 Why Distribution-Free Methods in Phase I?

Performances of Phase I methods are usually evaluated in terms of alarm probabilities. In particular, the control limits of Phase I control charts are determined so that, at least approximately, the FAP, i.e., the overall probability of giving at least one false alarm, attains a nominal value. Control limits are often computed under the assumption of a known underlying probability distribution, such as normal, exponential, gamma, etc. However, as anticipated in the Introduction, in Phase I stability with respect to a parametric model is often tested when a little information is available to validate distributional assumptions. A misspecification of the underlying IC process distribution may result in inflated false alarm probabilities but also in an incorrect classification of an observation as an "outlier" or "out-of-control" point. Indeed:

1. The attained FAP can be very different from the nominal value when the real process distribution deviates from the assumed parametric model. For example in the univariate case, when $m = 50$ and $n = 5$, the attained FAP of a retrospective Shewhart $\overline{X}$-S control chart, designed to give a FAP equal to 0.05 under the normality assumption, is equal to 0.528 and 0.749 when Phase I data actually come from a Student's t_5 and an Exponential, respectively. The IC performance is even more degraded in the multivariate framework. For example a T^2 control chart designed to give a FAP equal to 0.05 for multivariate normally distributed data, provides an attained FAP equal to 0.72 ($m = 50$, $n = 5$) and 0.97 (m=100 and $n = 5$) when is applied to data coming from a five dimensional Student's t_3. Even when more Phase I data are available ($m = 100$ and $n = 10$), the attained FAP reaches an unacceptably high value equal to 0.87.
2. On the other hand, the classification of an observation as an "outlier" strictly depends on the strength of its evidence against the model chosen as more appropriate for representing a stable process. The standard Phase I practice, consisting in iteratively identifying, removing OC points and recomputing control limits, leads to a "reference" sub-sample easily consistent with the hypothesized parametric model but not necessarily representative of the true

stable underlying probability distribution (see Capizzi (2015), for an example and additional discussions).

3 Distribution-Free Phase I Control Charts: A Brief Review

A distribution-free (or nonparametric) control chart is defined in terms of its IC behavior. If the IC properties are the same for (at least) all continuous distributions, the resulting control charts are called distribution-free (see Chakraborti et al. (2001), Chakraborti (2007), Chakraborti et al. (2009), and Chakraborti (2011), for some reviews covering much of the recent SPM nonparametric literature).

Two possible approaches can be followed for implementing a distribution-free Phase I analysis.

1. Plot distribution-free control statistics, such as mean ranks, sign statistics or median-based statistics. This approach has been adopted for example by Jones-Farmer et al. (2009), Jones-Farmer and Champ (2010) and Graham et al. (2010). When compared with control charts based on standard control statistics, such as the standard Shewhart-type $\overline{X}$ and S control charts, this approach can produce an inferior performance in the normal or nearly normal case which, however, is compensated by an efficiency gain when the process distribution strongly deviates from the normal assumption. A practical disadvantage associated with this approach is the need to learn and use new summary statistics not very familiar to users. Further, it is difficult to generalize distribution-free statistics, such as those based on the ranks, to the multivariate framework. Indeed, such a generalization only involves the family of elliptical IC probability distributions (see Oja (2010) for a general discussion and Bell et al. (2014) and Cheng and Shiau (2015) for two specific proposals).
2. Plot well-known control statistics, such as the subgroup means or the Hotelling T^2s, but modify the control limits to account for a possible non-normality of the process distribution. According to this approach, the distribution-free design of control charts does not require, also in the multivariate framework, any specification of the underlying process distribution. The control limits, computed via a resampling method (booststrap, permutation, etc.) are able, exactly or approximately, to guarantee the desired FAP both in the normal and nonnormal scenarios. The limits can be quickly computed also using a low-end personal computer. In particular, in `dfphase1`, we mainly consider the permutation approach (Pesarin 2001; Good 2005; Lehmann and Romano 2005) since it is able to exactly achieve a prescribed FAP regardless of the underlying process distribution, at least for independent and identically distributed observations. Furthermore, at least in many practical scenarios, there is no performance loss in using the permutation-based limits. Indeed, a Monte Carlo study has shown that the considered approach enjoys an *"oracle property"*, i.e., the resulting schemes perform at least as well as if the shape of the process distribution were known a priori and used to compute the control limits (e.g. Capizzi and Masarotto 2013a).

4 The dfphase1 Package

Table 1 summarizes the Phase I methods implemented in the package dfphase1. The package is written in R. The more computational demanding procedures have been written in C++ using the Rcpp interface (Eddelbuettel 2013).

Table 1 Phase I methods implemented in package dfphase1

1. Univariate methods
   a. Shewhart-type control charts
      (i) $\overline{X}$ control chart (Montgomery 2009, chapter 6), with permutation-based control limits
      (ii) S control chart (Montgomery 2009, chapter 6), with permutation-based control limits
      (iii) Rank-based control chart for location (Jones-Farmer et al. 2009)
      (iv) Rank-based control chart for scale (Jones-Farmer and Champ 2010)
      (v) Balanced combination of (i)–(ii) or (iii)–(iv) for simultaneously testing location and scale and giving a desired overall FAP (Capizzi 2015)
   b. Methods for change-point detection
      Sullivan and Woodall (1996) control chart, also adapted to subgrouped data, with permutation-based limits (see also Qiu 2013, chapter 6)
   c. Hybrid
      RS/P method (Capizzi and Masarotto 2013b)
2. Multivariate methods
   a. Shewhart-type control charts
      (i) Hotelling T^2 control chart, with permutation-based control limits (Montgomery 2009, chapter 11, equation 11.19)
      (ii) Normal likelihood control chart for monitoring process variability, with permutation-based control limits (Montgomery 2009, chapter 11, equation 11.34)
      (iii) Analogous of (i) and (ii) but based on the marginal ranks (Lung-Yut-Fong et al. 2011), spatial signs or ranks (Oja 2010) and signed ranks (Hallin and Paindaveine 2004, 2008)
      (iv) Balanced combination of the previous Shewhart-type schemes
   b. Methods for change-point detection
      (i) Sullivan and Woodall (2000) control chart, also adapted to subgrouped data, with permutation-based limits (see also Qiu 2013, chapter 6 and 7)
      (ii) Analogous control charts based on the marginal ranks (Lung-Yut-Fong et al. 2011), spatial signs or ranks (Oja 2010) and signed ranks (Hallin and Paindaveine 2004, 2008)
   c. Hybrid
      A model identification approach, based on the forward search and the LASSO algorithms, for detecting multiple location shifts with arbitrary patterns (Capizzi and Masarotto 2017)

Control statistics in Table 1 can be based on several estimates of the common process parameters. For example, as illustrated in Sect. 5, the multivariate Shewhart control chart can be based on the classical estimates of the multivariate location and variability (e.g. Montgomery 2009, equations 11.17a-c) but also on the highly robust minimum covariance determinant (MCD) estimate (Maronna et al. 2006; Jensen et al. 2007).

The package handles applications in the standard univariate and multivariate framework. However, as shown in Sect. 5, it can also be used in more complex situations where not necessarily the original observations have to be monitored, but some of their features, such as principal components, model parameters, etc. Nevertheless, in order to guarantee the validity of the implemented Phase I procedures, the "extraction" must be equivariant under a permutation of the original data.

The choice of the implemented methods reflects the idea that the detection of location and/or scale changes is of particular interest in most applications. Observe, that `dfphase1` also allows to implement two simultaneous control charts originally designed to detect separately location and scale shifts. As discussed by Capizzi (2015), the control limits of the two charts are adjusted so that

(a) The overall FAP is guaranteed, i.e.,

$$\text{Prob}\big(\text{one or both of the two charts give a false signal}\big) = FAP_0.$$

where FAP_0 is a desired value of the FAP.

(b) The FAP is evenly balanced between the two charts, i.e.,

$$\text{Prob}\big(\text{first chart gives a false signal}\big) = \text{Prob}\big(\text{second chart gives a false signal}\big).$$

The `R` functions are easy to use. The only needed arguments are the Phase I data, organized as follows.

- *Univariate control charts:* an $n \times m$ matrix, where n and m are the size of each subgroups and the number of subgroups, respectively. A vector of length m is accepted in the case of individual data, i.e., when $n = 1$.
- *Multivariate control charts:* a $p \times n \times m$ array, where p denotes the number of monitored variables. A $p \times m$ matrix is accepted in the case of individual data.

All the functions in `dfphase1` compute the control limits using Monte Carlo simulations. For the implemented Phase I methods, the default number of Monte Carlo replications has been differently set to provide an high accuracy of the attained FAP. However, users can run a different number Monte Carlo replications changing the default value of the `L` argument. For example, for the computation of the permutation-based control limits, users can set `L`$\propto (n \times m)!$ to run a number of Monte Carlo replications proportional to the number of permutations.

5 An Example

5.1 Description of the Data

To illustrate the use of the package, we consider a dataset of 564 near-infrared (NIR) gasoline spectra measured at wavelengths from 900 to 1700 nm (in 2 nm intervals). In particular, 12 gasoline samples have been collected each day for a period of 47 (consecutive) days in an oil refinery. The command

```
> NIR <- as.matrix(read.table("NIR"))
```

loads in memory a matrix, named NIR, of dimension

```
> dim(NIR)
[1] 564 401
```

containing the spectra (one for each row). Note that values are the logarithms of the absorbances.

Following the suggestions of the production engineers, each day is handled as a rational subsample. Hence, we assume that the dataset comprises

```
> m <- 47
```

subgroups of observations, each of size

```
> n <- 12
```

Figure 1a, b shows the plot of all the 564 spectra and of those collected during the first day, respectively. They have been obtained with the following commands.

```
> library(lattice)
> wavelength <- seq(900,1700,by=2)
> xyplot(NIR~rep(wavelength,rep(564,401)),
+        groups=rep(1:564,401), type="l",
```

(continued)

```
+        xlab="nm",ylab=expression(log(Absorbance)))
> samples <- reorder(rep(1:12,401),
+                    rep(c(9:12,5:8,1:4),401))
> xyplot(NIR[1:12,]~rep(wavelength,rep(12,401))|samples,
+        type="l",
+        xlab="nm", ylab=expression(log(Absorbance)))
```

Here, we are clearly facing a profile monitoring problem. As often done with functional data (see Ramsay and Silverman (2005), and Ramsay et al. (2009), for a

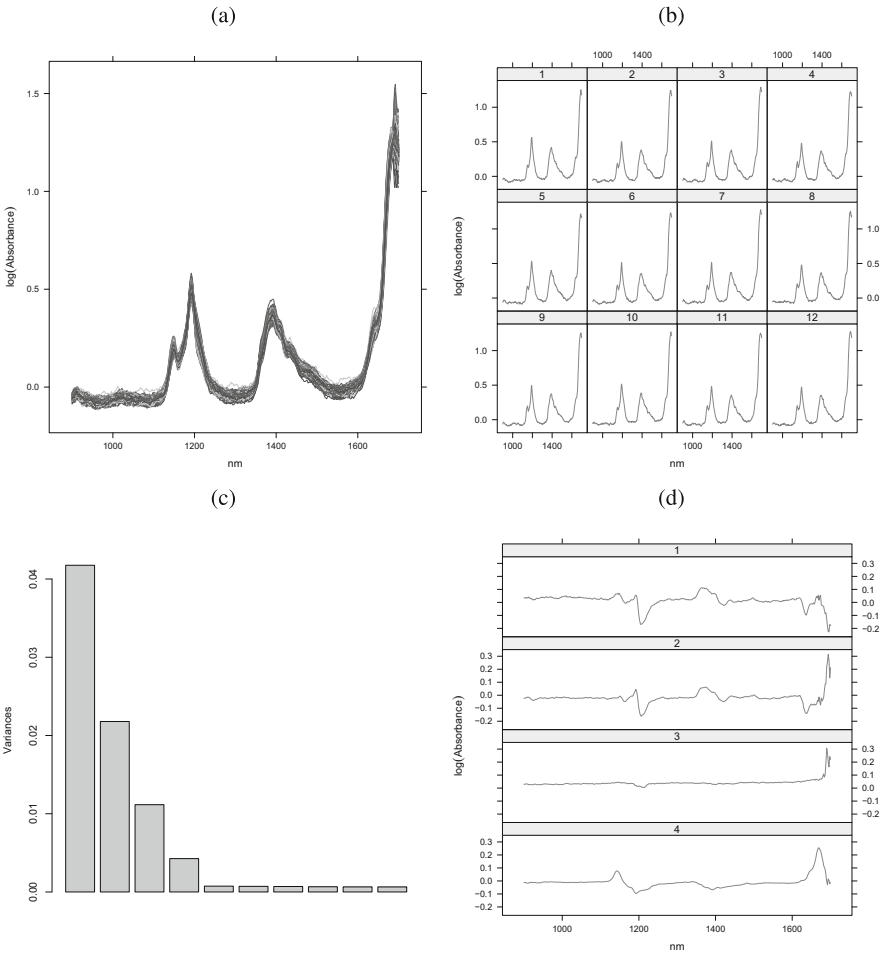


Fig. 1 Gasoline NIR spectra: (**a**) all the 564 spectra (superimposed); (**b**) 12 spectra collected during the first day; (**c**) scree plot; (**d**) first four eigenvectors

general discussion; Yu et al. (2012), for a specific application to SPM), we reduce the dimensionality of data via principal component analysis (PCA). Observe that

1. When *NC* components are retained, PCA provides the "regression-like" representation of the *i*th NIR spectrum

$$\text{NIR}_i(\text{nm}) = \mu(\text{nm}) + \sum_{j=1}^{NC} x_{i,j}\xi_j(\text{nm}) + r_i(\text{nm})$$

   where $x_{i,j}$ is the *j*th principal component, and $\text{NIR}_i(\text{nm})$, $\mu(\text{nm})$, $\xi_j(\text{nm})$ and $r_i(\text{nm})$ are the logarithm of the absorbance, the log-absorbance mean, the *j*th eigenvector and the residual term at the wavelength (nm). Hence, because for profile data, such as the gasoline spectra, the eigenvectors are relatively smooth functions, testing for the stability of the principal components $x_{i,j}$ over time is similar to testing for the stability of the coefficients of a (mixed) regression model describing the profiles.
2. At least in its standard implementation, the principal components are equivariant under a (row) permutation of the original dataset. Hence, permutation- and rank-based Phase I methods maintain their distribution-free properties.

   The scree plot of the `NIR` data, obtained with the command

```
> plot(pca <- prcomp(NIR),main="")
```

and displayed in Fig. 1c, suggests to retain the first 4 principal components. Figure 1d, displaying the corresponding eigenvectors, can be obtained using the following commands

```
> eigv <- gl(4,401,labels=4:1)
> xyplot(pca$rotation[,1:4]~rep(wavelength,4)|eigv,
+        type="l", layout=c(1,4),
+        xlab="nm",ylab=expression(log(Absorbance)))
```

   As often done in SPM, we also retain the additional variable

$$Q_i = \frac{1}{401} \sum_{nm=900,\ldots,1700} |r_i(\text{nm})|,$$

which reflects the size of the residual term. The following code "extracts" the first four principal components, computes "Q" and, as required by `dfphase1`,

organizes the results in a $5 \times n \times m$ array,

```
> fitted <- t(tcrossprod(pca$rotation[,1:4], pca$x[,1:4])
+             +pca$center)
> r <- NIR-fitted
> x <- array(rbind(t(pca$x[,1:4]),rowMeans(abs(r))),
+            c(5,n,m))
> dimnames(x) <- list(c(paste("PCA",1:4,sep=""),"Q"),
+                     NULL,NULL)
```

5.2 Phase I Analysis

The package can be loaded during an R session using

```
> library(dfphase1)
```

The mshewhart function can be used to obtain different multivariate Shewhart control charts (see Table 1). When the data array is the only argument,

```
> u <- mshewhart(x)
```

the function provides the graph displayed in Fig. 2a. The two panels show the standard control statistics used for monitoring the stability of the mean and dispersion of a multivariate normal distribution, respectively (see Montgomery 2009, equations 11.19 and 11.34). However, the control limits

```
> u$limits
[1] 25.802370  6.715327
```

are computed by permutation so that the desired FAP is guaranteed for each multivariate distribution. In dfphase1, the default value of the FAP is 5%, but it can be easily changed using the FAP argument.

The lower panel in Fig. 2a suggests that the dispersion was probably OC during days 33, 35, 36 and 37. Then, observe that control statistics of days 32 and 34 are below the limit but larger than the values of the other "in-control" days (see the lower panel). Indeed, when the observations collected on days 33, 35, 36 and 37 are

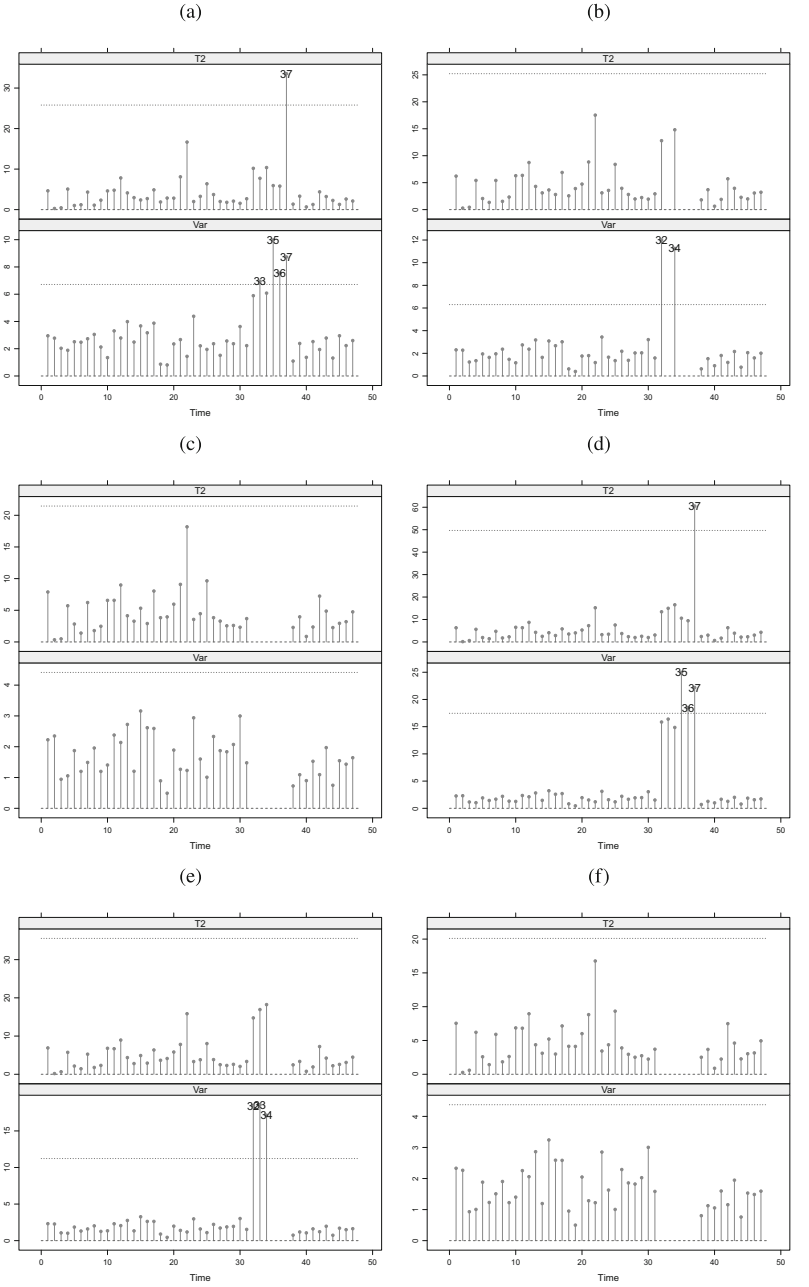


Fig. 2 Combination of an Hotelling T^2 control chart and a control chart for monitoring the stability of the covariance matrix, based on standard and MCD estimates of location and scatter parameters. Standard estimates: (**a**) all the data; (**b**) days 33, 35, 36, 37 deleted; (**c**) days from 32 to 37 deleted. MCD estimates: (**d**) all the data; (**e**) days 35, 36, 37 deleted; (**f**) days from 32 to 37 deleted

deleted, days 32 and 34 are flagged as OC for the dispersion. See Fig. 2b which is produced by the following command

```
> mshewhart(x,subset=-c(33,35:37))
```

As shown by Fig. 2c, obtained with the following command

```
> mshewhart(x,subset=-(32:37))
```

no other subgroup is flagged as OC when observations from day 32 to day 37 are not considered.

In `dfphase1`, alternative estimates of process parameters can be used adding the optional argument `loc.scatter` in the call to `mshewhart`. Figure 2d, e, and f, produced using the commands

```
> mshewhart(x,loc.scatter="MCD")
> mshewhart(x,subset=-(35:37),loc.scatter="MCD")
> mshewhart(x,subset=-(32:37),loc.scatter="MCD")
```

show that days from 32 to 37 are also flagged as OC when the standard estimates of multivariate location and dispersion (e.g. Montgomery 2009, equations 11.17a-c) are replaced with the high-breakdown MCD estimates (e.g. Maronna et al. 2006, chapter 6).

The `mchangepoint` function can be used to detect a sustained shift. An example is provided by Fig. 3a produced using

```
> mchangepoint(x,score="Signed Ranks")
```

The upper panel shows the control statistics for verifying the presence of a shift either in the multivariate location or dispersion (see Sullivan and Woodall 2000; Qiu 2013). Since, for many days, the values are greater than the permutation-based control limit, the hypothesis of a stable process is rejected, and, in particular, the graph points to a possible shift on day 32. The middle and lower panel show the decomposition of the control statistic in the two parts due to changes in the location and dispersion, respectively (see Sullivan and Woodall 2000). For the NIR data, these diagnostic graphs clearly point to a shift in the dispersion. Note the optional

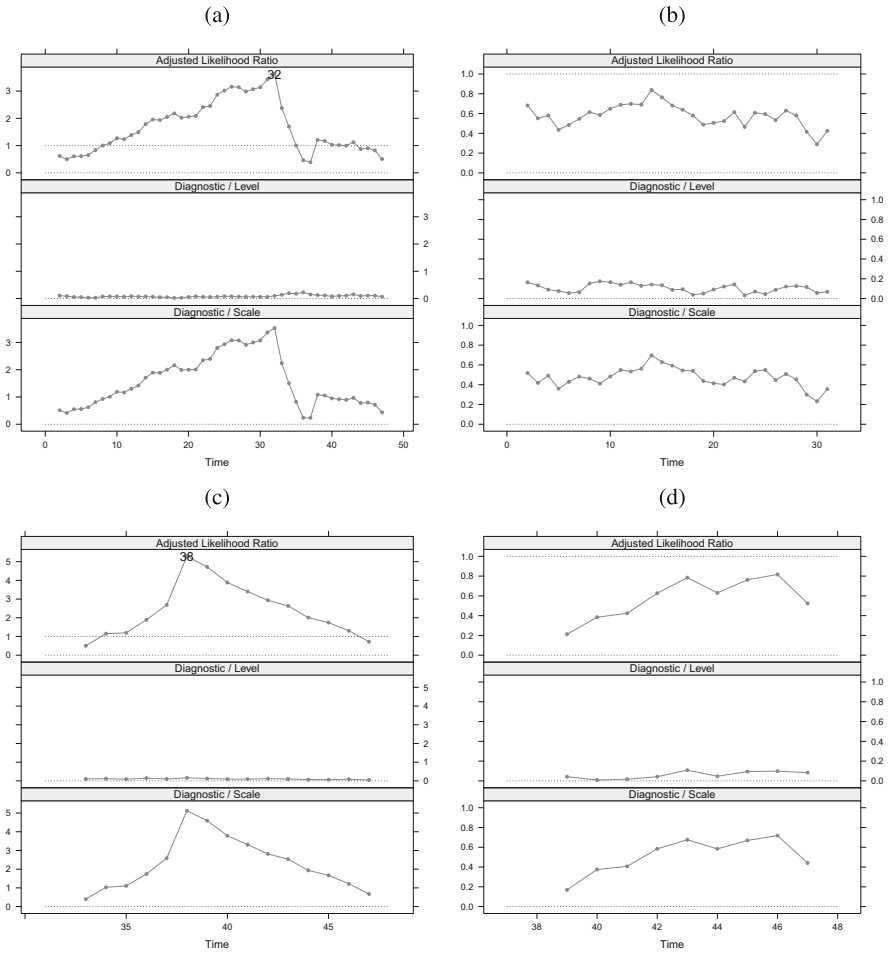


Fig. 3 Multivariate change-point detection: (**a**) all the data; (**b**) observations up to day 31; (**c**) observations after days 31; (**d**) observations after days 37

argument `score` which asks for a suitable multivariate rank transformation. An analogous argument can also be used for `mshewhart`.

Having divided the observations in two periods (before and after day 32), it is also useful to see if there is evidence of other shifts within these periods. Recursive application of `mchangepoint` suggests that the process was stable before days 32 (Fig. 3b) but that another dispersion shift was probably present starting on day 38 (Fig. 3c). No additional shift is detected after day 38 (Fig. 3d). The commands used to produce Fig. 3b–d are

```
> mchangepoint(x,subset=1:31,score="Signed Ranks")
> mchangepoint(x,subset=32:47,score="Signed Ranks")
> mchangepoint(x,subset=38:47,score="Signed Ranks")
```

The four principal components and the additional variable Q are expected to be (more or less) independent. Hence, for these data, process stability can be also assessed applying separately one or more univariate Phase I control charts to the five variables. For reasons of space, we will only show the application of three different schemes to the second principal component.

The `shewhart` function can be used to plot some univariate Shewhart-type control charts. Figure 4a–c shows the iterative application to the second principal component of the combined $\overline{X} - S$ control chart with control limits computed by permutation. Analogously, Fig. 4d and e illustrate the joint use of the two rank-based control charts proposed by Jones-Farmer et al. (2009) and Jones-Farmer and Champ (2010) for monitoring the univariate location and dispersion, respectively. These five Subfigures have been obtained using the commands

```
> shewhart(x[2,,])
> shewhart(x[2,,],subset=-c(32:33,35:37))
> shewhart(x[2,,],subset=-(32:37))
> shewhart(x[2,,],stat="Rank")
> shewhart(x[2,,],subset=-(32:37),stat="Rank")
```

The `rsp` function implements the RS/P method suggested by Capizzi and Masarotto (2013b). In particular, Fig. 4f can be obtained with the following command

```
> rsp(x[2,,])
```

Observe that no iterative use of `rsp` is usually needed since this method tries to detect multiple isolated and step changes.

Figure 4a–f indicates an increased variability in the second principal component for days from 32 to 37. Similar results have also been observed for the third and fourth principal components (but not for the first component and Q).

Globally speaking, the application of univariate and multivariate control charts signals the presence of an OC condition in the interval [32; 37]. The instability was attributed to a transitory malfunction of the automatic process adjustments. By deleting data collected in these days, the hypothesis of a stable process is accepted. Hence, observations up to day 31 and after day 37 can be used to study the process

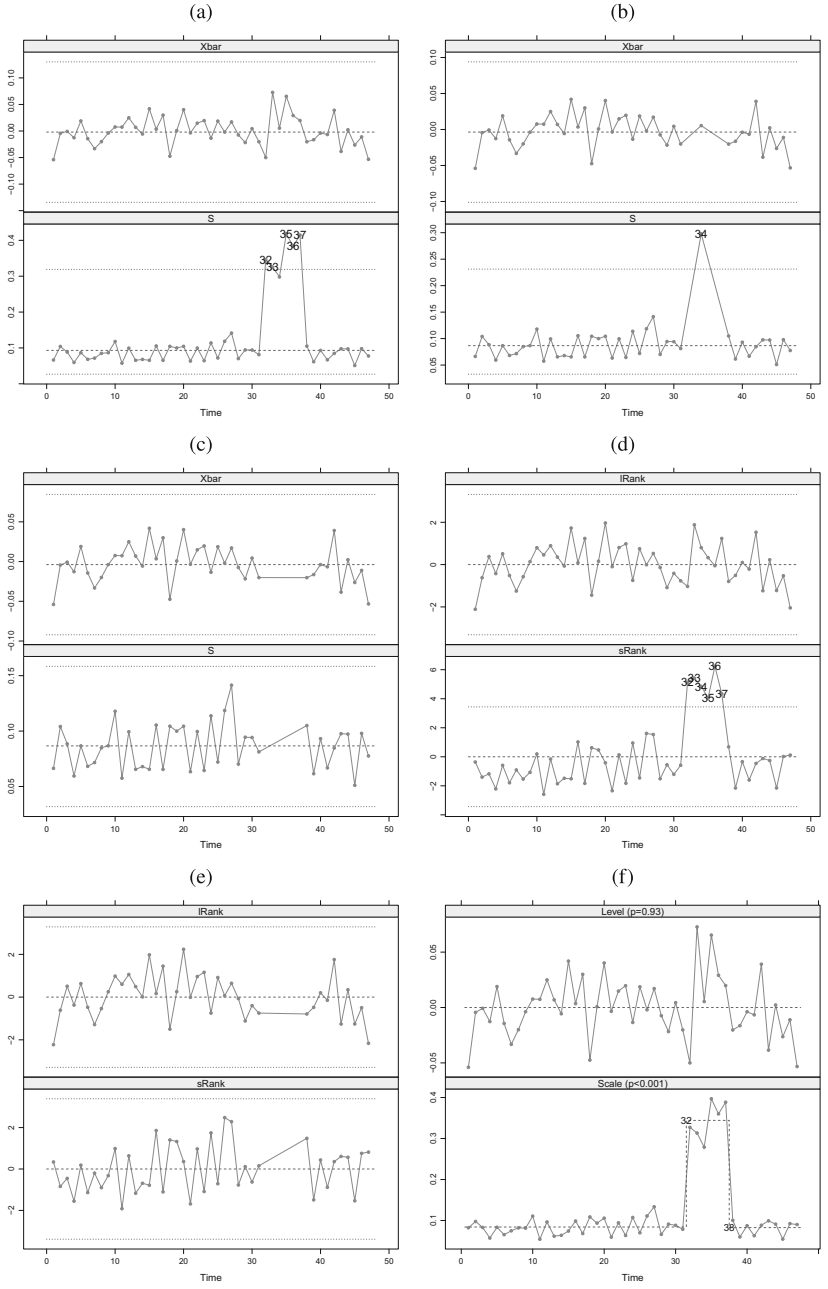


Fig. 4 Univariate control charts applied to the second principal component: (**a**) $\overline{X} - S$ control chart—all the data; (**b**) $\overline{X}-S$ control chart—days 32, 33, 35, 36, 37 deleted; (**c**) $\overline{X}-S$ control chart—days from 32 to 37 deleted; (**d**) rank-based control chart—all the data; (**e**) rank-based control chart—days from 32 to 37 deleted; (**f**) RS/P procedure

capability and design a Phase II control chart for prospectively monitoring the process.

6 Conclusions

We have illustrated the motivations and use of an `R` package developed for the distribution-free Phase I analysis of univariate and multivariate data. The package, which is available from the *Comprehensive R Archive Network* (https://cran.r-project.org/package=dfphase1), has been developed with the aim to facilitate and diffuse the use of distribution-free Phase I methods among practitioners.

Acknowledgements The authors thank the anonymous referee for her/his timely review and the helpful comments that improved the manuscript. This research was partially funded by UNIPD CPDA128413/12 grant.

References

Bell, R. C., Jones-Farmer, L. A., & Billor, N. (2014). A distribution-free multivariate Phase I location control chart for subgrouped data from elliptical distributions. *Technometrics, 56*(4), 528–538.

Capizzi, G. (2015). Recent advances in process monitoring: Nonparametric and variable-selection methods for Phase I and Phase II (with discussion). *Quality Engineering, 27*, 44–80.

Capizzi, G., & Masarotto, G. (2013a). Permutation-based design of the Phase I $\overline{X}$ control chart (with or without supplementary runs rules). In *3th International Symposium on Statistical Process Control*, July 9–11, 2013, University of Piraeus, Greece.

Capizzi, G., & Masarotto, G. (2013b). Phase I distribution-free analysis of univariate data. *Journal of Quality Technology, 45*(3), 273–284.

Capizzi, G., & Masarotto G. (2017). Phase I distribution-free analysis of multivariate data. *Technometrics, 59*(4), 484–495 (2017).

Chakraborti, S. (2007). Nonparametric control charts. In *Encyclopedia of statistics in quality and reliability* (pp. 415–429). New York: Wiley.

Chakraborti, S. (2011). Nonparametric (distribution-free) quality control charts. In *Encyclopedia of statistical sciences* (pp. 1–27). New York: Wiley.

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

Chakraborti, S., Van Der Laan, P., & Bakir, S. T. (2001). Non parametric control charts: An overview and some results. *Journal of Quality Technology, 33*, 304–315.

Cheng, C. R., & Shiau, J. J. H. (2015). A distribution-free multivariate control chart for Phase I applications. *Quality and Reliability Engineering International, 31*, 97–111.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York: Springer.

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). New York: Springer.

Graham, M. A., Human, S. W., & Chakraborti, S. (2010). A Phase I nonparametric Shewhart-type control chart based on the median. *Journal of Applied Statistics, 37*, 1795–1813.

Hallin, M., Paindaveine, D. (2004). Multivariate signed-rank tests in vector autoregressive order identification. *Statistical Science, 19*(4), 697–711.

Hallin, M., Paindaveine, D. (2008). Optimal rank-based tests for homogeneity of scatter. *The Annals of Statistics, 36*, 1261–1298.

Human, S. W., Chakraborti, S., Smit, C. F. (2010). Shewhart-type control charts for variation in Phase I data analysis. *Computational Statistics and Data Analysis, 54*, 863–874.

Jensen, W. A., Birch, J. B., & Woodall, W. H. (2007). High breakdown estimation methods for Phase I multivariate control charts. *Quality and Reliability Engineering International, 23*(5), 615–629.

Jones-Farmer, L. A., & Champ, C. W. (2010). A distribution-free Phase I control chart for subgroup scale. *Journal of Quality Technology, 42*, 373–387.

Jones-Farmer, L. A., Jordan, V., & Champ, C. W. (2009). Distribution-free Phase I control charts for subgroup location. *Journal of Quality Technology, 41*, 304–316.

Jones-Farmer, L. A., Woodall, W. H., Steiner, S. H., & Champ, C. W. (2014). An overview of Phase I analysis for process improvement and monitoring. *Journal of Quality Technology, 46*, 265–280.

Killick, R., Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software, 58*, 1–19.

Knoth, S. (2016). spc: Statistical Process Control – Collection of Some Useful Functions. R package version 0.5.3. https://CRAN.R-project.org/package=spc.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.

Lung-Yut-Fong, A., Lévy-Leduc, C., & Cappé, O. (2011). Homogeneity and change-point detection tests for multivariate data using rank statistics (preprint). arXiv:11071971.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics. Theory and methods*. Hoboken: Wiley.

Montgomery, D. C. (2009). *Introduction to statistical quality control* (6th ed.). New York: Wiley.

Oja, H. (2010). *Multivariate nonparametric methods with R. An approach based on spatial signs and ranks*. New York: Springer.

Pesarin, F. (2001). *Multivariate permutation tests: With applications in biostatistic*. New York: Wiley.

Qiu, P. (2013). *Introduction to statistical process control*. Boca Raton: Chapman & Hall/CRC Press.

Ramsay, J., & Silverman, B. (2005). *Functional data analysis*. New York: Springer.

Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB*. New York: Springer.

Gordon, R. J. (2015). Parametric and nonparametric sequential change detection in R: The cpm package. *Journal of Statistical Software, 66*, 1–20.

Scrucca, L. (2004). qcc: an R package for quality control charting and statistical process control. *R News, 4*, 11–17.

Sullivan, J. H., Woodall, W. H. (1996). A control chart for preliminary analysis of individual observations. *Journal of Quality Technology, 28*, 265–278.

Sullivan, J. H., Woodall, W. H. (2000). Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations. *IIE Transactions, 32*(6), 537–549.

Woodall, W. H. (2017). Bridging the gap between theory and practice in basic statistical process monitoring (with discussion). *Quality Engineering, 29*, 2–15.

Yu, G., Zou, C., Wang, Z. (2012). Outlier detection in functional observations with applications to profile monitoring. *Technometrics, 54*, 308–318.

Assessment of Shewhart Control Chart Limits in Phase I Implementations Under Various Shift and Contamination Scenarios

Murat Caner Testik, Christian H. Weiß, Yesim Koca, and Özlem Müge Testik

Abstract In Phase I implementations of control charts, the unknown parameters required in calculating the Phase II control limits are estimated. This retrospective study requires analyses of observations all at once to characterize a stable process by identifying and eliminating the observations corresponding to out-of-control process states. Hence, it is aimed to obtain an in-control reference set of observations for estimation. Nevertheless, there are many possibilities for parameter shifts and contaminations of observations due to out-of-control process states in industrial settings. As a simple but effective tool, Shewhart control charts are recommended in the literature for Phase I use.

In this study, the width of the Shewhart control limits is altered to assess the performance in Phase I implementations. Out-of-control states of a process are simulated using various mean shifts and contamination percentages of subgroups for normally distributed observations. Considering various combinations of the number of subgroups and the number of observations in each subgroup as control factors as well as the mean shifts and contamination percentages of the sample as uncontrollable factors, performance metrics such as true and false alarm percentages, number of iterations to complete estimation, and mean squared error of the parameter estimates are investigated. Robustness to uncontrollable factors through control limit width selection is studied.

Keywords Statistical process control · Phase I · Control limit widths · Estimation error · Contaminated data · Mean squared error

M. C. Testik (✉) · Y. Koca · Ö. M. Testik
Hacettepe University, Department of Industrial Engineering, Beytepe-Ankara, Turkey
e-mail: mtestik@hacettepe.edu.tr; yesimkoca@hacettepe.edu.tr; ozlemaydin@hacettepe.edu.tr

C. H. Weiß
Helmut Schmidt University, Department of Mathematics and Statistics, Hamburg, Germany
e-mail: weissc@hsu-hh.de

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_2

1 Introduction

Monitoring of a process or product characteristic using a control chart requires the determination of parameter values, such as the in-control process mean and standard deviation, to be used in the design of the chart. Since these parameters are often unknown in many real world applications, a retrospective control chart application, called Phase I, becomes necessary to estimate the unknown parameters. Utilizing a set of observations that are assumed to be clean from the effects of assignable causes of variation, unknown parameters are estimated and then used in designing the chart for online monitoring, i.e. for the Phase II application of a control chart. Although a Phase I study is frequently necessary in practical applications, most research on control charts focused on Phase II development and performance evaluation, treating the control chart parameters as known.

Recently, a topic that deserved considerable attention from researchers is the effect of parameter estimation on the control chart properties. Interested readers are referred to Jensen et al. (2006) for a review of the literature. In the studies, variability due to estimation is taken into account. Assuming control chart designs with estimated parameters, conditional and marginal performances in the Phase II of control charts are evaluated and some recommendations on sample sizes are provided (see, for example, Weiß and Testik 2011; Testik 2007; Testik et al. 2006; Jones et al. 2001), or adjustments to Phase II control limits are proposed (see, for example, Albers and Kallenberg 2004, 2005). In these works, it is emphasized that collecting representative samples of sufficient size will ensure the desired Phase II performance. Nevertheless, these studies intrinsically considered that the Phase I observations are clean in the sense that they are obtained from a statistically in-control process.

But obtaining a reference set of in-control observations from a process is often a difficult task for an analyst, and such judgments may require process knowledge. Even with a good process knowledge, the stability of a process should be checked. As Shewhart (1939) notes, "In the majority of practical instances, the most difficult job of all is to choose the sample that is to be used as the basis for establishing tolerance range. If one chooses such a sample without respect to the assignable causes present, it is practically impossible to establish a tolerance range that is not subject to a huge error. Before choosing the sample, therefore, it is desirable to try to detect the presence of assignable causes and to discover the nature of these so that their influence may be foretold." Yet, research on methods to obtain an in-control reference set of observations has received less emphasis (Jones-Farmer et al. 2014).

In this study, obtaining a reference sample that is representative of the in-control state of the process is considered. Parameter estimation in Phase I applications is studied by simulating different levels of contaminations to represent effects of the presence of assignable causes of variation in fixed-size sets of samples of observations. The iterative steps of the Phase I analysis using conventional Shewhart control charts are simulated for detecting the presence of assignable causes of variation, and these are then removed to obtain statistically in-control reference

sets of observations. To better handle the detection of assignable causes, Shewhart control limits' width is adjusted as a control factor. The Mean Squared Error (MSE) criterion is used to evaluate the estimates of the unknown parameters.

The chapter is organized as follows. In Sect. 2, the Phase I application of control charts is explained. Design of the Shewhart control charts for monitoring the mean and the variability of a normally distributed characteristic is given next (Sect. 3). Following the description of the simulation methodology in Sect. 4, results are discussed and recommendations on the control limit widths to obtain approximately optimal MSE for the mean are given (Sects. 5 and 6). Then the chapter is concluded.

2 Phase I Application of Control Charts

Phase I applications are essential if the values of parameters required for designing a control chart are unknown. The aim is to identify the in-control state of the process for characterizing the quality characteristic and designing the control chart to be used in Phase II for online monitoring. For this purpose, robust estimators or change-point procedures can be considered as alternatives in Phase I. The readers are referred to Schoonhoven and Does (2012) and Zwetsloot et al. (2014) for studies on the use of robust estimators in Phase I, and to Samuel et al. (1998a,b) for the use of change point procedures to estimate the time of a process change. Yet, the standard textbook recommendation (see, for example Montgomery 2009) and the most commonly used approach in practice is a retrospective application of standard Shewhart control charts in Phase I. In the Phase I stage of a control chart implementation, a fixed-size set of observations on a characteristic to be monitored is gathered during a time period when the process is considered to be in-control. To assess the process stability, the set of observations is iteratively tested for the presence of assignable causes of variation. Hence, Phase I is an attempt to determine out-of-control situations retrospectively (Montgomery 2009; Quevedo et al. 2016).

Phase I analysis for parametric control charts begins with the identification of an appropriate probabilistic model for the characteristic of interest and the corresponding control statistic to be monitored. Then the parameters required for designing a Phase I control chart selected are estimated. Shewhart type control charts are often recommended in Phase I, and the iterative implementation of the use of these charts is as follows (see, Weiß and Testik 2015; Dasdemir et al. 2016). Initial parameter estimates are obtained first by using a fixed-size Phase I set of observations, and trial control limits are determined. Then the trial control limits are used to test if there are control statistic values exceeding the limits. Observations corresponding to control statistic values that exceed the limits are investigated for the presence of assignable causes of variation. Corrective actions are taken for the identified assignable causes of variation and the corresponding observations are ignored in the subsequent iteration. Using the remaining observations in the Phase I set of observations, parameter estimates and the control limits are revised, control statistic values are tested with the revised control limits, corrective actions are taken

for identified assignable causes of variation, and the observations corresponding to these are ignored in the subsequent iteration. This is iterated until all the points are within the control limits. Final parameter estimates are then used in designing the Phase II control chart. Note that, if observations corresponding to out-of-control states of a process are not omitted when estimating the parameters, these will reflect themselves in the parameter estimates, effects of which will be propagated to the Phase II performance of the chart.

3 $\bar{x}$ and s Control Charts

In practice, a widely used approach for determining a reference set of observations for estimation in Phase I is to use Shewhart control charts. This is due to the ease of construction and interpretation, effectiveness in detecting large sustained shifts in the parameters, outliers, measurement errors, data recording errors and the like (Montgomery 2009). Therefore, following the recommendations of standard textbooks, the Shewhart control charts $\bar{x}$ and s are considered for the Phase I analysis in this study.

Suppose that m subgroups of size n observations on a characteristic x are used at an iteration of a Phase I application. Let $\bar{x}$ be the subgroup average,

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

and s be the subgroup standard deviation,

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

The upper control limit (UCL), center line (CL), and lower control limit (LCL) for the $\bar{x}$ control chart are as follows;

$$\begin{aligned} \text{UCL} &= \bar{\bar{x}} + \frac{L\bar{s}}{c_4\sqrt{n}}, \\ \text{CL} &= \bar{\bar{x}}, \\ \text{LCL} &= \bar{\bar{x}} - \frac{L\bar{s}}{c_4\sqrt{n}}, \end{aligned}$$

where L is the distance of a control limit from the center line in terms of standard deviation units, $\bar{\bar{x}}$ is the average of m subgroup averages,

$$\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_i,$$

$\overline{s}$ is the average of m subgroup standard deviations,

$$\overline{s} = \frac{1}{m}\sum_{i=1}^{m} s_i,$$

and $c_4 = c_4(n)$ is a constant depending on n such that $\overline{s}/c_4$ is an unbiased estimator of the process standard deviation σ: $c_4(n) = \sqrt{2/(n-1)}\,\Gamma(\frac{n}{2}) \,/\, \Gamma(\frac{n-1}{2})$.

Since the standard deviation of s is $\sigma\sqrt{1-c_4^2}$, a common textbook recommendation for the UCL, CL, and LCL of the s control chart are,

$$\begin{aligned} \text{UCL} &= \overline{s} + L\frac{\overline{s}}{c_4}\sqrt{1-c_4^2}, \\ \text{CL} &= \overline{s}, \\ \text{LCL} &= \overline{s} - L\frac{\overline{s}}{c_4}\sqrt{1-c_4^2}, \end{aligned}$$

respectively. The distribution of s is not symmetric, so the use of symmetric control limits for the s chart (see recommendation in, e.g., Section 6.3.1 in Montgomery (2009)) is an approximation. Note that the $\overline{x}$ control chart is used to detect shifts in the mean, whereas the s control chart is used to detect shifts in the variability. In Phase I applications, the practitioner simultaneously uses these control charts, and the signals of the charts are evaluated for the presence of assignable causes of variation. The iterative use of these charts is as described in the previous section.

Besides the parameters m, n defining the number and size of subgroups, the practitioner also has to choose a value for the factor L of the control limits. This design parameter is often selected to be 3 in conventional use. To study the effect of the choice of L on the Phase I performance of the charts, we shall vary its value in the simulation study presented below and compare the results to those of the standard choice $L = 3$ (benchmark value for L).

4 Phase I Simulations Using Shewhart Control Charts

Two-sided $\overline{x}$ and s control charts are considered in this study to identify the presence of assignable causes of variation in Phase I. For a fixed-size set of Phase I observations, it is assumed that the cleaner the set of observations, the better will be the parameter estimates. The process model, simulation details and metrics used in evaluations are provided below.

Let the observations on the characteristic of interest be independent and normally distributed. For simplicity but without loss of generality, consider that the in-control mean is $\mu_0 = 0$ and the in-control standard deviation is $\sigma_0 = 1$. As a sampling strategy, it is assumed that consecutive samples are taken to minimize the chance of variability due to assignable causes within a subgroup and to maximize the chance

of variability between subgroups when assignable causes of variation are present, i.e. rational subgroups are considered. Consequently, process state changes due to assignable causes of variation are assumed to be shifts in the means between two successive subgroups, which affect all the observations within a subgroup after the change. Note that this is the snapshot approach as discussed in Montgomery (2009).

In the simulations, an out-of-control process due to the presence of assignable causes was modeled by a change in the mean from $\mu_0 = 0$ to a new level μ_1, but the standard deviation $\sigma_0 = 1$ was kept constant. Considering m initial subgroups each having a size of n observations, and a total of $N = mn$ observations in the Phase I dataset, a percentage c of the total observations (i.e. mc subgroups) were contaminated such that the contaminated subgroup mean becomes μ_1.

As controllable factors, the distance of a control limit from the center line in terms of standard deviation units (L), initial number of subgroups (m), and the number of observations within each subgroup (n) were studied in the simulation experiments. In addition, the contamination percentage (c) and shifted mean (μ_1) were considered as uncontrollable factors (although these factors are controllable for the purpose of simulation tests). Considered cases of interest are: $L = 1$–5 with increment size 0.1, $m = 25, 50, 100, 200, 1000$, $n = 5, 10$, $\mu_1 = 1, 2, 3$, and $c = 0, 4$ and 8%. Note that $c = 0$ corresponds to the case where there are no effects of assignable causes of variation in the Phase I data set.

For each combination of controllable and uncontrollable factors, a simulation study with 100,000 replications was performed by generating a Phase I data set for each replication, and by iterating the Phase I steps until all the points were within the control limits. Here, it is assumed that a point exceeding the Phase I limits is an indication of the existence of an assignable cause and is therefore ignored in the calculations of the subsequent iteration. Several performance metrics were calculated as follows for evaluating the Phase I applications:

- *Average Number of Iterations (ANI)* for evaluating the computational effort to reach a final decision that the process is in-control, is obtained by counting the number of iterations in each replication, and by taking the average of these.
- *True Alarm Percentage (TAP)* for determining the power in detecting assignable causes is

$$E\left[\frac{T}{mc}\right] \times 100,$$

  where T is the number of subgroups that trigger an out-of-control signal with either of the charts when the subgroup actually corresponds to an assignable cause. The ratio of the number of true signals to the number of contaminated samples is taken in each replication, and the average of these is calculated.
- *False Alarm Percentage (FAP)* for determining the performance in falsely detecting in-control subgroups as if assignable causes are present is

$$E\left[\frac{F}{m - mc}\right] \times 100,$$

where F is the number of subgroups that trigger an out-of-control signal with either of the charts when the subgroup actually corresponds to an in-control process. The ratio of the number of false signals to the number of clean samples is taken in each replication, and the average of these is calculated.

- *Mean squared error (MSE)* for determining the accuracy of estimation for the mean and standard deviation is

$$\mathrm{MSE}(\bar{\bar{x}}) = E\left[(\bar{\bar{x}} - \mu_0)^2\right],$$
$$\mathrm{MSE}(\bar{s}/c_4) = E\left[(\bar{s}/c_4 - \sigma_0)^2\right],$$

respectively. The square of the difference of the final parameter estimate from its true value was calculated in each replication, and the average of these was taken.

5 Results of Simulations

Subgroups are tested for the presence of assignable causes of variation in Phase I applications. The trial control limits are revised as out-of-control process states are detected, and subgroups corresponding to these are removed. This is iterated until all the points are within the control limits. To understand the effect of the distance L of a control limit from the center line on ANI, simulation results are provided in Figs. 1 and 2 for $c = 4\%$, respectively for $n = 5$ and $n = 10$. Additional results with $c = 8\%$ can be found in Figs. 10 and 11 in Appendix 1. In the figures, each panel corresponds to a mean value μ_1, where the left panels (shift $= 0$) are the in-control results (without contamination) for comparison. In each panel, the ANI graphs for the tested number of initial subgroups m are provided against L.

With regard to Figs. 1 and 2, as well as to Figs. 10 and 11, it can be seen that the ANI has a peak in between L values of 1 and 2 for all of the cases. This is due to many points plotting outside the tight control limits. Note that the points outside the control limits are not necessarily true signals for the existence of assignable causes of variation. True and false signal percentages, as well as their effect on the accuracy of estimation, will be discussed later. As expected, up to a point, ANI decreases with increasing L, which corresponds to wider control limits and less signals. Furthermore, the larger the number of initial subgroups m, the higher the ANI for a given L. Note that more signals can be expected as the number of observations increase, which, in turn, lead to more iterations.

For the shift values $\mu_1 = 0$ and $\mu_1 = 1$, the ANI approaches 1 in the interval 3–5 for L. That is, the estimation is completed in a single iteration on average since the charts do not signal much with the wider control limits. Yet, the convergence to this limiting ANI value is slower with the shift value 1, compared to 0.

On the other hand, for the moderate shift value $\mu_1 = 2$ and $n = 5$, there are local minimums greater than the limiting value 1 for the ANI graphs in the interval 2–4 for L. With the L values greater than the ones corresponding to the local minimums,

Fig. 1 ANI at the end of Phase I analyses for various shifts in the mean (panels) with contamination $c = 4\%$, number of subgroups m (graphs), and distance L of control limit from center line, when the number of observations n within subgroups is 5

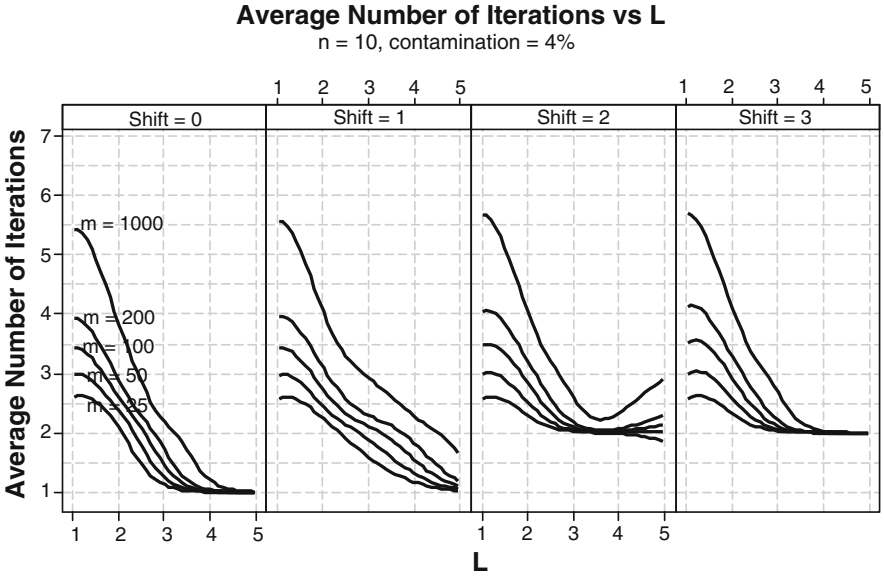


Fig. 2 ANI at the end of Phase I analyses for various shifts in the mean (panels) with contamination $c = 4\%$, number of subgroups m (graphs), and distance L of control limit from center line, when the number of observations n within subgroups is 10

the ANI may slightly increase first and then decrease below this local minimum. This slight increase in the ANI with an increase in L from the corresponding local minimum can be explained by having less signals in the initial iterations but as these are detected and removed, slowly the control limits get tighter for detection of the remaining points in the subsequent iterations. Hence, the number of iterations increase up to a point. Then, with the decrease of detection capability due to wider control limits, the ANI starts to decrease with less and less signals to trigger new iterations. Considering the shift value of $\mu_1 = 2$ but with $n = 10$, the minimums of ANI are achieved in the interval 2–4 for L, where the global minimums are approximately 2 for ANI. With an increase of the detection power in contrast to $n = 5$, the effect of increasing L from the corresponding minimum value of ANI is more visible. Although there is a decrease of detection capability due to wider control limits with increasing L, the ANI starts to increase with less signals that are detected. So the control limits get tighter slowly, resulting in the detection of the remaining points in the subsequent iterations. The shift value of $\mu_1 = 3$, both with n equal to 5 and 10, achieves the minimums of ANI in the interval 3–5 for L, where the global minimums are again approximately 2. Taking altogether, the benchmark choice $L = 3$ usually goes along with a relatively low ANI, which certainly is attractive for practice.

Now consider the effect of the distance L of a control limit from the center line on the type of signals of the control chart. Figure 3 plots the FAP results for the combinations of c and n in a panel, with the graphs in the panels representing

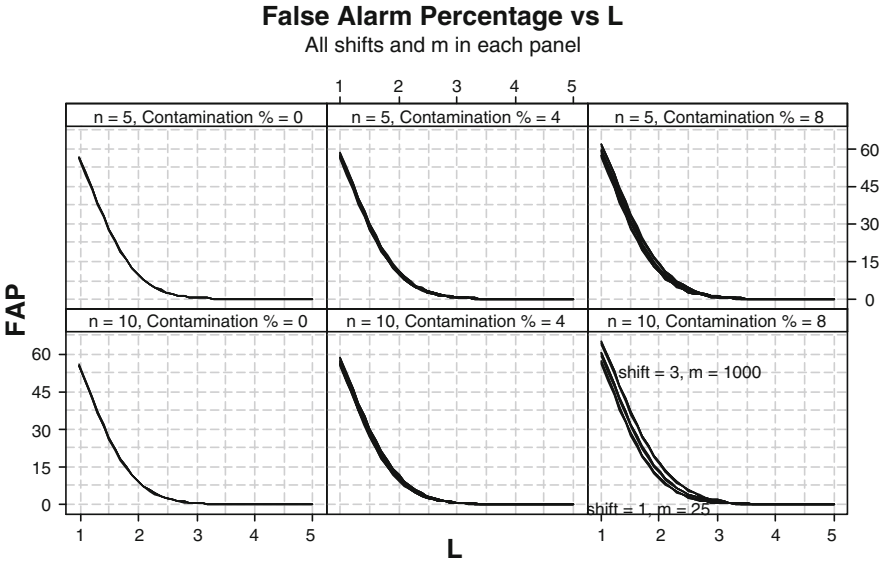


Fig. 3 FAP at the end of Phase I analyses against L values for the combinations of number of observations within subgroups n and contamination percentage c (panels) for the shifts considered (graphs)

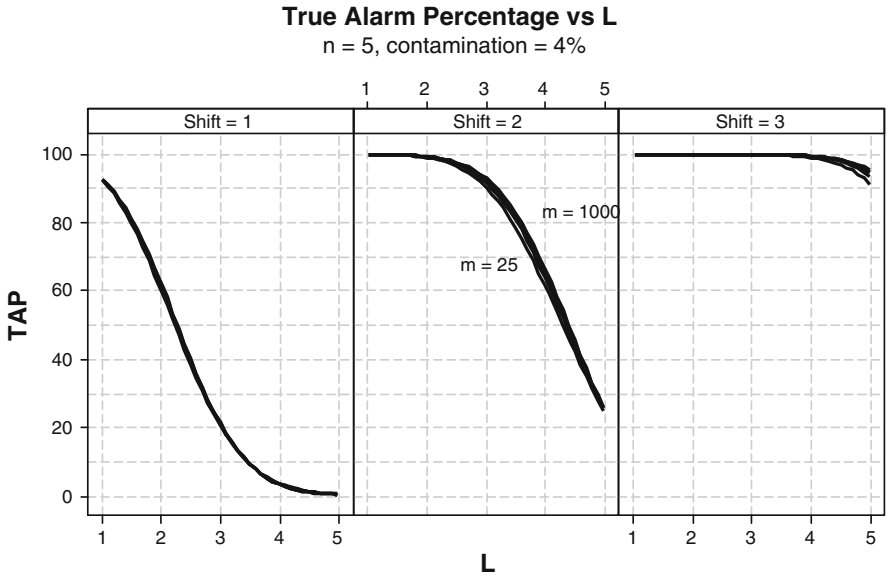


Fig. 4 TAP at the end of Phase I analyses for various shifts in the mean (panels) with contamination $c = 4\%$, number of subgroups m (graphs), and distance L of control limit from center line, when the number of observations n within subgroups is 5

the combinations of m and shift values. Simulation results for the TAP are provided in Figs. 4 and 5 for $c = 4\%$, respectively for $n = 5$ and $n = 10$. Additional results with $c = 8\%$ can be found in Figs. 12 and 13 in Appendix 2. In these figures, the FAP and TAP graphs in each panel are very close to each other and, hence, FAP and TAP are sensitive mostly to the shifted values of the mean, μ_1, but not to the number m of subgroups.

The FAP of the charts under the different cases considered are all similar. The FAP graphs are monotonically decreasing as L gets larger. At $L = 1$, FAP is close to 60% and with the L value of 2.5 or greater (including the benchmark choice $L = 3$), these graphs are close to 0, indicating that false signals are rare. Yet, a slightly larger FAP can be observed with the larger shift value $\mu_1 = 3$.

While the behavior of the FAP is quite unique and reasonably low for $L > 2.5$, the TAP results are more complex. Consider first the shift value $\mu_1 = 1$. While most of the out-of-control subgroups are detected in a Phase I study with an L value in the interval 1–2, there is a steep decrease of the true alarm percentages in the interval 2–3 for L when $n = 5$. However, this decrease in power is slower with $n = 10$. As L approaches 5, TAP approaches 0 indicating that the charts lose the capability to detect true signals when the shift in the mean is small.

For the moderate sized shift value $\mu_1 = 2$, most of the out-of-control subgroups are detected with the L values in the interval 1–3 when $n = 5$. As L approaches 5, TAP approaches 30%. For $n = 10$, the interval for a good detection power gets

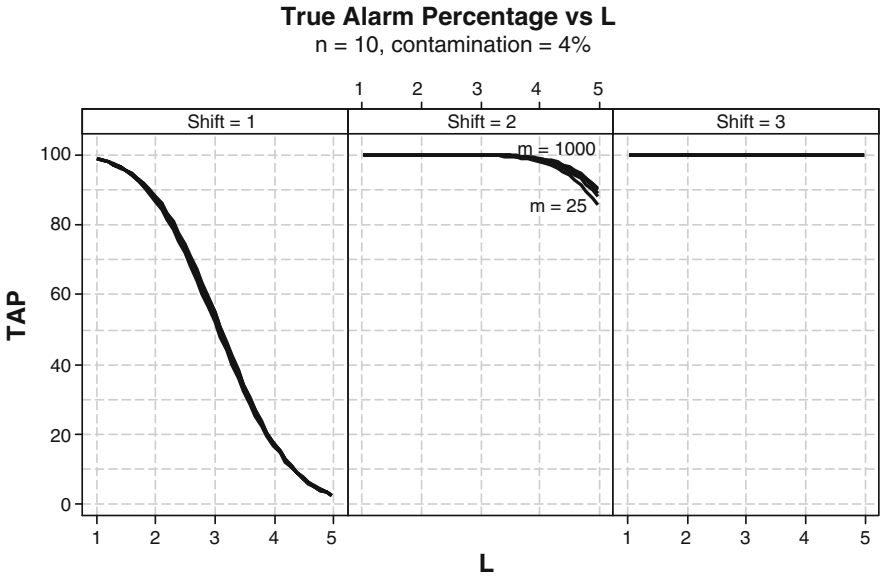


Fig. 5 TAP at the end of Phase I analyses for various shifts in the mean (panels) with contamination $c = 4\%$, number of subgroups m (graphs), and distance L of control limit from center line, when the number of observations n within subgroups is 10

visibly larger, and L values between 1 and 4 result in a very high TAP. In this case, as L approaches 5, TAP approaches 90%. Hence, there is a clear performance advantage in true signals when $n = 10$, compared to $n = 5$. So while the number m of subgroups has only little effect on the TAP, their size n should be carefully chosen. On the other hand, for the large shift case $\mu_1 = 3$, almost all of the out-of-control subgroups are detected when an L value is selected to be between 1 and 5. When L is 5, TAP is close to 90% with $n = 5$, and even close to 100% with $n = 10$. Altogether, in view of the FAP, the value of L should not fall below 2.5, while an upper bound as implied by the TAP depends on the size of the shifts to be considered. For moderate to large shifts, the benchmark choice $L = 3$ appears to be a reasonable choice in view of FAP and TAP.

A control chart signal may be either true or a false. The total number of signals at the end of a Phase I study is composed of both the true and false signals. The average of the total number of signals is a monotonically decreasing function of L, since the control limits get wider as L increases and the probability of a point plotting outside these limits gets smaller. Consider again the ANI graphs for the moderate shift value $\mu_1 = 2$ with contamination 4%. These graphs have a local minimum around 3. Here, the total number of signals, dominated by false signals, starts decreasing as L increases. With the increase of L to 3, although the total number of signals is less, these are mostly true signals. When L is increased more, the ANI starts to increase first, since a high percentage of true out-of-control subgroups are detected but with more iterations. With larger L, the true alarm percentage starts

dropping steeper and the true out-of-control subgroups become mostly undetectable with the wider control limits. Hence, the ANI starts to decrease since there are less signals to trigger subsequent iterations.

Finally, let us turn to the mean squared errors of the mean and standard deviation estimates as obtained to the end of Phase I analysis. Before discussing the simulation results, let us point out that the limit $L \to \infty$ corresponds to the case of computing estimates from the full initial data set. For the shift scenarios considered here, and using the well-known identity MSE = Var + Bias^2, the corresponding MSE values are computed exactly as

$$\mathrm{MSE}(\bar{\bar{x}}) = \frac{\sigma_0^2}{m\,n} + \left(\frac{c}{100}\,(\mu_1 - \mu_0)\right)^2 = \frac{1}{m\,n} + \left(\frac{c\,\mu_1}{100}\right)^2,$$

$$\mathrm{MSE}(\bar{s}/c_4) = \frac{\sigma_0^2\,(1 - c_4^2)}{m\,c_4^2} + 0^2 = \frac{1 - c_4^2}{m\,c_4^2}.$$

So the limiting behaviour of the mean estimates' MSE depends on the shift size μ_1. Since these limiting values express the MSE if not doing any data filtering, these can be interpreted as benchmark values; such values of L are desirable that lead to an MSE close to or preferably below the respective limiting value.

For each fixed-size initial set of Phase I observations, the MSEs for the mean estimates are presented in Figs. 6 and 7 for the $c = 4\%$ with $n = 5$ and $n = 10$,

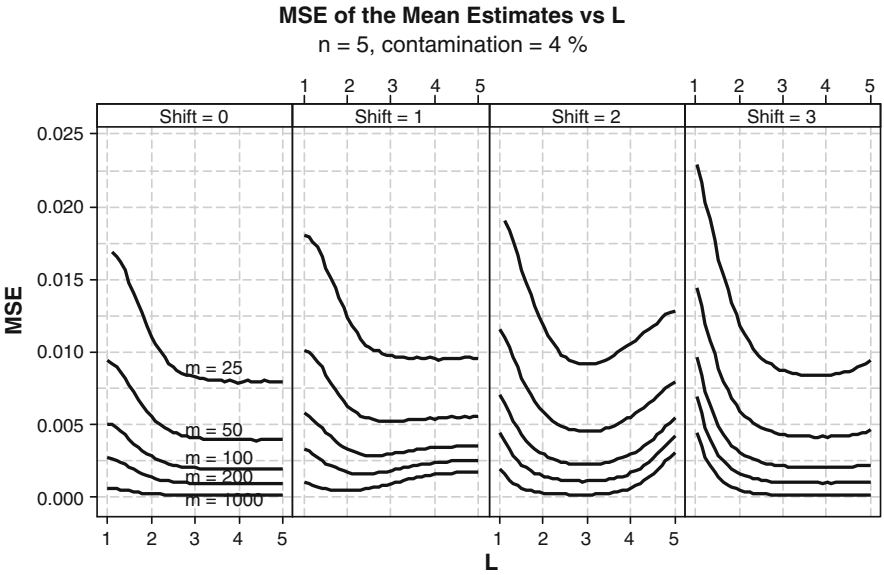


Fig. 6 MSE of the mean estimates at the end of Phase I analyses for various shifts in the mean (panels) with contamination $c = 4\%$, number of subgroups m (graphs), and distance L of control limit from center line, when the number of observations n within subgroups is 5

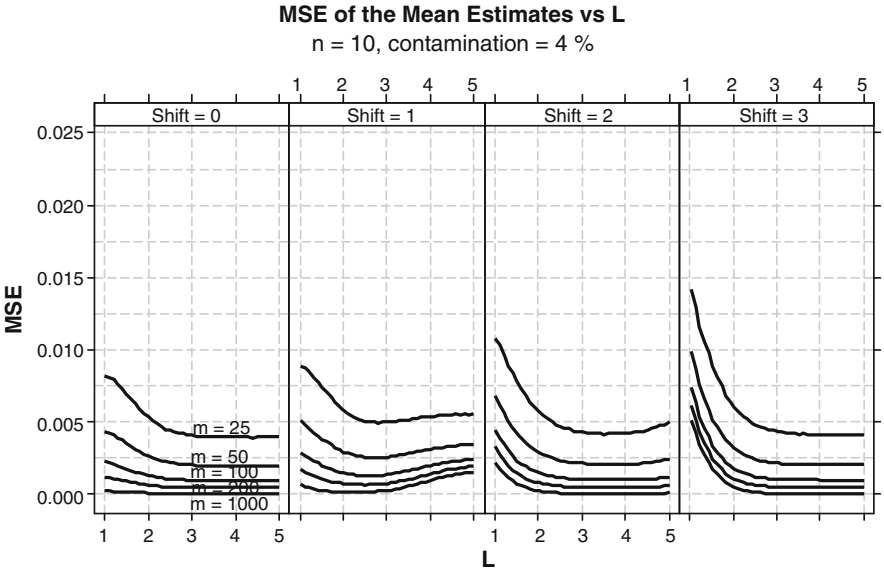


Fig. 7 MSE of the mean estimates at the end of Phase I analyses for various shifts in the mean (panels) with contamination $c = 4\%$, number of subgroups m (graphs), and distance L of control limit from center line, when the number of observations n within subgroups is 10

respectively. The results for $c = 8\%$ cases are provided in Figs. 14 and 15 in Appendix 3. In these figures, the left panels are the in-control cases for comparisons. In each panel, the MSE graphs for the tested number of initial subgroups m are provided against the distance L of control limits from the center line.

The sensitivity of the MSE to L can be seen by a comparison of the MSE graphs within and between panels in each figure. For the out-of-control cases (panels with shifts ≥ 1), the MSE graphs are U-shaped, with a large MSE either for small L (then too many in-control subgroups are removed from the Phase I data set), or for large L (then contaminated subgroups are not detected). In particular, there always exist values of L leading to an MSE below the limiting value, i.e., these Phase I analyses lead to an improvement compared to the initial situation. A further comparison indicates that larger shifts in the mean may result in smaller MSE values for a given value of L and m. This is due to the increase of the detection performance of the $\overline{x}$ chart with larger mean shifts, and the effect can be seen better with smaller subgroup sizes m. However, for a given out-of-control shift and L value, the MSE can be larger with the higher contamination percentage $c = 8\%$, since some observations representing the out-of-control process cannot be detected.

For the larger shifts $\mu_1 = 2$ and $\mu_1 = 3$, the minimum of the MSE value is located close to the benchmark choice $L = 3$. On the other hand, flatness of the MSE graphs can be observed around the L values that provide the minimum MSE value. Accordingly, one can identify alternative L values for given m and n pairs

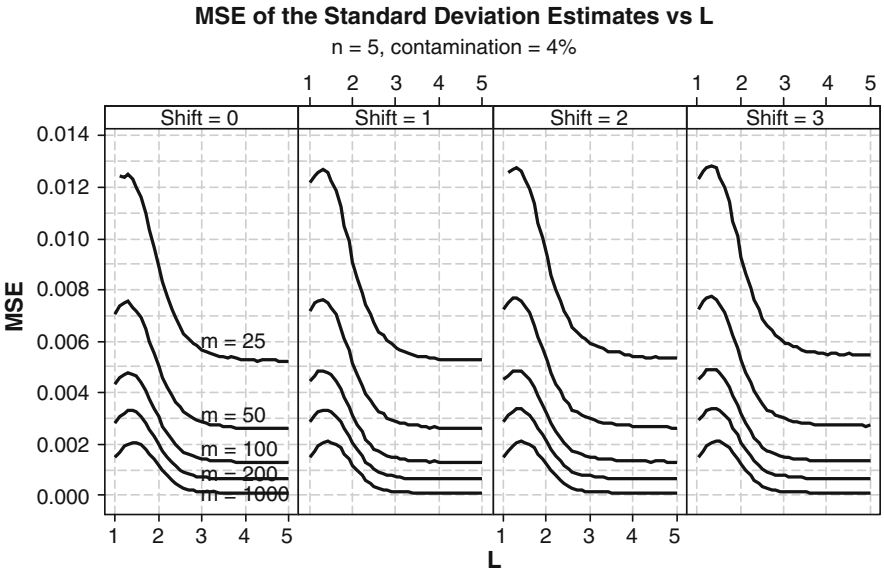


Fig. 8 MSE of the standard deviation estimates at the end of Phase I analyses for various shifts in the mean (panels) with contamination $c = 4\%$, number of subgroups m (graphs), and distance L of control limit from center line, when the number of observations n within subgroups is 5

to yield minimum or close to minimum MSE of mean estimates. As expected, the greater the initial number m of subgroups or the subgroup size n, the smaller the MSE.

Now consider the MSE graphs for the standard deviation estimates given in Figs. 8 and 9 for the $c = 4\%$ cases with $n = 5$ and $n = 10$, respectively, and the $c = 8\%$ cases presented in Figs. 16 and 17 in Appendix 3. Since rational subgroups are considered and the mean shifts are assumed to be between subgroups, the standard deviation estimates are essentially only affected by the number of samples used for estimation for a given m and n. The number of subgroups used for estimation at the end of a Phase I application decreases with an increase of the mean shift, since more signals are expected to be triggered by the $\overline{x}$ chart. Therefore, the MSE of the standard deviation estimates are slightly greater in a figure, for a given m and L, as the shift increases. On the other hand, flatness of the MSE graphs around the minimum MSE can be clearly seen for a wide interval of L values. The MSE graphs generally have an approximate minimum with the L values greater than 3 for $n = 5$, whereas the approximate minimum is achieved with L values greater than 2.3 for $n = 10$.

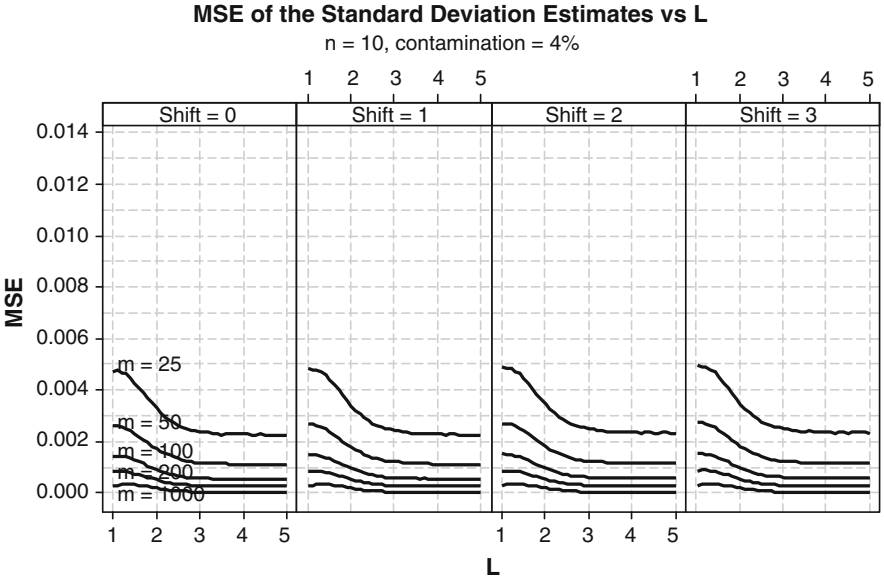


Fig. 9 MSE of the standard deviation estimates at the end of Phase I analyses for various shifts in the mean (panels) with contamination $c = 4\%$, number of subgroups m (graphs), and distance L of control limit from center line, when the number of observations n within subgroups is 10

6 MSE Optimal and Robust L Values for Phase I Charts

In industrial settings, there are many possibilities of shift sizes and contamination levels for out-of-control observations from a process. Therefore, a generalization for the L value used in the estimation under various levels of these uncontrollable factors is important for practical use. In the following, L values that are robust to a variety of uncontrollable factor levels in estimating the mean are investigated.

It was mentioned earlier that MSE graphs are almost flat around the L values that provide the minimum MSE, and therefore alternative L values can be identified for given m and n pairs to yield minimum or close to minimum MSE of mean estimates. To provide suggestions for practitioners, we searched for L values that are robust over the considered contamination percentages $c = 4$ and 8%, the shifts $\mu_1 = 0, 1, 2, 3$, as well as their combinations. For this purpose, an upper deviation bound of 5% from the minimum MSE value is selected, and L values satisfying this condition were searched. Hence, intervals for L for each combination of m, n, c and μ_1 were identified. For each m and n pair, which are controllable factors for practitioners, L values were determined by an interval intersection rule developed. According to this rule, intersection of the intervals for L values were searched. For example, if one is looking for robust L values over all of the considered shifts and contamination percentages, there are seven intervals (1 in-control and 3 out-of-control shift cases each with 2 contamination percentages) for each of the m

and n pairs. If an interval of L that satisfies all of these 7 intervals could be found, this interval is considered to have a robust performance in estimation. Since this is not possible for all cases, majority voting is used when an intersection of all intervals cannot be found. Note that one should also consider the MSE of the standard deviation estimates in selecting an L value for Phase I analysis. Among the alternative L values for the mean, one can select the larger ones to reduce the MSE of the standard deviation estimates.

Considering the joint operation of $\overline{x}$ and s charts with the use of the same L value for both charts, L values that provide approximately MSE optimal estimates of the mean for rational subgroups are provided in Table 1. In the table, "*" indicates that majority voting is used. Note also that an upper deviation bound of 5% is used here. Since the change in the MSE may be very small outside this bound in some cases, alternative bounds can also be considered to reduce the variation in L among different cases.

Overall, it becomes clear that the MSE optimal L values are often close to the benchmark choice of $3\,\sigma$ limits ($L = 3.0$), and such a choice is further supported in view of having a low false alarm rate (see the above FAP results), and of having a reasonable power to detect the moderate to large shifts (TAP results for $\mu_1 = 2, 3$). Table 1 indicates that the MSE optimal L should be slightly lower than 3.0 for situations where at most small mean shifts are to be expected, while it should be slightly larger than 3.0 in the case of larger mean shifts.

7 Conclusions

The Phase I implementation of Shewhart control charts is generally essential, especially in practical settings, to design various control charts for monitoring in Phase II. In fact, the Phase II performance of control charts in terms of false alarms and detection often depends on the design of the chart with the use of estimated parameters.

In this study, the Phase I implementation of the Shewhart control chart for a normally distributed process is simulated under the assumption of rational subgroups. Considering scenarios for mean shifts and contamination percentages of the initial subgroups, the distance L of a control limit from the center line, as the Shewhart control chart parameter, is altered for various numbers of initial subgroups and observations in each subgroup.

The computational requirements for the Phase I implementation are expressed through the average number of iterations metric. It is shown that if the subgroups have no mean shift or a mean shift of 1 standard deviation, Phase I implementations will be completed in a few iterations when the conventional $3\,\sigma$ limits are used. With larger shifts in the mean, the number of iterations required may be larger.

Since a control chart signal may be either true or a false, and the total number of signals at the end of a Phase I study is composed of both the true and false signals, the performance of Shewhart control charts in Phase I implementations is

Table 1 *L* values for approximately MSE optimal Phase I design for estimating mean

		Shifts in Phase I with contaminations 4 and 8%								
m	*n*	0	1	2	3	0 or 1	0, 2 or 3	2 or 3	1, 2 or 3	0, 1, 2 or 3
25	5	[3.1; 5.0]	[2.8; 3.4]	[2.5; 3.1]	[3; 4.3]	[3.1; 3.4]	3.1	[3; 3.1]	[3; 3.1]	3.1
25	10	[2.5; 5.0]	[2.4; 2.8]	[2.9; 4.1]	[3.2; 5]	2.8	[3.2; 4.1]	[3.2; 4.1]	3.2*	3.2*
50	5	[2.9; 5.0]	[2.4; 2.8]	[2.6; 3.1]	[3.1; 4.4]	2.8 or 2.9*	3.1	3.1	3.1*	3.1*
50	10	[2.7; 5.0]	[2.3; 2.7]	[3; 4.2]	[3.2; 5]	2.7	[3.2; 4.2]	[3.2; 4.2]	3.2*	3.2*
100	5	[2.7; 5.0]	[2.2; 2.4]	[2.5; 3.2]	[3; 4.5]	2.7*	[3; 3.2]	[3; 3.2]	3.0*	3.0*
100	10	[2.9; 5.0]	[2.2; 2.6]	[2.8; 4.4]	[3.4; 5]	2.6 or 2.9*	[3.4; 4.4]	[3.4; 4.4]	3.0 or 3.4*	2.9 or 3.0*
200	5	[2.9; 5.0]	[2.1; 2.2]	[2.5; 3.1]	[2.9; 4.8]	2.9*	[2.9; 3.1]	[2.9; 3.1]	2.9 or 3.1*	2.9*
200	10	[2.6; 5.0]	[2.2; 2.5]	[2.8; 4.5]	[3.1; 5]	2.5 or 2.6*	[3.1; 4.5]	[3.1; 4.5]	3.1*	3.1*
1000	5	[2.3; 5]	[1.7; 2.0]	[2.4; 3.2]	[3.2; 4.6]	2.5*	3.2	3.2	3.2*	3.2*
1000	10	[1.9; 5]	[1.9; 2.5]	[2.9; 4.5]	[3.2; 5]	[1.9; 2.5]	[3.2; 4.5]	[3.2; 4.5]	3.2*	3.2*

"*" indicates that majority voting is used

also studied by using the true alarm percentage (power) and false alarm percentage (size) metrics. It is observed that the false alarm percentages of the charts under the different cases considered are similar, and they monotonically decrease as L gets larger. Furthermore, the true alarm percentages may be dramatically low for a mean shift of 1 standard deviation.

In order to evaluate the accuracy of parameter estimation at the end of a Phase I implementation, mean squared errors for the mean and standard deviation estimates were computed. Robust values of L for estimating the mean by minimizing the mean squared error metric were investigated. It turned out that often, the optimal choice for L is close to 3, i.e. the conventional $3\,\sigma$ limits appear to be a reasonable guideline for Phase I chart design in view of having a small MSE together with low false alarm percentages and reasonably large true alarm percentages.

Appendix 1: Average Number of Iterations for the Cases of $c = 8\%$

See Figs. 10 and 11.

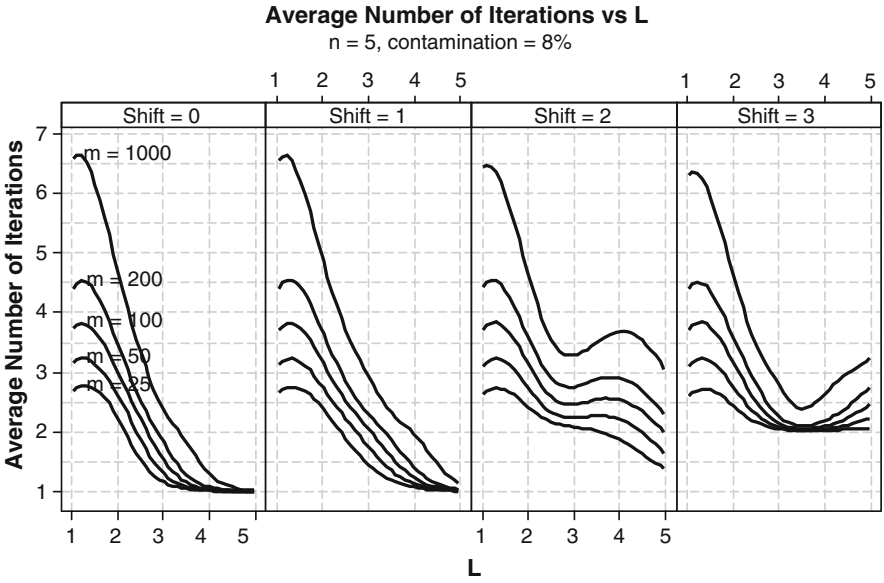


Fig. 10 ANI at the end of Phase I for various shifts in the mean with contamination $c = 8\%$ and when the number of observations within subgroups n is 5

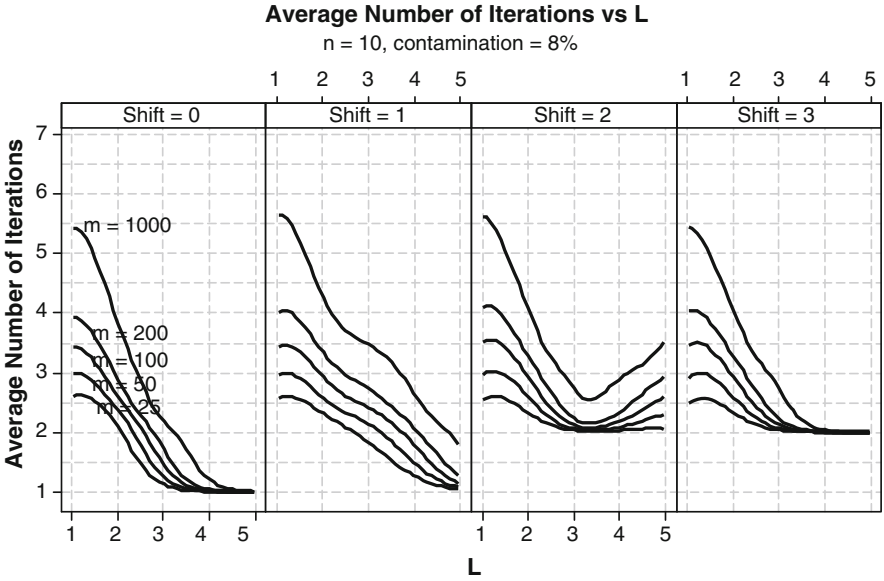


Fig. 11 ANI at the end of Phase I for various shifts in the mean with contamination $c = 8\%$ and when the number of observations within subgroups n is 10

Appendix 2: True Alarm Percentages for the Cases of $c = 8\%$

See Figs. 12 and 13.

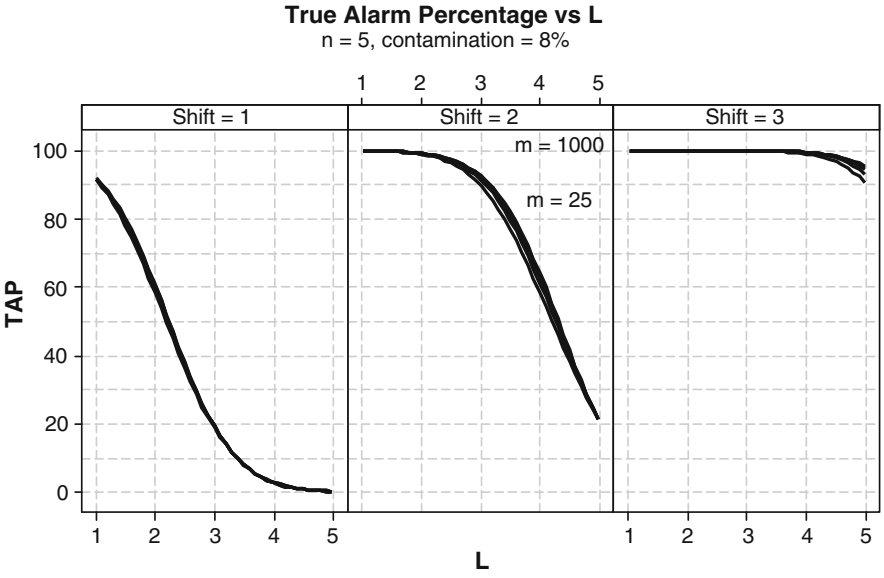


Fig. 12 TAP at the end of Phase I for various shifts in the mean with contamination $c = 8\%$ and when the number of observations within subgroups n is 5

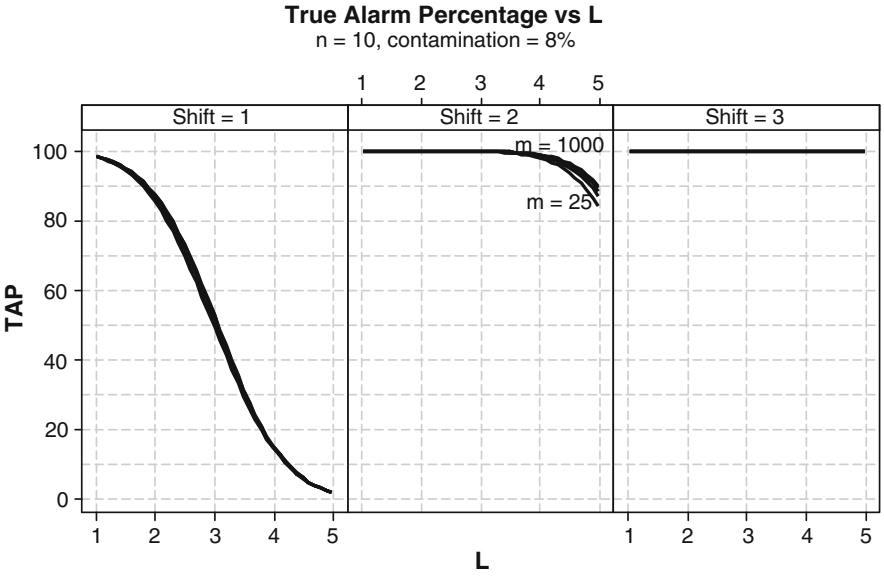


Fig. 13 TAP at the end of Phase I for various shifts in the mean with contamination $c = 8\%$ and when the number of observations within subgroups n is 10

Appendix 3: Mean Square Errors for the Cases of $c = 8\%$

See Figs. 14, 15, 16, and 17.

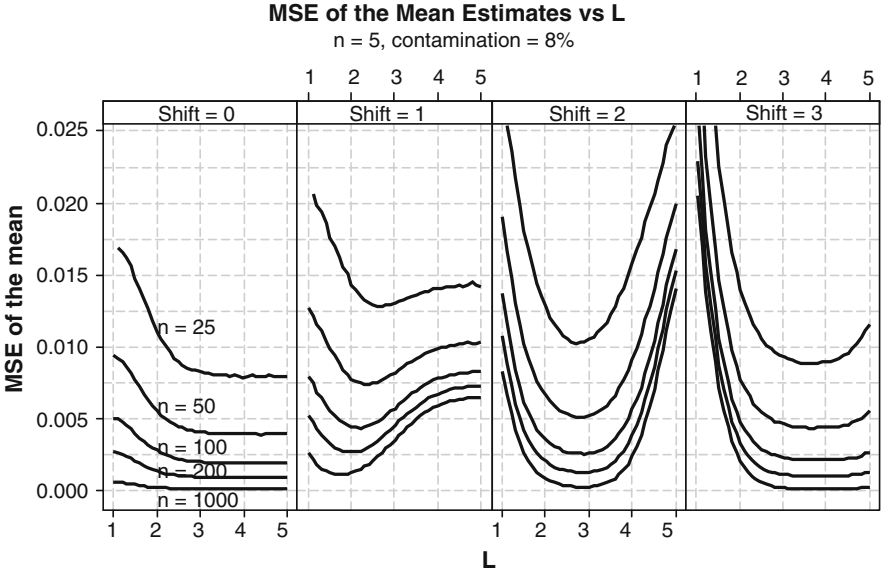


Fig. 14 MSE of the mean estimates at the end of Phase I for various shifts in the mean with contamination $c = 8\%$ and when the number of observations within subgroups n is 5

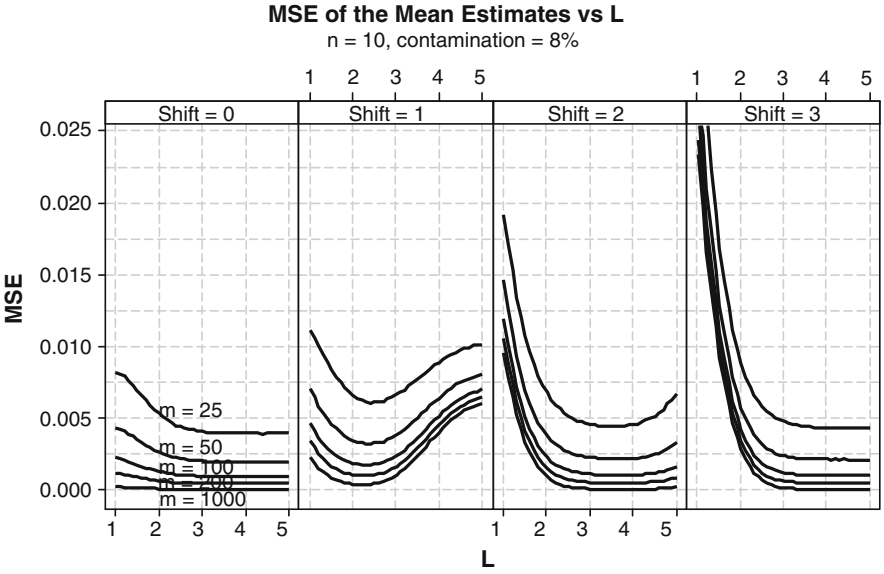


Fig. 15 MSE of the mean estimates at the end of Phase I for various shifts in the mean with contamination $c = 8\%$ and when the number of observations within subgroups n is 10

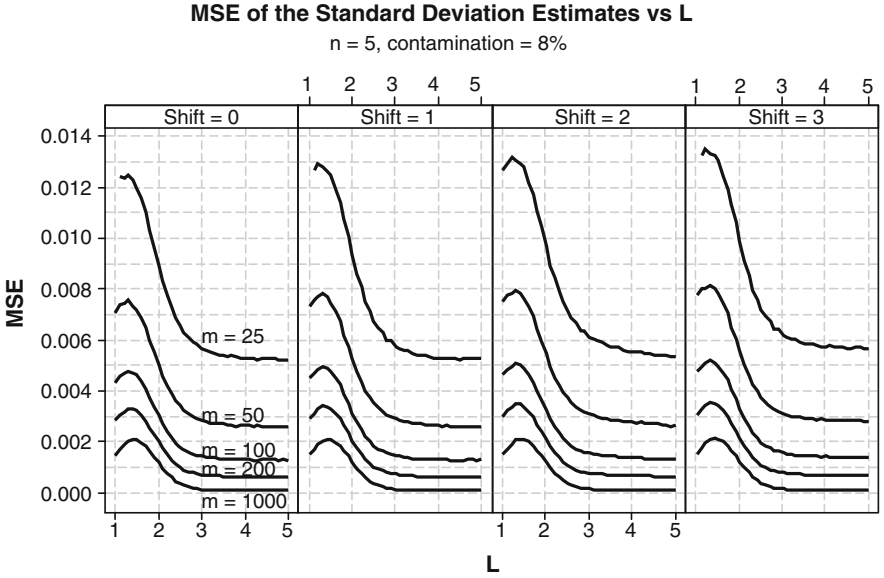


Fig. 16 MSE of the standard deviation estimates at the end of Phase I for various shifts in the mean with contamination $c = 8\%$ and when the number of observations within subgroups n is 5

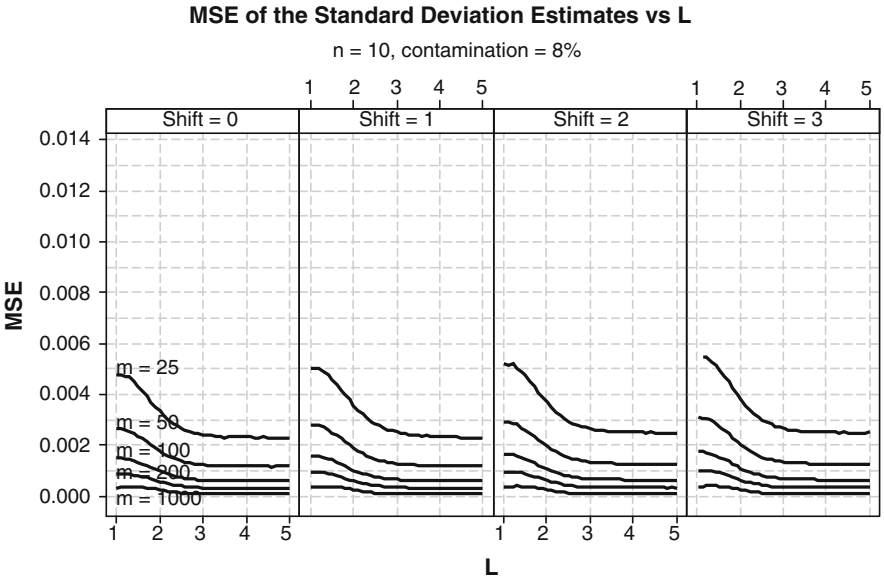


Fig. 17 MSE of the standard deviation estimates at the end of Phase I for various shifts in the mean with contamination $c = 8\%$ and when the number of observations within subgroups n is 10

References

Albers, W., & Kallenberg, W. C. (2004). Estimation in Shewhart control charts: Effects and corrections. *Metrika, 59*(3), 207–234.

Albers, W., & Kallenberg, W. C. (2005). New corrections for old control charts. *Quality Engineering, 17*(3), 467–473.

Dasdemir, E., Weiß, C. H., Testik, M. C., & Knoth, S. (2016). Evaluation of Phase I analysis scenarios on Phase II performance of control charts for autocorrelated observations. *Quality Engineering, 28*(3), 293–304.

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

Jones, L. A., Champ, C. W., & Rigdon, S. E. (2001). The performance of exponentially weighted moving average charts with estimated parameters. *Technometrics, 43*(2), 156–167.

Jones-Farmer, L. A., Woodall, W. H., Steiner, S. H., & Champ, C. W. (2014). An overview of Phase I analysis for process improvement and monitoring. *Journal of Quality Technology, 46*(3), 265–280.

Montgomery, D. C. (2009). *Introduction to statistical quality control* (6th ed.). New York: Wiley.

Quevedo, V., Vegas, S., & Vining, G. (2016). A tutorial on an iterative approach for generating Shewhart control limits. *Quality Engineering, 28*(3), 305–312.

Samuel, T. R., Pignatiello Jr, J. J., & Calvin, J. A. (1998a). Identifying the time of a step change with $\bar{X}$ control charts. *Quality Engineering, 10*(3), 521–527.

Samuel, T. R., Pignatiello Jr, J. J., & Calvin, J. A. (1998b). Identifying the time of a step change in a normal process variance. *Quality Engineering, 10*(3), 529–538.

Schoonhoven, M., Does, R. J. (2012). A robust standard deviation control chart. *Technometrics, 54*(1), 73–82.

Shewhart, W. A. (1939). *Statistical Methods from the Viewpoint of Quality Control*. Reprinted 1986, Mineola: Dover Publications.

Testik, M. C. (2007). Conditional and marginal performance of the Poisson CUSUM control chart with parameter estimation. *International Journal of Production Research, 45*(23), 5621–5638.

Testik, M. C., McCullough, B. D., & Borror, C. M. (2006). The effect of estimated parameters on Poisson EWMA control charts. *Quality Technology and Quantitative Management, 3*(4), 513–527.

Weiß, C. H., & Testik, M. C. (2011). The Poisson INAR(1) CUSUM chart under overdispersion and estimation error. *IIE Transactions, 43*(11), 805–818.

Weiß, C. H., & Testik, M. C. (2015). On the Phase I analysis for monitoring time-dependent count processes. *IIE Transactions, 47*(3), 294–306.

Zwetsloot, I. M., Schoonhoven, M., & Does, R. (2014). A robust estimator for location in Phase I based on an EWMA chart. *Journal of Quality Technology, 46*(4), 302–316.

New Results for Two-Sided CUSUM-Shewhart Control Charts

Sven Knoth

Abstract Already Yashchin (IBM J Res Dev 29(4):377–391, 1985), and of course Lucas (J Qual Technol 14(2):51–59, 1982) 3 years earlier, studied CUSUM chart supplemented by Shewhart limits. Interestingly, Yashchin proposed to calibrate the detecting scheme via $P_\infty(RL > K) \geq 1 - \alpha$ for the run length (stopping time) RL in the in-control case. Calculating the RL distribution or related quantities such as the ARL (Average Run Length) are slightly complicated numerical tasks. Similarly to Capizzi and Masarotto (Stat Comput 20(1):23–33, 2010) who utilized Clenshaw-Curtis quadrature to tackle the ARL integral equation, we deploy less common numerical techniques such as collocation to determine the ARL. Note that the two-sided CUSUM chart consisting of two one-sided charts leads to a more demanding numerical problem than the single two-sided EWMA chart.

Keywords Average Run Length · Fredholm Integral Equation of the Second Kind · Collocation · Numerical Accuracy

1 Introduction

It is a more or less established pattern, that Shewhart charts are powerful tools to detect large changes quickly, while the more complex EWMA (exponentially weighted moving average) or CUSUM (cumulative sum) charts are well suited to trace small and medium size changes. All three have been on the market for a long time now—Shewhart (1926), Roberts (1959) and Page (1954) initiated the research and usage a long time ago. Then a combination of the simple and among the three most popular device, the Shewhart chart, with one of the more subtle siblings seems to be a good idea. To the best of our knowledge, Westgard

S. Knoth (✉)
Institute of Mathematics and Statistics, Department of Economics and Social Sciences, Helmut Schmidt University Hamburg, Hamburg, Germany
e-mail: Sven.Knoth@hsu-hh.de

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_3

et al. (1977) introduced it into statistical process control (SPC) literature. For an application in clinical chemistry, they proposed CUSUM-Shewhart combinations. However, their two-sided CUSUM chart is not the well-known pair of two one-sided schemes. It resembles a CUSUM phenotype which was described later on in Crosier (1986) explicitly. Moreover, Westgard et al. (1977) provided an unorthodox presentation of CUSUM charts, calculated an operations characteristic look-alike measure via 1/ARL (Average Run Length) and performed many Monte Carlo studies to supply, eventually, nomograms for further application of the new scheme. Afterwards, Lucas (1982) and Yashchin (1985b) discussed the combination of two-sided Shewhart charts with the more common construction of a two-sided CUSUM procedure by running two one-sided CUSUMs. Both authors analyzed one-sided designs as well. While Lucas (1982) calculated the zero-state ARL for normal distribution by modifying the popular Markov chain approximation, did Yashchin (1985b) a more elaborated study by dealing with the zero- and steady-state ARL and RL quantiles for normal, χ^2 (normal variance) and Poisson data. He applied Markov chain approximation too. More publications regarding distributions different to normal are Abel (1990) for Poisson, Morais and Pacheco (2006) and Henning et al. (2015) for binomial and Qu et al. (2011) for exponentially distributed data. For the more popular normal case, Starks (1988), Blacksell et al. (1994), and Gibbons (1999) reported application cases, while Reynolds and Stoumbos (2005) and Abujiya et al. (2013) provided more methodological insights and developments. This is, of course, not a complete list of references. Definitely, CUSUM-Shewhart combos became part of standard quality literature, see, for example, Montgomery (2009), chapter 9.1.5. But it is not a popular strand of SPC research. In particular, the ARL calculation was not questioned so far after its first treatment in Lucas (1982) and Yashchin (1985b). This is, more or less, the aim of this contribution. We start with the simpler case of one-sided combos, before the subtle two-sided scheme is touched. Examples are provided, technical details moved into Appendix, and some conclusions complete the chapter.

2 One-Sided CUSUM-Shewhart Chart

Henceforth, denote $\{X_i\}$ a sequence of independent and normally distributed data with mean μ which is under risk to change, and with some known and fixed variance σ^2 that is set to 1 without loosing generality. In this section, we are interested in detecting *increases* in the mean from $\mu_0 = 0$ to $\mu_1 = \delta > 0$. This is done by combining the very popular Shewhart X chart and one of the more known "modern" competitors, the CUSUM chart. First, some math is collected to provide the necessary notions.

$$\text{Shewhart rule} \quad \ell_S = \inf\{i \geq 1 : X_i > c_S\}\,.$$

$$Z_0 = z_0 = 0\,,\ Z_i = \max\{0, Z_{i-1} + X_i - k\}\,,$$

$$\begin{aligned} \text{CUSUM rule} \quad \ell_c &= \inf\{i \geq 1 : Z_i > h\}\,. \\ \text{combo rule} \quad \ell &= \min\{\ell_S, \ell_c\}\,. \\ \text{ARL} \quad &= E_\mu(\ell)\,. \\ \text{ARL function} \quad \mathcal{L}(z) &= E_\mu(\ell \mid z_0 = z)\,. \end{aligned}$$

The terms ARL and ARL function label the well-known Average Run Length both per se and as function of the initializing value z_0. Apparently, the CUSUM-Shewhart combo consists of three parameters, the alarm thresholds c_S (Shewhart) and h (CUSUM), and CUSUM's reference value k, which is typically set to $(\mu_0+\mu_1)/2 = \delta/2$. In all, they control the detection performance of the combo. Typically, some in advance chosen large false alarm level, here denoted by A, and several prominent shifts, δ, are utilized to find an effective triple (c_S, h, k) so that $E_0(\ell) = A$, and $\{E_\delta(\ell)\}$, in some way, are minimized.

Proper choice of c_S implies $k < c_S < h + k$. For $c_S \leq k$, the above combo would be reduced to a pure Shewhart chart. This is due to the fact that as long as the Shewhart component is not signaling, hence $X_n \leq c_S \leq k$, the CUSUM statistic Z_n will not increase. Thus the Shewhart component will never signal after the CUSUM component. Moreover, a CUSUM chart with $h = 0$ and $k > 0$ is equivalent to a Shewhart chart (by setting $k = c_S$). Therefore, the reference value k of a proper ($h > 0$) CUSUM chart is smaller than the alarm threshold c_S with the same in-control ARL. On the other hand, if $h + k \leq c_S$ then the combo is equivalent to a standalone CUSUM chart. Namely, each X_n that triggers a Shewhart chart alarm is now larger than $h+k$ so that the corresponding $Z_n \geq Z_{n-1}+X_n-k > Z_{n-1}+h \geq h$. Hence, the CUSUM component signals too. Basically, the $k < c_S < h+k$ condition is needed for technical reasons.

For a standalone CUSUM chart, Fig. 1 illustrates the relationship between k and h for an in-control ARL of 1000. The reference value k is usually much smaller than the Shewhart threshold c_S. The actual interval of admissible c_S values is even tighter—the lower limit is given by the threshold of a standalone Shewhart chart, the normal quantile $\Phi^{-1}(1-1/A)$, the upper one by the threshold h_{alone} of a standalone CUSUM chart increased by k:

$$\Phi^{-1}(1-1/A) \leq c_S \leq h_{\text{alone}}(k, A) + k\,. \tag{1}$$

In the sequel we assume that (1) is fulfilled. From Fig. 2 we see that for small $k < 1$, the interval could be even more reduced, because for $c_S > 4.5$ the threshold h does not really change anymore.

Let $\varepsilon = c_S - k$ with $0 < \varepsilon < h$. Then the ARL function of the combo solves the following integral equation:

$$\mathcal{L}(s) = 1 + \Phi(k-s)\mathcal{L}(0) + \int_0^{\min\{h,\varepsilon+s\}} \varphi(z+k-s)\mathcal{L}(z)\,dz\,. \tag{2}$$

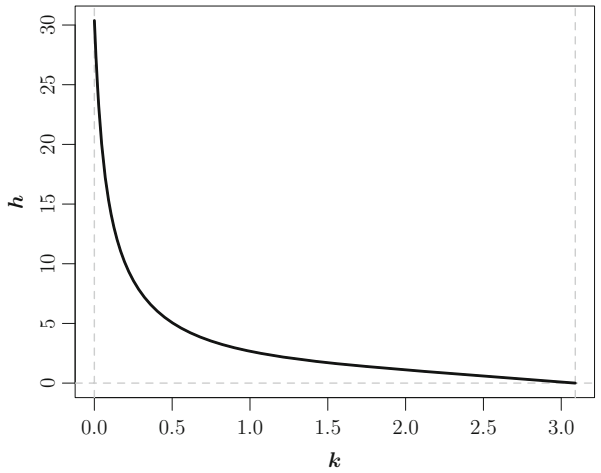


Fig. 1 CUSUM setup: relationship between reference value k and threshold h for an in-control ARL 1000. Admissible k values belong to the interval $\left(0, \Phi^{-1}(1 - 1/1000) = 3.09\right)$

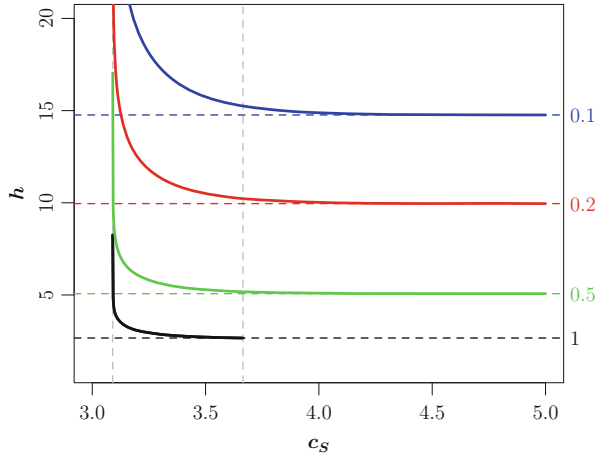


Fig. 2 Combinations of Shewhart threshold c_S and CUSUM's h ($k \in \{1, 0.5, 0.2, 0.1\}$) for an overall (of the combo) in-control ARL 1000

The functions $\Phi()$ and $\varphi()$ constitute the cumulative distribution and probability density function of a standard normal distribution. Replacing the upper integral limit with the constant value h leads to the well-known equation from Page (1954), Lucas (1976), and Vance (1986). Numerical solution of the above integral equation with an integral limit depending on the argument s is not straightforward. See, for instance, Capizzi and Masarotto (2010) for a similar treatment of the EWMA-Shewhart combo. They applied an aptly chosen Clenshaw-Curtis quadrature to obtain satisfying numerical accuracy. Most of the work for combo charts rely

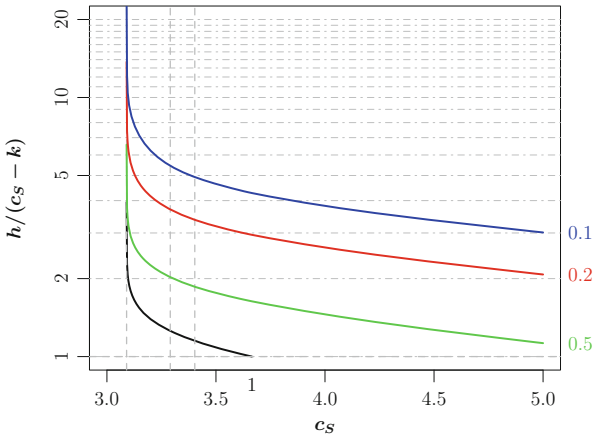


Fig. 3 The number of intervals, $r = \lceil h/(c_S - k) \rceil$, for combinations of Shewhart threshold c_S, CUSUM's h ($k \in \{1, 0.5, 0.2, 0.1\}$), and in-control ARL 1000

on modified Markov chain approximations—see, e. g., Lucas (1982), Yashchin (1985b), Reynolds and Stoumbos (2005), and Wu et al. (2008). Here, we want to exercise collocation with piecewise defined Chebyshev polynomials—see Knoth (2006) for their successful application in case of calculating the ARL of CUSUM charts deploying the sample variance S^2. First, we decompose the interval $[0, h]$ in r subintervals.

$$[0, h] = \big[0, h - (r-1)\varepsilon\big] \cup \big(h - (r-1)\varepsilon, h - (r-2)\varepsilon\big] \cup \ldots \cup \big(h - \varepsilon, h\big] .$$

The integer r is determined from $r = \lceil h/\varepsilon \rceil = \lceil h/(c_S - k) \rceil$. From Fig. 3 one concludes, that for large $k = 1$ (and $k = 0.5$ too), the value $r = 2$ seems to be the typical value, at least for the chosen $A = 1000$. Returning to the subinterval design we ascertain that except the usually shorter first one, all subintervals have the same width ε. The Chebyshev polynomials are defined on all these r intervals accordingly. The collocation framework is described for the simple case $r = 2$—the general case is taken care of in Appendix. Hence, we distinguish for $\mathcal{L}(s)$ the intervals $0 \le s \le h - \varepsilon$ and $h - \varepsilon < s \le h$. The constant $\mathcal{L}(0)$ seems to be another value to be calculated, but because of the continuity of the ARL function it is covered by the first interval. Now, we approximate $\mathcal{L}(s)$ on the mentioned intervals with two different linear combinations of Chebyshev polynomials up to order $N-1$, namely with

$$\sum_{j=1}^{N} c_{1j} T_{1j}(s) \qquad \text{and} \qquad \sum_{j=1}^{N} c_{2j} T_{2j}(s) \qquad .$$

The polynomials $T_{1j}(s)$ and $T_{2j}(s)$ are derived from the standard Chebyshev polynomials

$$u_j(z) = \cos\left(j \arccos(z)\right) \, , \, j = 0, 1, \ldots, N-1 \, , \, z \in [-1, 1]$$

thru mapping $[-1, 1]$ to the related subinterval and adjusting the numbering in j $(0 \to 1)$.

For the first interval, $0 \leq s \leq h - \varepsilon$, we obtain

$$\sum_{j=1}^{N} c_{1j} T_{1j}(s) = 1 + \Phi(k-s)\mathcal{L}(0) + \sum_{j=1}^{N} c_{1j} \int_0^{h-\varepsilon} \varphi(z+k-s) T_{1j}(z)\, dz$$
$$+ \sum_{j=1}^{N} c_{2j} \int_{h-\varepsilon}^{\varepsilon+s} \varphi(z+k-s) T_{2j}(z)\, dz \, ,$$

while for the second one, $h - \varepsilon < s \leq h$, we receive

$$\sum_{j=1}^{N} c_{2j} T_{2j}(s) = 1 + \Phi(k-s)\mathcal{L}(0) + \sum_{j=1}^{N} c_{1j} \int_0^{h-\varepsilon} \varphi(z+k-s) T_{1j}(z)\, dz$$
$$+ \sum_{j=1}^{N} c_{2j} \int_{h-\varepsilon}^{h} \varphi(z+k-s) T_{2j}(z)\, dz \, .$$

Both equations are evaluated at the roots of the Chebyshev polynomial $u_N(z)$ shifted to each of the considered intervals. Hence, a linear equation system with dimension $2N$ has to be solved, eventually. The remaining unknown constant $\mathcal{L}(0)$ is substituted by

$$\mathcal{L}(0) = \sum_{j=1}^{N} c_{1j} T_{1j}(0) = \sum_{j=1}^{N} c_{1j} (-1)^{j+1} \, .$$

In the following subsection the framework is applied for some examples.

2.1 Examples for One-Sided Designs

In order to demonstrate the numerical performance of the collocation design, we look firstly at one configuration utilized in Yashchin (1985b): $k = 1, h = 3, c_S = 3.5$. Consequently, $r = \lceil 3/(3.5-1) \rceil = 2$. With $n = 10$ (matrix dimension 20) we obtain the final ARL approximation, 1510.0 (Monte Carlo with 10^9 replicates

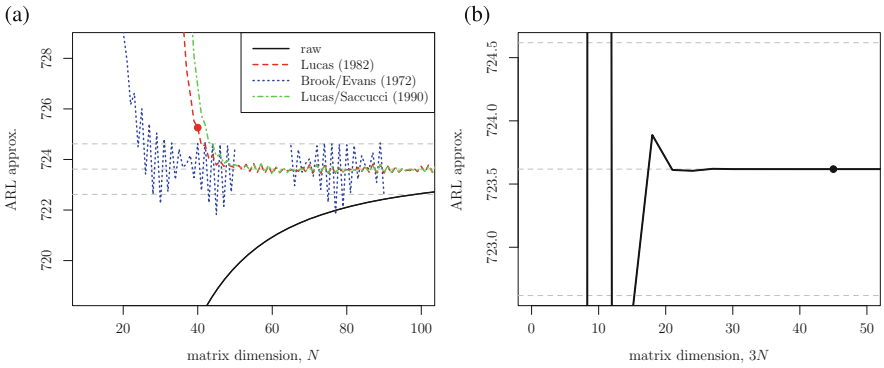


Fig. 4 (In-control) ARL approximation vs. matrix dimension; $k = 0.25, h = 8, c_S = 4$. (**a**) Markov chain. (**b**) Collocation

resulted in 1509.94 with s.e. 0.048), which differs considerably from the value from Yashchin (1985b) in the table printed as Fig. 4, 1507.3.

To illustrate potential accuracy issues, we study the more elaborated results from Lucas (1982) and consider $k = 0.25$ (the smaller k the more severe are the accuracy problems), $h = 8$ and $c_S = 4$ which results in $\varepsilon = 2.13$ and $r = 3$ intervals. In Fig. 4 the related ARL approximations are plotted versus matrix dimension. In Fig. 4a we display besides the "raw" Markov chain values three popular frameworks to improve convergence—the designs deployed by Lucas (1982), Brook and Evans (1972) and Lucas and Saccucci (1990). These utilize 4, 3 and 5 single Markov chain results, respectively, and plug them into the same linear model. For the sake of visibility, we omit some segments for the highly varying profile following Brook and Evans (1972). From Fig. 4a and b we conclude that collocation is more powerful in terms of accuracy. The two bullets mark the selections of N used in Lucas (1982) and for the comparison done in Table 1. In Fig. 5 we illustrate the complete ARL function, based on collocation. The three intervals are marked. Moreover, we want to compare the highly accurate numerical procedure with the Markov chain based results in Lucas (1982). From Lucas (1982), Table 2/Part 3 we take some numbers from the first block. Note that Lucas (1982) calculated his results adjusting all entries within the transition matrix of the Markov chain which correspond to an observation that would violate the Shewhart limit c_S. Then, by calculating the ARL approximation for 10, 20, 30 and 40 states and plugging the results into a simple regression model, he obtained the final results which surprisingly well match the collocation based numbers.

Table 1 Some ARL results from Table 2/Part 3 (upper entry) in Lucas (1982) vs. collocation (middle entry) and Monte Carlo simulation (lower entry, 10^9 rep.)

Parameters			Shift δ								
h	k	c_S	0.00	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
6	0.25	3	202.1	48.19	20.16	11.94	8.344	5.062	3.461	2.471	1.820
			202.0	48.17	20.16	11.93	8.340	5.058	3.458	2.469	1.819
			202.0	48.17	20.15	11.93	8.340	5.058	3.458	2.469	1.819
6	0.25	3.5	241.7	50.79	20.77	12.29	8.640	5.384	3.852	2.910	2.239
			241.8	50.81	20.77	12.29	8.642	5.387	3.855	2.914	2.244
			241.8	50.81	20.77	12.29	8.642	5.387	3.855	2.914	2.244
6	0.25	4	249.7	51.27	20.89	12.36	8.713	5.487	4.013	3.143	2.525
			249.7	51.28	20.89	12.36	8.712	5.487	4.013	3.142	2.525
			249.7	51.27	20.89	12.36	8.712	5.487	4.013	3.142	2.525
8	0.25	3	395.8	73.98	27.04	15.42	10.58	6.194	4.045	2.733	1.912
			396.0	74.02	27.05	15.43	10.58	6.196	4.045	2.732	1.911
			396.0	74.02	27.05	15.43	10.58	6.196	4.045	2.732	1.911
8	0.25	3.5	646.4	82.12	28.44	16.17	11.20	6.835	4.760	3.451	2.511
			645.5	82.12	28.43	16.17	11.20	6.834	4.760	3.450	2.511
			645.5	82.12	28.43	16.17	11.20	6.834	4.760	3.450	2.511
8	0.25	4	725.2	83.75	28.72	16.34	11.36	7.051	5.082	3.888	3.010
			723.6	83.74	28.72	16.34	11.36	7.048	5.078	3.883	3.005
			723.6	83.74	28.72	16.34	11.36	7.048	5.078	3.882	3.005
10	0.25	3	571.3	101.7	33.68	18.75	12.68	7.202	4.509	2.903	1.955
			571.7	101.7	33.68	18.74	12.68	7.202	4.511	2.905	1.956
			571.7	101.7	33.68	18.74	12.68	7.202	4.511	2.905	1.956
10	0.25	3.5	1441	119.9	36.09	20.00	13.71	8.222	5.584	3.897	2.706
			1436	119.9	36.10	20.01	13.71	8.227	5.591	3.904	2.711
			1436	119.9	36.10	20.01	13.71	8.227	5.591	3.904	2.711
10	0.25	4	1974	124.0	36.62	20.31	13.99	8.596	6.119	4.586	3.445
			1956	124.0	36.62	20.31	13.99	8.594	6.116	4.583	3.441
			1956	124.0	36.62	20.31	13.99	8.594	6.116	4.583	3.441

Two further figures illustrate the detection performance of the combo in terms of the zero-state ARL. Thereby, we consider two different $k \in \{0.5, 0.2\}$. Three different c_S are selected: 5, 10 or 20% within the interval of admissible c_S measured from the lower bound (threshold of the standalone Shewhart chart 3.090 and, for example, 3.214, 3.338, 3.586 for $k = 0.5$). From the profiles in Fig. 6 we conclude that for smaller k the impact of c_S is more specific. For both k in $\{0.2, 0.5\}$, unquestionably, adding a Shewhart limit improves considerably the detection performance for changes larger than 2.5. In summary, it looks like a handy improvement of the prim CUSUM procedure.

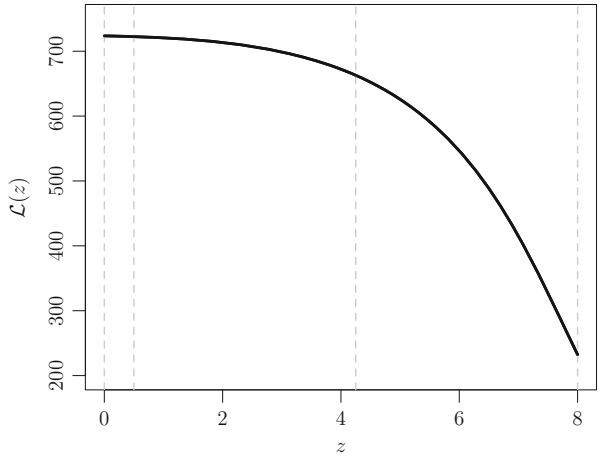


Fig. 5 Complete function $\mathcal{L}(z)$ for $k = 0.25$, $h = 8$, $c_S = 4$, $\varepsilon = 3.75$, $r = 3$ intervals

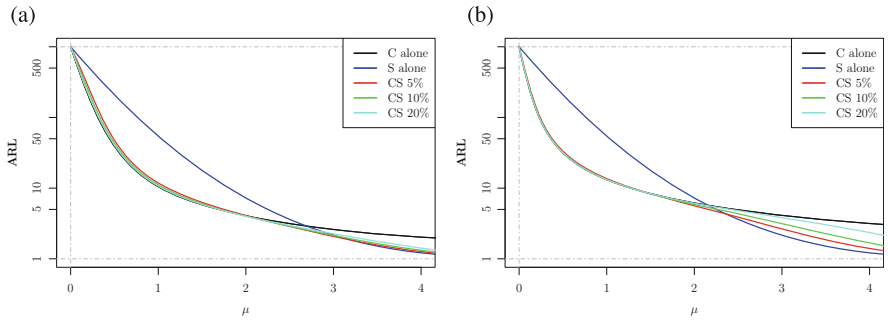


Fig. 6 ARL performance of different CUSUM-Shewhart combos and standalone charts. (**a**) $k = 0.5$. (**b**) $k = 0.2$

3 Two-Sided Case

First prominent discussions of two-sided CUSUM-Shewhart combo's ARL are Lucas and Crosier (1982) and Yashchin (1985a,b). Before we return to them in more detail, we introduce further notation:

$$\text{Shewhart rule } \ell_S^{(2)} = \inf\{i \geq 1 : |X_i| > c_S\}\,.$$

$$Z_0^+ = z_0^+ = 0\,,\ Z_n^+ = \max\{0, Z_{n-1}^+ + X_n - k\}\,,$$

$$\text{upper CUSUM rule } \ell_c^+ = \inf\{n \geq 1 : Z_n^+ > h\}\,.$$

$$Z_0^- = z_0^- = 0\,,\ Z_n^- = \max\{0, Z_{n-1}^- - X_n - k\}\,,$$

$$\text{lower CUSUM rule}\ \ \ell_c^- = \inf\{n \geq 1 : Z_n^- > h\}\,.$$

$$\text{2-sided CUSUM rule}\ \ \ell_c^{(2)} = \min\{\ell_c^+, \ell_c^-\}\,.$$

$$\text{combo rule}\ \ \ell^{(2)} = \min\{\ell_S^{(2)}, \ell_c^{(2)}\}\,.$$

Note that we restrict ourselves to the simple and quite popular CUSUM setup where both reference values (k) and thresholds (h) are equal. The validity of the here presented findings for the general case has to be proved yet.

First, we consider the ARL function for the two-sided CUSUM chart alone, $\ell_c^{(2)}$. By writing $\mathcal{L}(s^+, s^-)$ for the corresponding ARL function, we report the following ARL integral equation, which was derived by considering the values of X (within the usual total probability arguments) and not, as usual, the values of the CUSUM statistic:

$$\begin{aligned}\mathcal{L}(s^+, s^-) = 1 &+ \int_{\max\{k-s^+, s^--k\}}^{h+k-s^+} \varphi(x)\mathcal{L}(s^+ + x - k, 0)\,dx \\ &+ \big(\Phi(k - s^+) - \Phi(s^- - k)\big)\mathcal{L}(0,0) \quad \text{(vanishes if } 2k \leq s^+ + s^-\text{)} \\ &+ \int_{-h-k+s^-}^{\min\{k-s^+, s^--k\}} \varphi(x)\mathcal{L}(0, s^- - x - k)\,dx \\ &+ \int_{\max\{k-s^+, -h-k+s^-\}}^{\min\{s^--k, h+k-s^+\}} \varphi(x)\mathcal{L}(s^+ + x - k, s^- - x - k)\,dx\,.\end{aligned}$$

It turns out that it is reasonable to distinguish the cases (i) $s^+ + s^- \leq 2k$, (ii) $2k < s^+ + s^- \leq h + 2k$, and (iii) $h + 2k < s^+ + s^- \leq 2h$. Starting with (i), we write

$$\begin{aligned}\mathcal{L}(s^+, s^-) = 1 &+ \int_{k-s^+}^{h+k-s^+} \varphi(x)\mathcal{L}(s^+ + x - k, 0)\,dx \\ &+ \big(\Phi(k - s^+) - \Phi(s^- - k)\big)\mathcal{L}(0,0) \\ &+ \int_{-h-k+s^-}^{s^--k} \varphi(x)\mathcal{L}(0, s^- - x - k)\,dx\,. \qquad (3)\end{aligned}$$

Hence, for $s^+ + s^- \leq 2k$, the ARL function is driven exclusively from $\mathcal{L}(\cdot, 0)$, $\mathcal{L}(0, \cdot)$, and $\mathcal{L}(0, 0)$. For slightly larger $s^+ + s^-$, case (ii), we observe

$$\begin{aligned}\mathcal{L}(s^+, s^-) = 1 &+ \int_{s^- - k}^{h+k-s^+} \varphi(x)\mathcal{L}(s^+ + x - k, 0)\, dx \\ &+ \int_{k-s^+}^{s^- - k} \varphi(x)\mathcal{L}(s^+ + x - k, s^- - x - k)\, dx \\ &+ \int_{-h-k+s^-}^{k-s^+} \varphi(x)\mathcal{L}(0, s^- - x - k)\, dx\,. \end{aligned} \tag{4}$$

And the most simple and practically less important case, (iii), yields the following identity:

$$\mathcal{L}(s^+, s^-) = 1 + \int_{-h-k+s^-}^{h+k-s^+} \varphi(x)\mathcal{L}(s^+ + x - k, s^- - x - k)\, dx\,. \tag{5}$$

Conveniently, the arguments of $\mathcal{L}()$ in case (iii) do not appear in the integrals of cases (i) and (ii). Hence, to determine the ARL for all possible head-starts, it is sufficient to solve (i) and (ii). Then we deploy the fact that in case (iii) the sum of arguments in $\mathcal{L}()$ under the integral is $s^+ + s^- - 2k$, hence the original $s^+ + s^-$ is shrunk. This is already smaller than $h + 2k$ or another observation has to be considered. In the most extreme case, $s^+ + s^- = 2h$, $\lceil h/(2k) - 1 \rceil$ steps has to be taken. Finally, by using the solution of $\mathcal{L}(s^+, s^-)$ for $s^+ + s^- \leq h + 2k$, one iterates up to the initial extreme pair (s^+, s^-).

From Lucas and Crosier (1982) we take the much nicer formula eq. (A.1) for $s^+ + s^- \leq h + 2k$—hence (i) and (ii), but not (iii)—to link $\mathcal{L}(s^+, s^-)$ to the ARL function of the simpler one-sided CUSUM chart

$$\mathcal{L}(s^+, s^-) = \frac{\mathcal{L}^+(s^+)\mathcal{L}^-(0) + \mathcal{L}^+(0)\mathcal{L}^-(s^-) - \mathcal{L}^+(0)\mathcal{L}^-(0)}{\mathcal{L}^+(0) + \mathcal{L}^-(0)}\,. \tag{6}$$

It turns out that it solves the integral equation (i)+(ii)—for details see Appendix. Moreover, the restriction introduced by Lucas and Crosier (1982) does not block the simple calculation of the ARL for even more extreme head-start values as in case (iii). As already mentioned, by using the solution for the less extreme values from (i)+(ii) and some quadrature rule based iteration procedure for the integrals, the complete set of possible head-start values could be treated.

Now, we want to modify the integral equation framework in order to incorporate the impact of the additional Shewhart limit c_S. Essentially, $\max\{-c_S, \mathit{lower}\}$ and $\min\{c_S, \mathit{upper}\}$ replace the original limits *lower* and *upper*. Second, in case (ii) only the limits of integrals with $\mathcal{L}(\cdot, 0)$ or $\mathcal{L}(0, \cdot)$ are changed. In case (i), this is true by construction.

Could it be possible that using the results from the previous section and formula (6) would work? For some configurations it would. The subtlety consists in the fact that in (4) for $s^- > c_S + k$ the first integral vanishes. Similarly the third integral gets zero for $s^+ > c_S + k$. Then the mechanism of the proof elicited for the standalone CUSUM in Appendix could not be used anymore. For the configuration of the combo in Yashchin (1985b), it stays intact because for $k = 1$, $h = 3$, and $c_S = 3.5$ the aforementioned inequalities never hold. In general, for $h \leq c_S + k$ the formula (6) provides again an accurate ARL calculation tool. For the configurations considered in Table 1 from Lucas (1982), $k = 0.25$, $h \in \{6, 8, 10\}$, and $c_S \in \{3, 3.5, 4\}$, this does not hold anymore. However, we could use it as handy approximation. In the following section it is demonstrated, how it works in both scenarios.

Hence, the zero-state ARL of a two-sided CUSUM-Shewhart scheme could be calculated as for the standalone two-sided CUSUM chart by deploying the nice formula (A.1) in Lucas and Crosier (1982)—here (6).

3.1 *Examples for Two-Sided Designs*

Again we start with a result from Yashchin (1985b). We re-collect some numbers from Yashchin's Fig. 6 and new results in Table 2. Both, the results by Yashchin (1985b) and the new ones look convincing. The first ones, because despite being 30 years old they are quite close to the true values, and the last ones while being nicely matched by the Monte Carlo confirmation runs. At least the latter should be not too surprising because their accuracy could be "proved".

Turning to similar calculations in Lucas (1982), we have to face two problems. First, Lucas' results seem to be less accurate than Yashchin's ones. Second, the new results based on "believing" the nice rule (6) differ to the Monte Carlo derived results. Again, this does not surprise because we already announced that for the herewith considered configuration, rule (6) is an approximation only. However, the results in Table 3 are reasonably well. To indicate the nearly non-visible differences between rule (6) and the Monte Carlo numbers, all significant (5 % level) differences are marked with bold letters.

Table 2 Two-sided CUSUM-Shewhart ARL results from Yashchin (1985b) and new ones, numerical and Monte Carlo (10^9 rep.); $k = 1$, $h = 4$, $c_S = 3.5$

z_0^+	z_0^-	Yashchin (1985b)	Numerical	MC	MC s.e.
0	0	753.6	754.98	754.98	0.024
1.63	1.63	725.3	726.45	726.46	0.024
1.63	1.83	718.1	719.30	719.32	0.024

Table 3 Some ARL results from Table 2/Part 1 (upper entry) in Lucas (1982) vs. collocation (middle entry) and Monte Carlo simulation (lower entry, 10^9 rep.)

Parameters			Shift δ								
h	k	c_S	0.00	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
6	0.25	3	99.05	45.51	19.86	11.81	8.244	4.974	3.382	2.419	1.789
			101.0	46.12	20.05	11.92	8.338	5.058	3.458	2.469	1.819
			101.0	**46.13**	**20.05**	**11.92**	8.338	5.058	3.458	2.469	1.819
6	0.25	3.5	121.6	49.78	20.79	12.31	8.667	5.419	3.901	2.977	2.319
			120.9	49.63	20.75	12.29	8.641	5.387	3.855	2.914	2.244
			120.9	49.63	**20.75**	12.29	8.641	5.387	3.855	2.914	2.244
6	0.25	4	124.8	50.22	20.86	12.35	8.704	5.474	3.990	3.105	2.473
			124.9	50.24	20.87	12.36	8.712	5.487	4.013	3.142	2.525
			124.8	50.24	20.88	12.36	8.712	5.487	4.013	3.142	2.525
8	0.25	3	188.9	68.01	26.08	14.96	10.24	5.933	3.855	2.640	1.893
			198.0	70.76	26.88	15.40	10.58	6.196	4.045	2.732	1.911
			198.1	**70.82**	**26.89**	**15.41**	**10.58**	6.196	4.045	2.732	1.911
8	0.25	3.5	325.4	81.65	28.51	16.23	11.25	6.894	4.829	3.520	2.567
			322.8	81.19	28.41	16.17	11.20	6.834	4.760	3.450	2.511
			322.8	**81.20**	**28.41**	16.17	11.20	6.834	4.760	3.450	2.511
8	0.25	4	361.4	83.27	28.69	16.32	11.34	7.021	5.034	3.820	2.942
			361.8	83.30	28.71	16.34	11.36	7.048	5.078	3.883	3.005
			361.8	83.30	28.71	16.34	11.36	7.048	5.078	3.883	3.005
10	0.25	3	301.5	101.9	34.92	19.49	13.25	7.628	4.797	3.032	1.987
			285.8	96.02	33.41	18.71	12.67	7.202	4.511	2.905	1.956
			286.0	**96.17**	**33.44**	**18.72**	**12.67**	7.202	4.511	2.904	1.956
10	0.25	3.5	704.1	117.2	35.72	19.78	13.49	7.958	5.263	3.585	2.511
			718.1	118.6	36.06	20.00	13.71	8.227	5.591	3.904	2.711
			718.2	**118.6**	**36.06**	**20.00**	13.71	8.227	5.591	3.904	2.711
10	0.25	4	975.5	124.4	36.73	20.36	14.03	8.651	6.224	4.759	3.682
			978.3	123.7	36.61	20.31	13.99	8.594	6.116	4.583	3.441
			978.3	123.7	36.61	20.31	13.99	8.594	6.116	4.583	3.441

The accuracy problems are more pronounced for the head start results in Table 4—the head start is set to half of the alarm threshold h.

Hence, the here presented method provides quite good approximations, but they do not attain the traditional accuracy of ARL integral equation related methods. Compared, however, to the much more demanding bivariate Markov chain model of Lucas (1982), which is much less accurate in particular for large values of h, it works convincingly well.

Table 4 Some ARL results from Table 2/Part 2 (upper entry) in Lucas (1982) vs. collocation (middle entry) and Monte Carlo simulation (lower entry, 10^9 rep.); CUSUM part with head-start at $h/2$

Parameters		Shift δ									
h	k	c_S	0.00	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
6	0.25	3	79.33	33.53	12.94	7.219	4.951	3.063	2.228	1.752	1.453
			81.47	34.15	13.08	7.286	5.004	3.125	2.310	1.846	1.538
			81.49	**34.17**	**13.08**	**7.287**	**5.005**	3.125	2.310	1.846	1.538
6	0.25	3.5	96.75	36.33	13.34	7.366	5.053	3.175	2.391	1.983	1.732
			96.66	36.35	13.33	7.361	5.048	3.168	2.371	1.932	1.638
			96.67	**36.35**	13.33	7.361	5.049	3.168	2.371	1.932	1.638
6	0.25	4	99.14	36.57	13.35	7.364	5.048	3.166	2.374	1.952	1.684
			99.63	36.72	13.37	7.371	5.053	3.169	2.372	1.932	1.638
			99.63	36.72	13.37	7.371	5.053	3.169	2.372	1.932	**1.638**
8	0.25	3	161.9	51.10	16.63	8.959	6.070	3.667	2.585	1.954	1.550
			171.8	53.62	17.15	9.201	6.247	3.815	2.722	2.068	1.630
			171.9	**53.68**	**17.16**	**9.203**	**6.247**	3.815	2.722	2.068	**1.630**
8	0.25	3.5	278.3	60.74	17.77	9.406	6.385	3.951	2.881	2.233	1.774
			277.4	60.58	17.72	9.377	6.363	3.931	2.871	2.254	1.838
			277.5	**60.58**	**17.72**	9.377	6.363	3.930	**2.871**	2.254	1.838
8	0.25	4	306.4	61.72	17.80	9.402	6.379	3.949	2.897	2.290	1.895
			310.4	61.96	17.82	9.409	6.387	3.962	2.927	2.355	1.992
			310.4	61.96	17.82	9.409	6.387	3.962	2.928	2.355	1.992
10	0.25	3	275.8	78.59	21.93	11.49	7.777	4.765	3.412	2.560	1.926
			261.0	73.72	20.95	11.04	7.452	4.478	3.123	2.306	1.757
			261.2	**73.89**	**20.98**	**11.05**	**7.452**	**4.478**	**3.123**	2.306	1.757
10	0.25	3.5	633.5	87.78	21.73	11.26	7.574	4.575	3.231	2.426	1.880
			650.3	89.18	21.91	11.36	7.667	4.691	3.381	2.598	2.041
			650.4	**89.22**	**21.91**	11.36	7.668	4.692	3.380	**2.598**	2.041
10	0.25	4	877.2	92.93	22.16	11.45	7.732	4.772	3.514	2.785	2.241
			884.3	92.65	22.09	11.42	7.715	4.752	3.475	2.738	2.225
			884.3	92.65	22.09	11.42	7.715	4.752	3.475	2.738	2.225

4 Conclusions

New numerical methods are presented that provide high and medium accuracy for the ARL of one- and two-sided CUSUM-Shewhart schemes, respectively, for detecting changes in the normal mean over a broad range of potential shifts. A more elaborated numerical algorithm (two-dimensional) in the lines of the collocation procedure in Sect. 2 could and should be created to resolve the remaining accuracy gap. Notwithstanding, the numerical performance is sufficient for practical problems.

Appendix 1: Collocation Design for More Than $r = 2$ Intervals

Here we provide the generalization from $r = 2$, dealt with in Sect. 2, to general $r \in \{2, 3, \ldots\}$. To start with the first interval, $0 \le s \le h - \varepsilon$, we simply state that the shape of the collocation design does not change for greater r. For the succeeding intervals, $h-(r-m+1)\varepsilon < s \le h-(r-m)\varepsilon$ with $m = 2, \ldots, r-1$, the following structure is utilized.

$$\begin{aligned}\sum_{j=1}^{N} c_{mj} T_{mj}(s) &= 1 + \Phi(k-s)\mathcal{L}(0) \\ &\quad + \sum_{j=1}^{N} c_{1j} \int_0^{h-(r-1)\varepsilon} \varphi(z+k-s) T_{1j}(z)\, dz \\ &\quad + \sum_{t=2}^{m} \sum_{j=1}^{N} c_{tj} \int_{h-(r-t+1)\varepsilon}^{h-(r-t)\varepsilon} \varphi(z+k-s) T_{tj}(z)\, dz \\ &\quad + \sum_{j=1}^{N} c_{m+1,j} \int_{h-(r-m+1)\varepsilon}^{\varepsilon+s} \varphi(z+k-s) T_{m+1,j}(z)\, dz\,.\end{aligned}$$

Note that these equations are not present for $r = 2$. However, the last interval, $h-\varepsilon < s \le h$, is considered for all $r \ge 2$. The general structure of the corresponding collocation equation is similar to the above one (now with $m = r$) except for the upper limit of the last integral where $\varepsilon + s$ has to be replaced by h.

Appendix 2: Two-Sided CUSUM Chart

Starting with case (i), $s^+ + s^- \le 2k$, and re-writing the corresponding integral equation results in:

$$\begin{aligned}\mathcal{L}(s^+, s^-) &= \Phi(k-s^+)\mathcal{L}(0,0) + \int_{k-s^+}^{h+k-s^+} \varphi(x)\mathcal{L}(s^+ + x - k, 0)\, dx \\ &\quad + \Phi(k-s^-)\mathcal{L}(0,0) + \int_{k-s^-}^{h+k-s^-} \varphi(-x)\mathcal{L}(0, s^- + x - k)\, dx \\ &\quad + 1 - \mathcal{L}(0,0)\,.\end{aligned}$$

Setting s^+ or s^- to zero in (6) yields:

$$\mathcal{L}(0,0) = \frac{\mathcal{L}^+(0)\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)},$$
$$\mathcal{L}(s^+,0) = \frac{\mathcal{L}^+(s^+)\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}, \quad \mathcal{L}(0,s^-) = \frac{\mathcal{L}^+(0)\mathcal{L}^-(s^-)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}$$

Using this, the first line of the integral equation's right-hand side changes to

$$\frac{\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}\left(\Phi(k-s^+)\mathcal{L}^+(0) + \int_{k-s^+}^{h+k-s^+} \varphi(x)\mathcal{L}^+(s^+ + x - k)\,dx\right)$$

Substituting $x = z + k - s$ in (2) while replacing the upper limit by h results in

$$\mathcal{L}(s) - 1 = \Phi(k-s)\mathcal{L}(0) + \int_{k-s}^{h+k-s} \varphi(x)\mathcal{L}(s+x-k)\,dx, \tag{7}$$

so that the line under analysis simplifies heavily to

$$\frac{\mathcal{L}^-(0)(\mathcal{L}^+(s^+)-1)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}.$$

In a similar way we treat the second line ending in

$$\frac{\mathcal{L}^+(0)(\mathcal{L}^-(s^-)-1)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}.$$

For the second line we made use of $\varphi(-x) = \varphi(x)$ in the in-control case ($\delta = 0$), while for $\delta \neq 0$ we have to change the sign of δ, hence $\varphi_\delta(-x) = \varphi_{-\delta}(x)$. All together resembles (the "1" consumes the disturbing parts of the above two ratios)

$$\frac{\mathcal{L}^+(s^+)\mathcal{L}^-(0) + \mathcal{L}^-(s^-)\mathcal{L}^+(0) - \mathcal{L}^+(0)\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \quad \text{which confirms (6).}$$

Turning to case (ii), $2k < s^+ + s^- \leq h + 2k$, we recall the shape of the related integral equation:

$$\begin{aligned}\mathcal{L}(s^+,s^-) = 1 &+ \int_{s^- - k}^{h+k-s^+} \varphi(x)\mathcal{L}(s^+ + x - k, 0)\,dx \\ &+ \int_{k-s^+}^{s^- - k} \varphi(x)\mathcal{L}(s^+ + x - k, s^- - x - k)\,dx \\ &+ \int_{-h-k+s^-}^{k-s^+} \varphi(x)\mathcal{L}(0, s^- - x - k)\,dx.\end{aligned}$$

First we plug (6) into and transform the second line

$$\begin{aligned}
&\int_{k-s^+}^{s^--k} \varphi(x)\mathcal{L}(s^+ + x - k, s^- - x - k)\,dx \\
&\quad = \frac{\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \int_{k-s^+}^{s^--k} \varphi(x)\mathcal{L}^+(s^+ + x - k)\,dx \\
&\qquad + \frac{\mathcal{L}^+(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \int_{k-s^+}^{s^--k} \varphi(x)\mathcal{L}^-(s^- - x - k)\,dx \\
&\qquad - \frac{\mathcal{L}^-(0)\mathcal{L}^+(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \int_{k-s^+}^{s^--k} \varphi(x)\,dx\,,
\end{aligned}$$

with the last integral subsequently reduced to $\Phi(s^- - k) - \Phi(k - s^+)$. We rewrite the first line as for case (i) and merge, borrowing $\mathcal{L}^-(0)/\big(\mathcal{L}^+(0)+\mathcal{L}^-(0)\big)$,

$$\begin{aligned}
&\frac{\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \\
&+ \frac{\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \Phi(k - s^+)\mathcal{L}^+(0) \\
&+ \frac{\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \int_{k-s^+}^{s^--k} \varphi(x)\mathcal{L}^+(s^+ + x - k)\,dx \\
&+ \frac{\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \int_{s^--k}^{h+k-s^+} \varphi(x)\mathcal{L}^+(s^+ + x - k)\,dx
\end{aligned}$$

to get

$$\frac{\mathcal{L}^+(s^+)\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}$$

by applying again (7). Exploiting $\Phi(s^- - k) = 1 - \Phi(k - s^-)$ we proceed in a similar way with the third line by collecting after transforming both integrals as in the first case

$$\begin{aligned}
&\frac{\mathcal{L}^+(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \\
&+ \frac{\mathcal{L}^+(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \Phi(k - s^-)\mathcal{L}^-(0)
\end{aligned}$$

$$+\frac{\mathcal{L}^+(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}\int_{s^--k}^{k-s^+}\varphi(x)\mathcal{L}^-(s^-+x-k)\,dx$$
$$+\frac{\mathcal{L}^+(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}\int_{k-s^+}^{h+k-s^-}\varphi(x)\mathcal{L}^-(s^-+x-k)\,dx$$

which results in

$$\frac{\mathcal{L}^-(s^-)\mathcal{L}^+(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}.$$

The two "borrowed" terms

$$-\frac{\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}-\frac{\mathcal{L}^+(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}=-1$$

are compensated with the 1 on the right-hand side of the original equation. The last remaining term forms together with the two others

$$\frac{\mathcal{L}^+(s^+)\mathcal{L}^-(0)+\mathcal{L}^-(s^-)\mathcal{L}^+(0)-\mathcal{L}^+(0)\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)}.$$

This completes the proof.

References

Abel, V. (1990). On one-sided combined Shewhart-CUSUM quality control schemes for Poisson counts. *Computational Statistics Quarterly, 6*(1), 31–39.

Abujiya, M. R., Riaz, M., & Lee, M. H. (2013). Improving the performance of combined Shewhart-cumulative SumControl charts. *Quality and Reliability Engineering International, 29*(8), 1193–1206.

Blacksell, S. D., Gleeson, L. J., Lunt, R. A., & Chamnanpood, C. (1994). Use of combined Shewhart-CUSUM control charts in internal quality control of enzyme-linked immunosorbent assays for the typing of foot and mouth disease virus antigen. *Revue Scientifique et Technique, 13*(3), 687–699.

Brook, D., & Evans, D. A. (1972). An approach to the probability distribution of CUSUM run length. *Biometrika, 59*(3), 539–549.

Capizzi, G., & Masarotto, G. (2010). Evaluation of the run-length distribution for a combined Shewhart-EWMA control chart. *Statistics and Computing, 20*(1), 23–33.

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

Gibbons, R. D. (1999). Use of combined Shewhart-CUSUM control charts for ground water monitoring applications. *Ground Water, 37*(5), 682–691.

Henning, E., Konrath, A. C., da Cunha Alves, C., Walter, O. M. F. C., & Samohyl, R. W. (2015). Performance of a combined Shewhart-Cusum control chart with binomial data for large shifts in the process mean. *International Journal of Engineering Research and Application, 5*(8), 235–243.

Knoth, S. (2006). Computation of the ARL for CUSUM-S^2 schemes. *Computational Statistics and Data Analysis, 51*(2), 499–512.

Lucas, J. M. (1976). The design and use of V-mask schemes. *Journal of Quality Technology, 8*(1), 1–12.

Lucas, J. M. (1982). Combined Shewhart-CUSUM quality control schemes. *Journal of Quality Technology, 14*(2), 51–59.

Lucas, J. M., & Crosier, R. B. (1982). Fast initial response for CUSUM quality-control schemes: Give your CUSUM a head start. *Technometrics, 24*(3), 199–205.

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

Montgomery, D. C. (2009). *Statistical Quality Control: A Modern Introduction* (6th ed., internat. student version edition). Hoboken: Wiley.

Morais, M. C., & Pacheco, A. (2006). Combined CUSUM-Shewhart schemes for binomial data. *Economic Quality Control, 21*(1), 43–57.

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

Qu, L., Wu, Z., & Liu, T.-I. (2011). A control scheme integrating the T chart and TCUSUM chart. *Quality and Reliability Engineering International, 27*(4), 529–539.

Reynolds, M. R., & Stoumbos, Z. G. (2005). Should Exponentially Weighted Moving Average and Cumulative Sum Charts Be Used With Shewhart Limits? *Technometrics, 47*(4), 409–424.

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

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

Starks, T. H. (1988). Evaluation of control chart methodologies for RCRA waste sites. Technical Report 37480, U.S. Environmental Protection Agency (EPA).

Vance, L. C. (1986). Average run lengths of cumulative sum control charts for controlling normal means. *Journal of Quality Technology, 18*(3), 189–193.

Westgard, J. O., Groth, T., Aronsson, T., & de Verdier, C.-H. (1977). Combined Shewhart-Cusum control chart for improved quality control in clinical chemistry. *Clinical Chemistry, 23*(10), 1881–1887.

Wu, Z., Yang, M., Jiang, W., & Khoo, M. B. C. (2008). Optimization designs of the combined Shewhart-CUSUM control charts. *Computational Statistics and Data Analysis, 53*(2), 496–506.

Yashchin, E. (1985a). On a unified approach to the analysis of two-sided cumulative sum control schemes with headstarts. *Advances in Applied Probability, 17*, 562–593.

Yashchin, E. (1985b). On the analysis and design of CUSUM-Shewhart control schemes. *IBM Journal of Research and Development, 29*(4), 377–391.

Optimal Design of the Shiryaev–Roberts Chart: Give Your Shiryaev–Roberts a Headstart

Aleksey S. Polunchenko

Abstract We offer a numerical study of the effect of headstarting on the performance of a Shiryaev–Roberts (SR) chart set up to control the mean of a normal process. The study is a natural extension of that previously carried out by Lucas and Crosier (Technometrics 24(3):199–205, 1982. https://doi.org/10.2307/1268679) for the CUSUM scheme. The Fast Initial Response (FIR) feature exhibited by a headstarted CUSUM turns out to be also characteristic of an SR chart (re-)started off a positive initial score. However, our main result is the observation that a FIR SR with a carefully designed *optimal* headstart is not just faster to react to an initial out-of-control situation, it is nearly *the* fastest *uniformly*, i.e., assuming the process under surveillance is equally likely to go out of control effective any sample number. The performance improvement is the greater, the fainter the change. We explain our optimization strategy, and tabulate the optimal initial score, control limit, and the corresponding "worst possible" out-of-control Average Run Length (ARL), considering mean-shifts of diverse magnitudes and a wide range of levels of the in-control ARL.

Keywords Quality control · Shirayev–Roberts chart · Fast Initial Response (FIR) feature

1 Introduction

The main problem addressed in this work is that of optimal design of the Shiryaev–Roberts (SR) chart, originally proposed by Shiryaev (1961, 1963) and Roberts (1966), and later generalized by Moustakides et al. (2011). Recall that the classical SR chart set up to detect a possible change in the baseline mean of a series of

A. S. Polunchenko (✉)
Department of Mathematical Sciences, State University of New York at Binghamton, Binghamton, NY, USA
e-mail: aleksey@binghamton.edu

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_4

independent samples $X_1, X_2, \ldots$ drawn from a normal unit-variance population at regular time intervals involves sequential evaluation of the SR statistic $\{R_n\}_{n\geq 0}$ using the recurrence $R_n = (1 + R_{n-1}) \exp\{S_n\}$, $n = 1, 2, \ldots$, with $R_0 = 0$, and where the quantity

$$S_n \triangleq \mu \left(X_n - \frac{\mu}{2} \right) \tag{1}$$

is a numerical score that captures the severity of the deviation of the n-th sample point X_n from the target mean-value in either direction; the score function S_n assumes that the intended (target) mean-value of the data is zero, but it is anticipated to change abruptly and permanently to a known off-target value $\mu \neq 0$. The n-th observation X_n might represent a single reading or the average of a batch of observations from a designated routine sampling plan. The chart triggers an alarm at the first stage, $\mathcal{S}_A$, such that $R_{\mathcal{S}_A} \geq A$, where $A > 0$ is a control limit (detection threshold) set in advance in accordance with the desired level of the false alarm risk; more formally, $\mathcal{S}_A \triangleq \min\{n \geq 1 \colon R_n \geq A\}$, where $A > 0$ is given. Hence the process $\{X_n\}_{n\geq 1}$ is considered to be in control until stage $\mathcal{S}_A$. The random variable, $\mathcal{S}_A$, referred to as the run length, is the stage at which sampling stops and appropriate action is taken. A brief account of the history of the SR chart was recently offered by Pollak (2009). For an up-to-date summary of the classical as well as Generalized SR charts' optimality properties, see, e.g., Polunchenko and Tartakovsky (2012).

Though nowhere nearly as known and as widespread as Page's (1954) celebrated CUSUM "inspection scheme", the SR chart did receive some attention in the applied literature. One of the earliest investigations of the chart's characteristics is due to Roberts (1966), who offered a performance comparison of the chart against a host of other statistical process control procedures, including the CUSUM scheme and the EWMA chart (also introduced by Roberts (1959)). A similar type of SR-vs-CUSUM comparison (but with respect to a different criterion and for a different data model) was also later performed by Mevorach and Pollak (1991). See also, e.g., Tartakovsky and Ivanova (1992), Tartakovsky et al. (2009), and Moustakides et al. (2009). Certain data-analytic advantages of the chart over the CUSUM scheme were pointed out by Kenett and Pollak (1996). Kenett and Pollak (1986) provided an example of an application of the SR chart in the area of software reliability.

In the (more theoretical) area of quickest change-point detection, the SR chart received far more attention. To a large extent this is due to the fundamental work of Shiryaev (1961, 1963) who proved that the chart solves a particular Bayesian version of the quickest change-point detection problem; see also Girshick and Rubin (1952). The chart then remained unnoticed until recently Pollak and Tartakovsky (2009) and Shiryaev and Zryumov (2009) discovered that it solves yet another so-called *multi-cyclic* or *generalized* Bayesian version of the quickest change-point detection problem; the multi-cyclic setup is instrumental in such applications as cybersecurity (see, e.g., Tartakovsky et al. 2013), financial monitoring (see, e.g., Pepelyshev and Polunchenko 2017), and economic design of control charts (which

is a major area of research in quality control that originated in the fundamental work of Duncan (1956)). This brought the SR chart back into the spotlight. Polunchenko et al. (2017) performed a robustness analysis of the SR chart's multi-cyclic capabilities when the post-change distribution involves a misspecified parameter. Moustakides et al. (2011) observed that by starting the SR statistic $\{R_n\}_{n\geq 0}$ off a positive initial value, i.e., setting $R_0 = r > 0$, the SR chart can be made nearly the best (in the minimax sense of Pollak (1985)). Roughly, this means the SR chart is almost the fastest to react to a change in the observations' distribution when the corresponding unknown change-point is equally likely to be any point in time; see Sect. 2 for a formal definition. As a matter of fact Polunchenko and Tartakovsky (2010) and Tartakovsky and Polunchenko (2010) demonstrated that in two specific change-point scenarios the SR chart with a carefully designed headstart is *the* fastest (in the sense of Pollak (1985)). This result was then extended by Tartakovsky et al. (2012) who proved that the SR chart whose headstart is selected in a specific fashion is almost the "best one can do" (again, in the sense of Pollak (1985)) asymptotically, as the false alarm risk tends to zero, in a general change-point scenario.

In spite of the aforementioned strong theoretically established optimality properties of the SR chart, and the fact that no such properties are exhibited by either the CUSUM scheme or the EWMA chart, applications of the SR chart in quality control remain very few. In part, this may be due to the lack of existing resources offering pre-computed, for a variety of cases, optimal headstart and control limit values. To the best of our knowledge, the work of Tartakovsky et al. (2009) and that of Polunchenko and Sokolov (2014) have heretofore been the only sources with such data (computed assuming the observations are exponential). This work's goal is to optimize the SR chart for yet another model, namely, the standard Gaussian model widely used in the quality control literature as a testbed for charts' performance analysis. The specific optimization strategy is presented in Sect. 2. The optimization itself is carried out in Sect. 3 using the numerical framework developed by Moustakides et al. (2011) and then improved upon by Polunchenko et al. (2014a,b). The obtained optimal headstart and control limit values are reported in Sect. 3 as well. Conclusions follow in Sect. 4.

2 The Shiryaev–Roberts Chart, Its Properties and Optimization

To control the mean of a standard Gaussian process, the headstarted tweak of the classical SR chart proposed by Moustakides et al. (2011) operates by sequentially updating the statistic $\{R_n^r\}_{n\geq 0}$ via the recurrence

$$R_n^r = (1 + R_{n-1}^r)\exp\{S_n\},\ n = 1, 2, \ldots \text{ with } R_0^r = r \geq 0, \tag{2}$$

where S_n is the score function defined in (1); the initial score $R_0^r = r \geq 0$ is a design parameter also referred to as the headstart, which is the original terminology of Lucas and Crosier (1982) who suggested to headstart the CUSUM scheme. The corresponding run length is as follows:

$$\mathcal{S}_A^r \triangleq \min\{n \geq 1 : R_n^r \geq A\}, \tag{3}$$

where $A > 0$ is the control limit (detection threshold) selected in advance so as to keep the chart's false alarm characteristics tolerably low. Note that if $r = 0$ then the chart is the classical SR chart (with no headstart) of Shiryaev (1961, 1963) and Roberts (1966). For this reason Tartakovsky et al. (2012) coined the term "*Generalized* SR chart" (or the GSR chart for short) to refer to the headstarted SR chart defined by (2) and (3). It is also worth reiterating that the score function (1)—and hence also the statistic (2)—is indifferent to the direction of the mean-shift, i.e., the sign of $\mu \neq 0$ is irrelevant.

In quality control, the operating characteristics of a control chart are customarily assessed by means of two major performance indices: the in-control Average Run Length (ARL) and the out-of-control ARL. It is of note that while the in-control ARL has a clear definition (it is simply the average number of samples taken by the chart before a false out-of-control signal), its out-of-control counterpart is not as straightforward, and can refer, e.g., to the zero-state ARL, or to the cyclical steady-state ARL, or to the conditional steady-state ARL. See, e.g., Knoth (2006) for an overview of the various ways to define the out-of-control ARL used in the quality control literature. In this work, we shall adapt the (more exhaustive) approach used in the quickest change-point detection literature. Let $\mathbb{P}_k$ ($\mathbb{E}_k$) denote the probability measure (expectation) induced by the data $\{X_n\}_{n\geq 1}$ assuming the change-point is at time moment $k = 0, 1, 2, \ldots, \infty$, i.e., assuming the process $\{X_n\}_{n\geq 1}$ is in-control until sample number k inclusive, and is out-of-control starting from sample number $k+1$ onward. The notation $k = 0$ ($k = \infty$) is to be understood as the case when the process under surveillance is out of control *ab initio* (never, respectively).

In change-point detection, the main in-control characteristic of a control chart is the Average Run Length (ARL) to false alarm $\mathrm{ARL}(T) \triangleq \mathbb{E}_\infty[T]$, As is evident from the definition, it is the average number of samples taken by the chart before an *erroneous* out-of-control signal is given. This is precisely what is known in the quality control literature as the *in-control* ARL. It is apparent that the higher the ARL to false alarm, the lower the level of the false alarm risk. For the GSR chart, the general inequality $\mathrm{ARL}(\mathcal{S}_A^r) \geq A - r$ can be used to design $A > 0$ and $r \in [0, A]$ so as to have $\mathrm{ARL}(\mathcal{S}_A^r)$ no lower than a desired margin $\gamma > 1$. It is of note that this inequality holds in general, whatever the statistical structure of the observations be. A more accurate result is the asymptotic (as $A \to +\infty$) approximation $\mathrm{ARL}(\mathcal{S}_A^r) \approx A/\xi - r$, which is actually known to be quite accurate, even if $A > 0$ is not high; see, e.g., (Pollak 1987, Theorem 1) or Tartakovsky et al. (2012). Here ξ denotes the so-called "limiting average exponential overshoot"—a model-dependent constant (taking values between 0 and 1) computable using nonlinear renewal-theoretic methods; see, e.g., Woodroofe (1982). For the Gaussian model considered in this

work it follows, e.g., from Woodroofe (1982, Example 3.1, pp. 32–33), that the following formula can be used:

$$\xi = \frac{2}{\mu^2} \exp\left\{-2\sum_{m=1}^{\infty} \frac{1}{m} \Phi\left(-\frac{|\mu|}{2}\sqrt{m}\right)\right\}, \tag{4}$$

where

$$\Phi(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

is the standard Gaussian cumulative distribution function. Note from the foregoing formula that ξ is an even function of $\mu \neq 0$. The formula was put to use by Woodroofe (1982) who computed ξ for various values of $\mu > 0$; see Woodroofe (1982, Table 3.1, p. 33) for the obtained results.

To quantify the capabilities of a control chart T when the process is no longer in control, Pollak (1985) suggested to use the "worst-case" (Supremum) Average Detection Delay (SADD), conditional on no false alarm having been sounded. Formally,

$$\mathrm{SADD}(T) \triangleq \max_{0 \le k < \infty} \mathrm{ADD}_k(T),$$

where $\mathrm{ADD}_k(T) \triangleq \mathbb{E}_k[T - k | T > k]$, $k = 0, 1, 2, \ldots$. Incidentally, the limiting ADD value $\lim_{k\to\infty} \mathrm{ADD}_k(T)$ is known in the quality control literature as the *conditional steady-state* ARL; see, e.g., Knoth (2006) and the references therein.

Pollak's (1985) criterion has a simple interpretation: for any fixed but finite $k = 0, 1, 2, \ldots$, the condition $T > k$ guarantees that it is an actual detection (i.e., not a false alarm), so that each $\mathrm{ADD}_k(T)$ is the average number of samples it takes the chart past the change-point k to realize that the process under surveillance is not in control anymore, and because k is unknown, it is reasonable to assume it equally likely to be any number $(0, 1, 2, \ldots)$ and consider the worst possible case, i.e., take the maximal of the $\mathrm{ADD}_k(T)$'s. For the CUSUM scheme with no headstart and for the classical SR chart (also headstart-free) it can be shown that $k = 0$ is when the ADD is the highest, i.e., $\mathrm{SADD}(T) = \mathrm{ADD}_0(T)$. As a result, it suffices to restrict attention to just $\mathrm{ADD}_0(T)$, and it is this quantity that the quality control community calls the zero-state out-of-control ARL. However, things are not as simple when the chart has a positive headstart, for, in that case, it is no longer obvious which of the delays $\mathrm{ADD}_k(\mathcal{S}_A^r)$'s for $k = 0, 1, 2, \ldots$ is the highest. As a matter of fact we shall see in the next section that the "bump" of the sequence $\{\mathrm{ADD}_k(\mathcal{S}_A^r)\}_{k\ge 0}$ has a highly unpredictable behavior in terms of its location on the time axis.

Let $\Delta(\gamma) \triangleq \{T\colon \mathrm{ARL}(T) \ge \gamma\}$ be the class of control charts (identified with a generic run length T) whose ARL to false alarm is at least as high as a desired pre-set level $\gamma > 1$. Pollak's (1985) minimax change-point detection problem consists in finding $T_{\mathrm{opt}} \in \Delta(\gamma)$ such that $\mathrm{SADD}(T_{\mathrm{opt}}) = \min_{T\in\Delta(\gamma)} \mathrm{SADD}(T)$ for any given

$\gamma > 1$. In general, this problem is still an open one, although there has been a continuous effort to solve it. To that end, for at least two specific data models, the answer was shown to be the GSR chart with "finetuned" threshold and headstart values; see Polunchenko and Tartakovsky (2010) and Tartakovsky and Polunchenko (2010). Moreover, for a general data model, the GSR chart (properly optimized) was also shown (by Tartakovsky et al. 2012) to solve Pollak's (1985) problem asymptotically as $\gamma \to +\infty$. Specifically, this means that if A and r are selected so that $\mathrm{ARL}(\mathcal{S}_A^r) \geq \gamma$ with $\gamma > 1$ given, i.e., $\mathcal{S}_A^r \in \Delta(\gamma)$, then

$$\mathrm{SADD}(\mathcal{S}_A^r) - \min_{T\in\Delta(\gamma)} \mathrm{SADD}(T) \to 0 \ \text{ as } \ \gamma \to +\infty, \tag{5}$$

provided, however, that $r/A \to 0$ as $A \to +\infty$; see Tartakovsky et al. (2012), who also supply a high-order large-γ expansion of $\mathrm{SADD}(\mathcal{S}_A^r)$. The foregoing is a strong optimality property known in the literature on change-point detection as asymptotic minimax optimality of order three, or asymptotic near minimaxity. It is noteworthy that the CUSUM chart, whether headstarted or not, does not have such strong "nearly-best" detection capabilities. Moreover, nor does the EWMA chart. Hence, our interest in the GSR chart. To provide an idea as to the difference made by a positive headstart, we remark that the classical SR chart (with zero headstart) is asymptotically (as $\gamma \to +\infty$) minimax of order two, i.e., the difference $\mathrm{SADD}(\mathcal{S}_A) - \min_{T\in\Delta(\gamma)} \mathrm{SADD}(T)$ goes to a positive constant as $\gamma \to +\infty$. Moreover, since the constant is the higher, the fainter the change, giving an SR chart a positive headstart is especially beneficial when the out-of-control behavior of the process differs from its in-control behavior only slightly.

Yet another strong optimality property of the GSR chart is its exact multi-cyclic or generalized Bayesian optimality. Specifically, Pollak and Tartakovsky (2009) and Shiryaev and Zryumov (2009) proved that the classical SR chart (with no headstart) minimizes the so-called Integral ADD

$$\mathrm{IADD}(T) \triangleq \sum_{k=0}^{\infty} \mathbb{E}_k[\max\{0, T-k\}], \tag{6}$$

and the so-called Relative IADD (RIADD)

$$\mathrm{RIADD}(T) \triangleq \mathrm{IADD}(T)/\,\mathrm{ARL}(T) = \sum_{k=0}^{\infty} \frac{\mathbb{P}_\infty(T>k)}{\mathrm{ARL}(T)}\,\mathrm{ADD}_k(T), \tag{7}$$

both inside the class $\Delta(\gamma)$ defined above, for any $\gamma > 1$. The meaning of this result can be explained by analyzing the structure of the definition (7) of $\mathrm{RIADD}(T)$. Specifically, on the one hand, the latter can be viewed as being the average of the delays $\mathbb{E}_k[\max\{0, T-k\}]$, $k = 0, 1, 2, \ldots$, taken with respect to the change-point k assuming that the latter has an improper uniform distribution on the set $\{0, 1, 2, \ldots\}$. The improper uniformity of the change-point is a core assumption of the generalized

Bayesian change-point detection problem. On the other hand, RIADD(T) can also be regarded as the average of the $\mathrm{ADD}_k(T)$'s taken with respect to k assuming that the probability mass function of k is given by the ratio $\mathbb{P}_\infty(T > k)/\,\mathrm{ARL}(T)$, $k = 0, 1, 2, \ldots$; note that $\mathbb{P}_k(T > k) \equiv \mathbb{P}_\infty(T > k)$ for any $k = 0, 1, 2, \ldots$, and that $\mathrm{ARL}(T) = \sum_{k=0}^{\infty} \mathbb{P}_\infty(T > k)$. For yet another, viz. multi-cyclic interpretation, see Pollak and Tartakovsky (2009) and Shiryaev and Zryumov (2009), who showed that the RIADD(T)-metric defined in (7) is mathematically equivalent to the so-called Stationary ADD (STADD) which is formally defined next.

The RIADD-optimality of the classical SR chart was generalized in (Polunchenko and Tartakovsky 2010, Lemma 1) where it was shown that the GSR chart, whose control limit $A > 0$ and headstart $r \geq 0$ are such that $\mathrm{ARL}(\mathcal{S}_A^r) \geq \gamma$ for a given $\gamma > 1$, minimizes the Stationary ADD (STADD)

$$\mathrm{STADD}(T) \triangleq (r\,\mathrm{ADD}_0(T) + \mathrm{IADD}(T)) \,\big/ (\mathrm{ARL}(T) + r) \tag{8}$$

inside class $\Delta(\gamma)$, for any $\gamma > 1$; recall that IADD(T) is as in (6). Formally, for any $\gamma > 1$, and any $A > 0$ and $r \geq 0$, it holds true that $\mathrm{STADD}(\mathcal{S}_A^r) = \min_{T \in \Delta(\gamma)} \mathrm{STADD}(T)$, provided that $\mathrm{ARL}(\mathcal{S}_A^r) \geq \gamma$ is satisfied. Also, observe that $\mathrm{STADD}(\mathcal{S}_A^r)$ reduces to $\mathrm{RIADD}(\mathcal{S}_A^r)$ when $r = 0$. It is also of note that STADD(T) is not the same as the limit $\lim_{k\to\infty} \mathrm{ADD}_k(T)$: in the quality control literature, the latter limit, as we indicated earlier, is known as the conditional steady-state ARL, while the STADD(T)-metric is known as the cyclical steady-state ARL. See, e.g., Pollak and Tartakovsky (2009), Shiryaev and Zryumov (2009), and Knoth (2006).

More importantly, it also turns out that the quantity $\mathrm{STADD}(\mathcal{S}_A^r)$ provides a universal lowerbound on the unknown value of $\min_{T \in \Delta(\gamma)} \mathrm{SADD}(T)$, and this lowerbound is valid for any $\gamma > 1$ and $r \geq 0$ such that $\mathcal{S}_A^r \in \Delta(\gamma)$. See Polunchenko and Tartakovsky (2010, Lemma 1 and Theorem 1). Specifically, introducing $\underline{\mathrm{SADD}}(\mathcal{S}_A^r) \equiv \mathrm{STADD}(\mathcal{S}_A^r)$, the following double inequality holds:

$$\underline{\mathrm{SADD}}(\mathcal{S}_A^r) \leq \min_{T \in \Delta(\gamma)} \mathrm{SADD}(T) \leq \mathrm{SADD}(\mathcal{S}_A^r), \tag{9}$$

for any $A > 0$ and $r \geq 0$ such that $\mathrm{ARL}(\mathcal{S}_A^r) \geq \gamma$, and any given $\gamma > 1$; cf. Moustakides et al. (2011, Inequality (2.12), p. 579).

A few important comments are now in order:

1. On the one hand, the double inequality (9), namely, its left part, implies that the lowerbound $\underline{\mathrm{SADD}}(\mathcal{S}_A^r) \equiv \mathrm{STADD}(\mathcal{S}_A^r)$, where STADD($T$) is defined in (8), can be used as a benchmark to get an idea as to how much room there is for improvement in the way of SADD for a chart of interest. Should it so happen that the SADD of the chart of interest with the ARL to false alarm level set to $\gamma > 1$ is only a tiny bit greater than $\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$ assuming $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma > 1$, then the chart is almost minimax optimal in the sense of Pollak (1985).
2. On the other hand, the double inequality (9) also suggests the following optimization strategy for the GSR chart: for a given $\gamma > 1$, pick the chart's detection threshold $A > 0$ and headstart $r \geq 0$ in such a way so as to make

the difference $\mathrm{SADD}(\mathcal{S}_A^r) - \underline{\mathrm{SADD}}(\mathcal{S}_A^r)$ as close to zero as is possible without violating the inequality $\mathrm{ARL}(\mathcal{S}_A^r) \geq \gamma$. More formally, the optimal detection threshold A^* and headstart r^* values are to be selected as follows:

$$(r^*, A^*) = \arg\min_{r,A\geq 0} \left\{ \mathrm{SADD}(\mathcal{S}_A^r) - \underline{\mathrm{SADD}}(\mathcal{S}_A^r) \right\}, \text{ but } \mathrm{ARL}(\mathcal{S}_A^r) = \gamma, \tag{10}$$

where $\gamma > 1$ is given; it goes without saying that both A^* and r^* are functions of $\gamma > 1$. The foregoing optimization strategy is originally due to Moustakides et al. (2011), and, in this work, we shall adapt it as well.

3. As we shall demonstrate in the next section, if the GSR chart's detection threshold A and initial score r are set to A^* and r^*, respectively, where A^* and r^* are as in (10) with $\gamma > 1$ given, then, conditional on $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma$, the difference $\mathrm{SADD}(\mathcal{S}_A^r) - \underline{\mathrm{SADD}}(\mathcal{S}_A^r)$ is nearly zero, even if $\gamma > 1$ is on the order of hundreds. Therefore, the GSR chart's third-order asymptotic optimality (5) does not necessarily require γ to be large.

While the constrained optimization problem (10) is generally infeasible to solve analytically, it can be solved *numerically* with any desired accuracy, e.g., with the aid of the numerical method proposed by Moustakides et al. (2011) and subsequently improved upon by Polunchenko et al. (2014a,b). This is precisely the object of the next section, and the results obtained in it are the main contribution of this work.

3 Experimental Results

We now examine the performance of the GSR chart given by (2) and (3) under different parameter settings, including (and especially) the optimal choice given by the solution of the constrained optimization problem (10). Specifically, the necessary performance characteristics of the GSR chart are computed *numerically* as solutions of certain integral (renewal) equations, which have previously been obtained by Moustakides et al. (2011) and by Polunchenko et al. (2014a,b). The need to treat the integral equations *numerically* is because an analytic solution is not an option. The specific numerical method used to solve the integral equations is a collocation-type method first proposed by Moustakides et al. (2011) and then also improved upon by Polunchenko et al. (2014a,b) who also provided tight error bounds for the method enabling one to judge the proximity of the numerical solution to the actual (infeasible to obtain) exact solution. As it may be relevant, we also note that we set up the numerical method so as to guarantee that its accuracy is on the order of a fraction of a percent.

We begin with an examination of the level of the ARL to false alarm, i.e., $\mathrm{ARL}(\mathcal{S}_A^r)$, treated as a function of the headstart $r \geq 0$, the detection threshold $A > 0$, and the magnitude of the change in the mean $\mu \neq 0$. With regard to the latter, for

lack of space, let us consider only two cases: $\mu = 0.2$ and $\mu = 0.5$. The former may be considered a faint change, while the latter is a moderate change. Figure 1 depicts $\mathrm{ARL}(\mathcal{S}_A^r)$ as a function of $r \in [0, A]$ and $A \in [0, 1000]$. Specifically, Fig. 1a is for $\mu = 0.2$ and Fig. 1b is for $\mu = 0.5$. As can be seen from either figure, the bivariate function $\mathrm{ARL}(\mathcal{S}_A^r)$ is almost linear in A (with r fixed) as well as in r (with A fixed). This is in perfect agreement with the aforementioned fact that $\mathrm{ARL}(\mathcal{S}_A^r) \approx A/\xi - r$ where ξ is given by (4). Since, according to Woodroofe (1982, Table 3.1, p. 33), the value of ξ for $\mu = 0.2$ is roughly 0.89004 versus approximately 0.74762 for $\mu = 0.5$, the sensitivity of the ARL to false alarm level to the detection threshold is higher, the stronger the change. Figure 1a, b also include contours (shown as bold dark curves) corresponding the different fixed levels $\gamma > 1$ of the ARL to false alarm. Specifically, each of the contours is the solution set (r, A) of the equation $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma$ for the appropriate value of $\gamma = \{100, 200, \ldots, 900, 1000\}$. These contours are important because the process of optimization of the GSR chart begins with picking a value for $\gamma > 1$, and then, with $\gamma > 1$ set and fixed, restricting attention to only those values of $A > 0$ and $r \geq 0$ for which the constraint $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma$ is satisfied. Due to space limitations, in this work we shall consider only three values of γ, namely, $\gamma = \{100, 500, 1000\}$.

Let us next look at Figs. 2 and 3 which show $\mathrm{ADD}_k(\mathcal{S}_A^r)$ as a function of $r \geq 0$ and $k = 0, 1, 2, \ldots$ under the constraint $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma$ with $\gamma = \{100, 500, 1000\}$. Specifically, Fig. 2 assumes $\mu = 0.2$ while Fig. 3 assumes $\mu = 0.5$. With regard to the level $\gamma > 1$ of the ARL to false alarm, Figs. 2a and 3a assume $\gamma = 100$, Figs. 2b and 3b are for $\gamma = 500$, and Figs. 2c and 3c assume $\gamma = 1000$. There are two important observations to make from either set of figures. First, it is evident that giving the SR chart a positive headstart equips the chart with the Fast Initial Response (FIR) feature: the chart becomes more sensitive to initial out-of-control situations. However, the flip side of the FIR feature is that the chart gets slower in situations when the process is initially in control but goes out of control later. It is worth reiterating that in order to retain the level of the ARL to false alarm assigning a higher value to the headstart is offset by an appropriate upward adjustment of the control limit. The second observation is that the maximal ADD, i.e., $\mathrm{SADD}(\mathcal{S}_A^r) \triangleq \max_{0 \leq k < \infty} \mathrm{ADD}_k(\mathcal{S}_A^r)$, is a sophisticated function of r, and the specific value of k at which the maximum is attained is hard to predict. As an aside, it is worth pointing out that the convergence of the ADD's to the steady-state regime is faster for $\mu = 0.5$ than for $\mu = 0.2$, which is consistent with one's intuition.

To better illustrate the FIR feature at work, let us look at Figs. 4 and 5, which are effectively the projections of the 3D surfaces shown in Figs. 2 and 3 onto the $(k, \mathrm{ADD}_k(\mathcal{S}_A^r))$-plane, made for a selection of values of r. Specifically, Fig. 4 assumes $\mu = 0.2$ and Fig. 5 is for $\mu = 0.5$. The corresponding levels $\gamma > 1$ of the ARL to false alarm are given in the figures' subtitles. The figures clearly demonstrate that, as the headstart increases, the performance of the GSR chart for initial of early out-of-control situation improves. However, the performance in situations when the process goes out of control later degrades. The interesting question is whether it is possible to optimize this tradeoff. This question is hard to answer properly without getting the lowerbound $\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$ involved, as is done in Figs. 6 and 7.

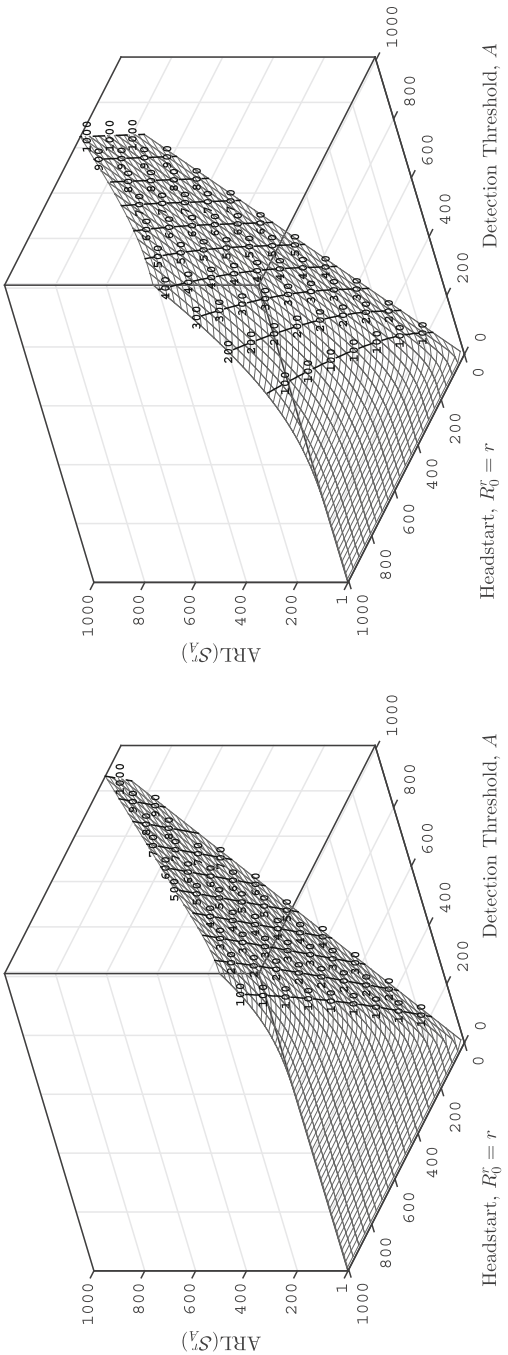


Fig. 1 ARL($\mathcal{S}_A^r$) as a function of the headstart $R_0^r = r \geq 0$ and the detection threshold $A > 0$ for $\mu = \{0.2, 0.5\}$. (**a**) $\mu = 0.2$. (**b**) $\mu = 0.5$

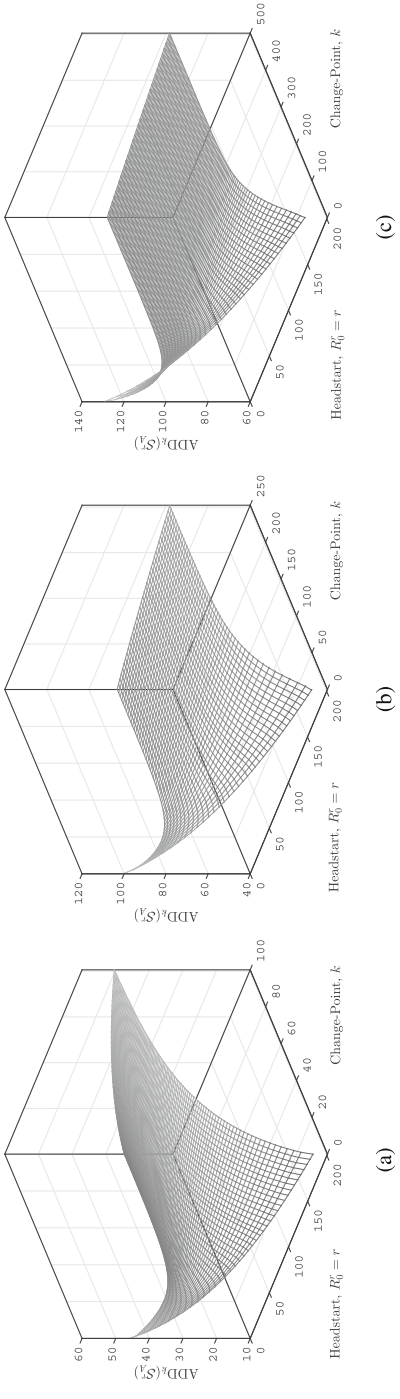


Fig. 2 $\mathrm{ADD}_k(\mathcal{S}_A^r)$ as a function of the headstart $R_0^r = r \geq 0$, the change-point $k = 0, 1, \ldots$, and the ARL to false alarm level $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma > 1$ for $\mu = 0.2$. (**a**) $\gamma = 100$. (**b**) $\gamma = 500$. (**c**) $\gamma = 1000$

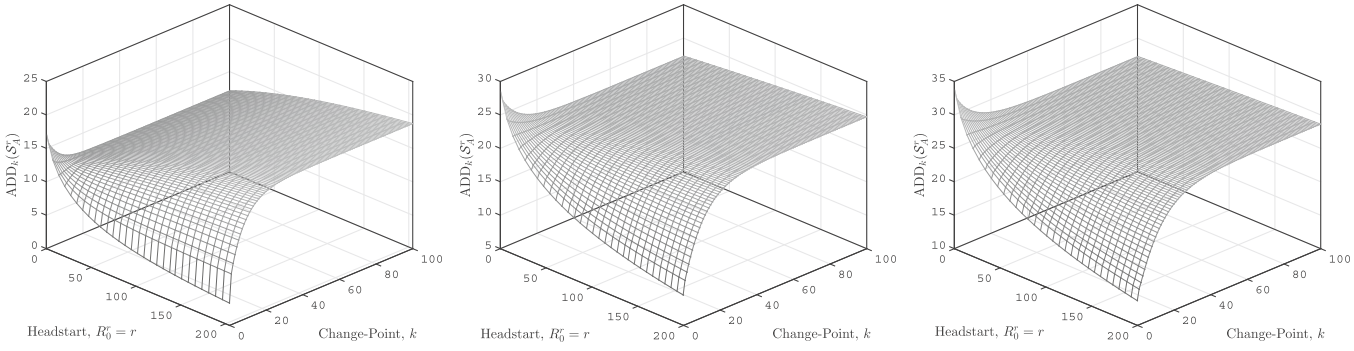


Fig. 3 $\mathrm{ADD}_k(\mathcal{S}_A^r)$ as a function of the headstart $R_0^r = r \geq 0$, the change-point $k = 0, 1, \ldots$, and the ARL to false alarm level $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma > 1$ for $\mu = 0.5$. (**a**) $\gamma = 100$. (**b**) $\gamma = 500$. (**c**) $\gamma = 1000$

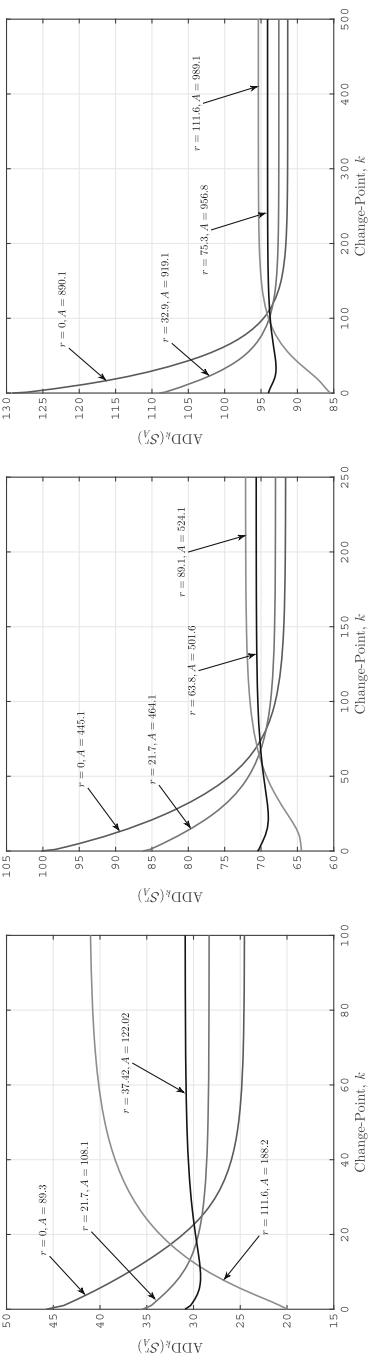


Fig. 4 $\mathrm{ADD}_k(\mathcal{S}_A^r)$ as a function of the headstart $R_0^r = r \geq 0$, the change-point $k = 0, 1, \ldots$, and the ARL to false alarm level $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma > 1$ for $\mu = 0.2$. **(a)** $\gamma = 100$. **(b)** $\gamma = 500$. **(c)** $\gamma = 1000$

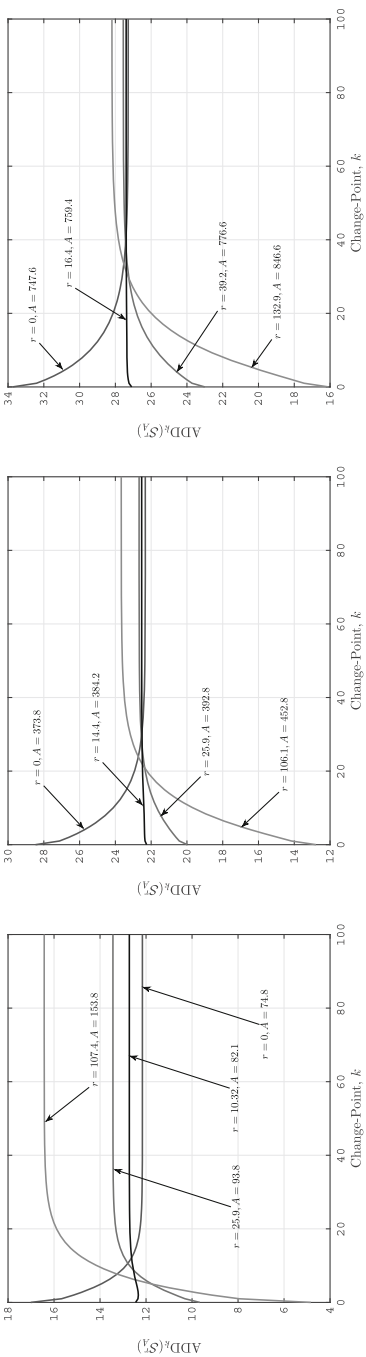


Fig. 5 $\mathrm{ADD}_k(\mathcal{S}_A^r)$ as a function of the headstart $R_0^r = r \geq 0$, the change-point $k = 0, 1, \ldots$, and the ARL to false alarm level $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma > 1$ for $\mu = 0.5$. (**a**) $\gamma = 100$. (**b**) $\gamma = 500$. (**c**) $\gamma = 1000$

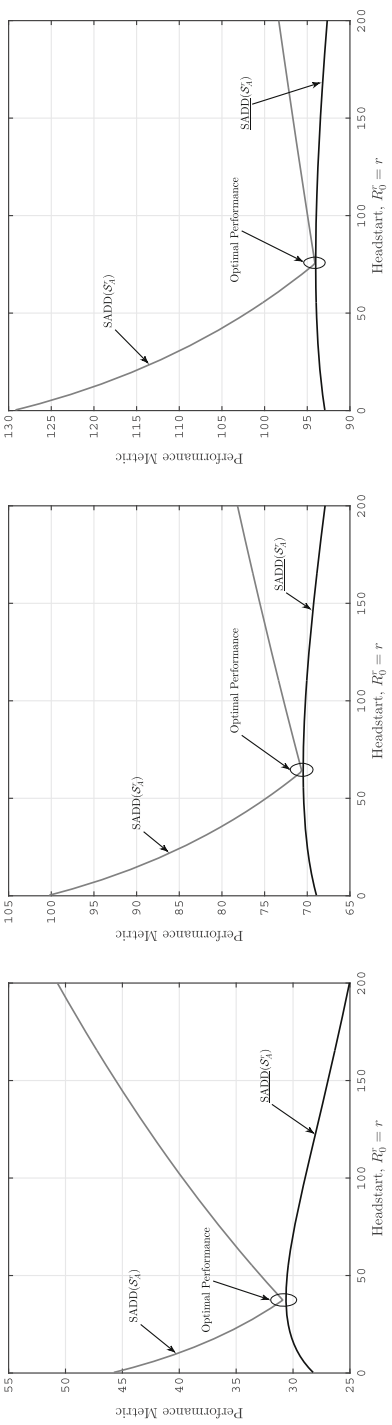


Fig. 6 SADD($\mathcal{S}_A^r$) and $\underline{\text{SADD}}(\mathcal{S}_A^r)$ as functions of the headstart $R_0^r = r$ and the ARL to false alarm level $\text{ARL}(\mathcal{S}_A^r) = \gamma > 1$ for $\mu = 0.2$. (**a**) $\gamma = 100$. (**b**) $\gamma = 500$. (**c**) $\gamma = 1000$

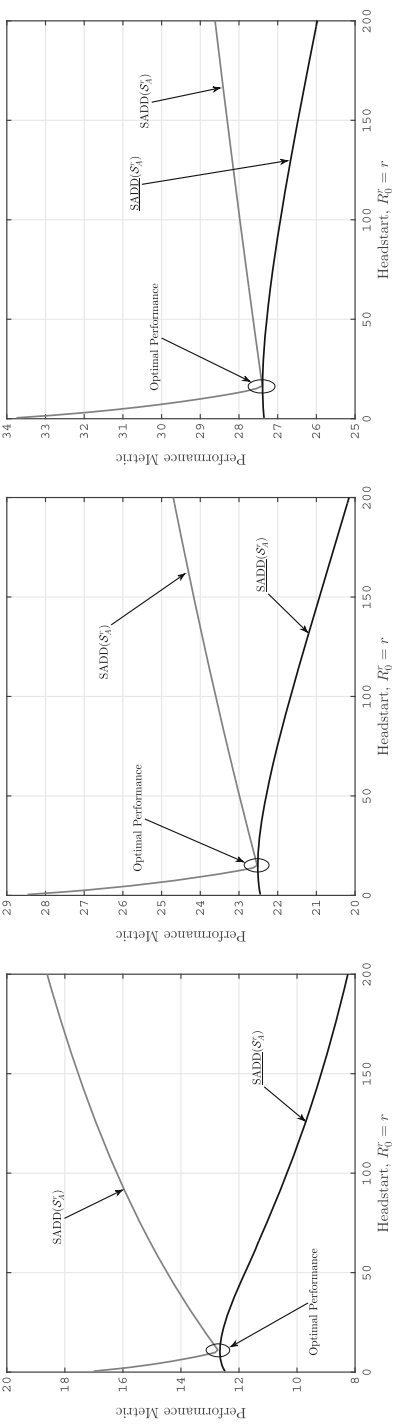


Fig. 7 SADD($\mathcal{S}_A^r$) and $\underline{\text{SADD}}(\mathcal{S}_A^r)$ as functions of the headstart $R_0^r = r$ and the ARL to false alarm level $\text{ARL}(\mathcal{S}_A^r) = \gamma > 1$ for $\mu = 0.5$. (**a**) $\gamma = 100$. (**b**) $\gamma = 500$. (**c**) $\gamma = 1000$

Specifically, Figs. 6 and 7 provide an idea as to the manner in which $\mathrm{SADD}(\mathcal{S}_A^r)$ and $\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$ each depend on the headstart, assuming, as before, that every change in the headstart is accompanied by the appropriate adjustment of the detection threshold, so that the ARL to false alarm constraint is kept intact. More specifically, Fig. 6 corresponds to $\mu = 0.2$ and Fig. 7 are for $\mu = 0.5$. The respective levels γ of the ARL to false alarm are again given in the subtitles.

It is evident from the figures that, regardless of the contrastness of the shift in the mean $\mu \neq 0$ and no matter the ARL to false alarm level $\gamma > 1$, the lowerbound is an upward arching smooth function of the initial score, and it has a distinct maximum. The figures also clearly indicate that the maximal ADD as a function of r has a minimum with the appearance of a down pointing cusp; the cusp is an indication that the way the maximal element of the sequence $\{\mathrm{ADD}_k(\mathcal{S}_A^r)\}_{k\geq 0}$ and its location within the sequence depend on the headstart is highly nonlinear. The essential observation is that the lowerbound appears to peak at approximately the same (slightly smaller actually) headstart value as that at which the maximal ADD is minimized. Moreover, although the maximal ADD's minimum is higher than the lowerbound's maximum, the difference is not practically significant, even if γ is as low as 100, and is smaller, the higher the value of γ. Therefore, any other chart with the same level of the ARL to false alarm cannot possibly detect the shift in the mean with a detection delay substantially lower than that delivered by the optimized GSR chart, especially if the shift in the mean is contrast.

To draw a line under this section, in Tables 1 and 2, we give the optimal headstart and detection threshold values that have been computed by solving the constrained optimization problem (10) for $\gamma = \{100, 200, \ldots, 900, 1000\}$ and $\mu = \{0.1, 0.2, \ldots, 0.9, 1.0\}$. Recall also that our data model is symmetric with respect to the sign of $\mu \neq 0$. The tables also include the corresponding $\mathrm{SADD}(\mathcal{S}_A^r)$ and $\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$ values. One can see from the tables that $\mathrm{SADD}(\mathcal{S}_A^r) \approx \underline{\mathrm{SADD}}(\mathcal{S}_A^r)$, which is to say that the detection capabilities of the optimized GSR chart are almost the best. One can also see that the effect of headstarting is the stronger, the fainter the anticipated shift in the mean. If the latter is fairly contrast, the optimal headstart value is close to zero. In addition, the tables also suggest that the optimal headstart value, as a function of the ARL to false alarm level $\gamma > 1$, has a *finite* limit as $\gamma \to +\infty$; the convergence to the limiting value is the slower, the weaker the change. However, a closed-form formula for this limiting value is prohibitively difficult to obtain.

4 Concluding Remarks

In summary we see that

1. Starting an SR chart off a nonzero initial score lessens the ARL to false alarm, so that the chart's in-control performance is worse than when no headstart is used. On the flip side, however, the chart becomes more sensitive to initial out-of-control situations. This is precisely the FIR phenomenon.

Table 1 Optimal headstart, $r^* > 0$, control limit, $A^* > 0$, maximal ADD, $\mathrm{SADD}(\mathcal{S}_A^r)$, and the lowerbound, $\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$, as functions of the shift magnitude, $\mu > 0$, and the ARL to false alarm level, $\mathrm{ARL}(\mathcal{S}_A^r) = \gamma > 1$, for $\gamma = \{100, 200, 300, 400, 500\}$

$\mathrm{ARL}(\mathcal{S}_A^r) = \gamma$	Performance characteristic	Change magnitude ($\mu > 0$)									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
100	r^*	83.93	37.42	21.96	14.53	10.32	7.66	5.89	4.64	3.74	3.05
	A^*	173.25	122.02	102.11	90.43	82.14	75.6	70.14	65.38	61.14	57.31
	$\mathrm{SADD}(\mathcal{S}_A^r)$	49.65	30.9	21.6	16.17	12.68	10.28	8.55	7.25	6.25	5.46
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	48.76	30.62	21.49	16.13	12.66	10.27	8.54	7.25	6.25	5.46
200	r^*	114.43	48.29	27.23	17.49	12.14	8.86	6.72	5.27	4.24	3.48
	A^*	296.37	220.71	190.51	172	158.26	147	137.28	128.63	120.79	113.58
	$\mathrm{SADD}(\mathcal{S}_A^r)$	79.79	45.39	30.1	21.75	16.61	13.19	10.79	9.03	7.69	6.65
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	78.7	45.14	30.03	21.72	16.6	13.19	10.79	9.03	7.69	6.65
300	r^*	135.53	55.1	30.29	19.12	13.1	9.54	7.27	5.7	4.6	3.77
	A^*	410.61	315.77	277.06	252.52	233.74	218.04	204.24	191.76	180.34	169.78
	$\mathrm{SADD}(\mathcal{S}_A^r)$	103.23	55.71	35.87	25.41	19.13	15.04	12.19	10.13	8.58	7.38
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	102.08	55.5	35.81	25.4	19.13	15.03	12.19	10.13	8.58	7.38
400	r^*	151.87	60.02	32.41	20.2	13.81	10.05	7.65	6.01	4.83	3.98
	A^*	520.37	409.15	362.81	332.61	309.04	288.95	271.07	254.81	239.82	225.94
	$\mathrm{SADD}(\mathcal{S}_A^r)$	122.8	63.86	40.29	28.17	21.02	16.4	13.22	10.93	9.22	7.91
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	121.65	63.68	40.25	28.16	21.01	16.39	13.22	10.93	9.22	7.91
500	r^*	165.27	63.84	33.98	21.03	14.36	10.45	7.95	6.25	5.03	4.14
	A^*	627.35	501.56	448.1	412.5	384.21	359.78	337.86	317.81	299.29	282.07
	$\mathrm{SADD}(\mathcal{S}_A^r)$	139.75	70.63	43.86	30.39	22.52	17.48	14.05	11.57	9.73	8.33
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	138.63	70.48	43.86	30.39	22.52	17.48	14.03	11.57	9.73	8.33

Table 2 Optimal headstart, $r^* > 0$, control limit, $A^* > 0$, maximal ADD, $\text{SADD}(\mathcal{S}_A^r)$, and the lowerbound, $\underline{\text{SADD}}(\mathcal{S}_A^r)$, as functions of the shift magnitude, $\mu > 0$, and the ARL to false alarm level, $\text{ARL}(\mathcal{S}_A^r) = \gamma > 1$, for $\gamma = \{600, 700, 800, 900, 1000\}$

$\text{ARL}(\mathcal{S}_A^r) = \gamma$	Performance characteristic	Change magnitude ($\mu > 0$)									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
600	r^*	176.63	66.94	35.24	21.73	14.81	10.78	8.19	6.45	5.2	4.28
	A^*	732.41	593.32	533.12	492.28	459.31	430.56	404.61	380.8	358.73	338.18
	$\text{SADD}(\mathcal{S}_A^r)$	153.8	76.46	46.95	32.26	23.77	18.38	14.71	12.09	10.15	8.67
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	153.71	76.33	46.92	32.25	23.77	18.37	14.71	12.09	10.15	8.67
700	r^*	186.49	69.53	36.25	22.32	15.21	11.06	8.42	6.61	5.34	4.39
	A^*	836.06	684.63	617.94	571.98	534.37	501.32	471.35	443.76	418.15	394.28
	$\text{SADD}(\mathcal{S}_A^r)$	168.37	81.58	49.59	33.87	24.85	19.15	15.29	12.54	10.51	8.96
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	167.32	81.47	49.57	33.86	24.85	19.14	15.28	12.54	10.51	8.96
800	r^*	195.2	71.73	37.14	22.84	15.55	11.3	8.6	6.77	5.47	4.5
	A^*	938.62	775.59	702.67	651.62	609.38	572.04	538.06	506.71	477.58	450.38
	$\text{SADD}(\mathcal{S}_A^r)$	180.76	86.16	51.94	35.29	25.79	19.82	15.79	12.93	10.82	9.22
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	179.75	86.06	51.92	35.28	25.79	19.82	15.79	12.93	10.82	9.22
900	r^*	202.99	73.65	37.93	23.31	15.86	11.53	8.78	6.91	5.57	4.59
	A^*	1040.31	866.3	787.3	731.22	684.37	642.75	604.77	569.66	536.98	506.46
	$\text{SADD}(\mathcal{S}_A^r)$	192.18	90.3	54.04	36.55	26.64	20.42	16.24	13.28	11.1	9.44
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	191.21	90.21	54.03	36.55	26.63	20.42	16.24	13.28	11.1	9.44
1000	r^*	210.04	75.34	38.62	23.71	16.14	11.73	8.93	7.04	5.68	4.66
	A^*	1141.3	956.81	871.86	810.77	759.35	713.44	671.46	632.6	596.38	562.54
	$\text{SADD}(\mathcal{S}_A^r)$	202.79	94.09	55.96	37.7	27.39	20.96	16.64	13.59	11.35	9.65
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	201.86	94.01	55.94	37.69	27.39	20.95	16.64	13.59	11.35	9.64

2. The drop in the ARL to false alarm caused by a positive headstart value can be compensated by an increase of the control limit. While this would negatively affect the chart's out-of-control performance, the magnitude of the effect appears to be not substantial.
3. The FIR feature comes at the price of poorer performance in situations when the process under surveillance is initially in control but goes out of control later. In particular, if the process is not expected to shift out of control for a long while, then no headstarting is necessary, because the SR chart's steady-state performance would degrade otherwise.

The same observations were previously made by Lucas and Crosier (1982) about the CUSUM chart.

Our additional and more important contribution consists in a deeper investigation of the headstart-vs-control-limit tradeoff: the overall performance of the GSR chart optimized not only with respect to the headstart but also with respect to the control limit proved to be nearly the best one can get amid complete uncertainty as to when the observed process may go out of control. This is a direct implication of the GSR chart's strong optimality properties established by Pollak and Tartakovsky (2009), Shiryaev and Zryumov (2009), Tartakovsky and Polunchenko (2010), Polunchenko and Tartakovsky (2010), and by Tartakovsky et al. (2012). The optimal headstart and control limit values, and the corresponding out-of-control performance and its lowerbound, for a variety of cases, are given in Tables 1 and 2.

The benefits of optimizing the GSR chart are the greater, the fainter the change. From a practical standpoint, this means that if one is interested in detecting a faint change, then the GSR chart with optimally selected control limit and headstart is the way to go. The size of the actual efficiency improvement can be estimated using Tables 1 and 2. However, if the anticipated change to be detected is more or less contrast, then the GSR chart, whether optimized or not, will not offer any substantial advantage (in terms of the speed of detection) over the CUSUM scheme or the EWMA chart.

Acknowledgements The author is thankful to the anonymous referee whose constructive feedback helped improve the manuscript.

The author's effort was partially supported by the Simons Foundation via a Collaboration Grant in Mathematics under Award # 304574.

References

Duncan, A. J. (1956). The economic design of $\bar{X}$ charts used to maintain current control of a process. *Journal of the American Statistical Association, 51*(274), 228–242. https://doi.org/10.2307/2281343

Girshick, M. A., & Rubin, H. (1952). A Bayes approach to a quality control model. *Annals of Mathematical Statistics, 23*(1), 114–125. https://doi.org/10.1214/aoms/1177729489

Kenett, R., & Pollak, M. (1986). A semi-parametric approach to testing for reliability growth, with application to software systems. *IEEE Transactions on Reliability, 35*(3), 304–311. https://doi.org/10.1109/TR.1986.4335439

Kenett, R., & Pollak, M. (1996). Data-analytic aspects of the Shiryayev-Roberts control chart: Surveillance of a non-homogeneous Poisson process. *Journal of Applied Statistics, 23*(1), 125–138. https://doi.org/10.1080/02664769624413

Knoth, S. (2006). The art of evaluating monitoring schemes – How to measure the performance of control charts? In: H.-J. Lenz, & P.-Th. Wilrich (Eds.), *Frontiers in Statistical Quality Control* (Vol. 8, pp. 74–99). Heidelberg: Physica. https://doi.org/10.1007/3-7908-1687-6_5

Lucas, J. M., & Crosier, R. B. (1982). Fast initial response for CUSUM quality-control schemes: Give your CUSUM a head start. *Technometrics, 24*(3), 199–205. https://doi.org/10.2307/1268679

Mevorach, Y., & Pollak, M. (1991). A small sample size comparison of the CUSUM and Shiryayev-Roberts approaches to changepoint detection. *American Journal of Mathematical and Management Sciences, 11*(3–4), 277–298. https://doi.org/10.1080/01966324.1991.10737312

Moustakides, G. V., Polunchenko, A. S., & Tartakovsky, A. G. (2011). A numerical approach to performance analysis of quickest change-point detection procedures. *Statistica Sinica, 21*(2), 571–596.

Moustakides, G. V., Polunchenko, A. S., & Tartakovsky, A. G. (2009). Numerical comparison of CUSUM and Shiryaev-Roberts procedures for detecting changes in distributions. *Communications in Statistics: Theory and Methods, 38*(16), 3225–3239. https://doi.org/10.1080/03610920902947774

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*(1–2), 100–115. https://doi.org/10.2307/2333009

Pepelyshev, A., & Polunchenko, A. S. (2017). Real-time financial surveillance via quickest change-point detection methods. *Statistics and Its Interface, 10*(1), 93–106. https://doi.org/10.4310/SII.2017.v10.n1.a9

Pollak, M. (1985). Optimal detection of a change in distribution. *Annals of Statistics, 13*(1), 206–222. https://doi.org/10.1214/aos/1176346587

Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *Annals of Statistics, 15*(2), 749–779. https://doi.org/10.1214/aos/1176350373

Pollak, M. (2009). The Shiryaev-Roberts changepoint detection procedure in retrospect – Theory and practice. In: *Proceedings of the 2nd International Workshop on Sequential Methodology*, University of Technology of Troyes, Troyes, France, 15–17 Jun, 2009.

Pollak, M., & Tartakovsky, A. G. (2009). Optimality properties of the Shiryaev-Roberts procedure. *Statistica Sinica, 19*, 1729–1739.

Polunchenko, A. S., & Sokolov, G. (2014). Toward optimal design of the Generalized Shiryaev-Roberts procedure for quickest change-point detection under exponential observations. In: *Proceedings of the 2014 International Conference on Engineering and Telecommunications*, Moscow Institute of Physics and Technology, Moscow, Russia, 26–28 Nov 2014; pp. 51–55. https://doi.org/10.1109/EnT.2014.37

Polunchenko, A. S., Sokolov, G., & Du, W. (2014). Efficient performance evaluation of the generalized Shiryaev-Roberts detection procedure in a multi-cyclic setup. *Applied Stochastic Models in Business and Industry, 30*(6), 723–739. https://doi.org/10.1002/asmb.2026

Polunchenko, A. S., Sokolov, G., & Du, W. (2014). An accurate method for determining the pre-change Run-Length distribution of the Generalized Shiryaev-Roberts detection procedure. *Sequential Analysis, 33*(1), 112–134. https://doi.org/10.1080/07474946.2014.856642

Polunchenko, A. S., Sokolov, G., & Du, W. (2017). On robustness of the Shiryaev-Roberts change-point detection procedure under parameter misspecification in the post-change distribution. *Communications in Statistics: Simulation and Computation, 46*(3), 2185–2206. https://doi.org/10.1080/03610918.2015.1039131

Polunchenko, A. S., & Tartakovsky, A. G. (2010). On optimality of the Shiryaev-Roberts procedure for detecting a change in distribution. *Annals of Statistics, 38*(6), 3445–3457. https://doi.org/10.1214/09-AOS775

Polunchenko, A. S., & Tartakovsky, A. G. (2012). State-of-the-art in sequential change-point detection. *Methodology and Computing in Applied Probability, 44*(3), 649–684. https://doi.org/10.1007/s11009-011-9256-5

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics, 1*(3), 239–250. https://doi.org/10.2307/1271439

Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics, 8*(3), 411–430. https://doi.org/10.2307/1266688

Shiryaev, A. N. (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Mathematics - Doklady, 2*, 795–799.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and Its Applications, 8*(1), 22–46. https://doi.org/10.1137/1108002

Shiryaev, A. N., & Zryumov, P. Y. (2009). On the linear and nonlinear generalized Bayesian disorder problem (discrete time case). In: F. Delbaen, M. Rásonyi, & Ch. Stricker (Eds.), *Optimality and Risk – Modern Trends in Mathematical Finance. The Kabanov Festschrift* (pp. 227–235). Berlin: Springer. https://doi.org/10.1007/978-3-642-02608-9_12

Tartakovsky, A. G., & Ivanova, I. V. (1992). Comparison of some sequential rules for detecting changes in distributions. *Problems of Information Transmission, 28*(2), 117–124.

Tartakovsky, A. G., Pollak, M., & Polunchenko, A. S. (2012). Third-order asymptotic optimality of the Generalized Shiryaev-Roberts changepoint detection procedure. *Theory Probability and Its Applications, 56*(3), 457–484. https://doi.org/10.1137/S0040585X97985534

Tartakovsky, A. G., Polunchenko, A. S., & Sokolov, G. (2013). Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal on Selected Topics in Signal Processing, 7*(1), 4–11. https://doi.org/10.1109/JSTSP.2012.2233713

Tartakovsky, A. G., & Polunchenko, A. S., (2010). Minimax optimality the Shiryaev-Roberts procedure. In: *Proceedings of the 5th International Workshop on Applied Probability*, Universidad Carlos III de Madrid, Colmenarejo Campus, Spain, 5–8 Jul, 2010.

Tartakovsky, A. G., Polunchenko, A. S., & Moustakides, G. V. (2009). Design and comparison of Shiryaev-Roberts- and CUSUM-type change-point detection procedures. In: *Proceedings of the 2nd International Workshop on Sequential Methodology*, University of Technology of Troyes, Troyes, France, 15–17 Jun, 2009.

Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*. Philadelphia: SIAM.

On ARL-Unbiased Charts to Monitor the Traffic Intensity of a Single Server Queue

Manuel Cabral Morais and Sven Knoth

Abstract We know too well that the effective operation of a queueing system requires maintaining the traffic intensity ρ at a target value ρ_0.

This important measure of congestion can be monitored by using control charts, such as the one found in the seminal work by Bhat and Rao (Oper Res 20:955–966, 1972) or more recently in Chen and Zhou (Technometrics 57:245–256, 2015).

For all intents and purposes, this chapter focus on three control statistics chosen by Morais and Pacheco (Seq Anal 35:536–559, 2016) for their simplicity, recursive and Markovian character. Since an upward and a downward shift in ρ are associated with a deterioration and an improvement (respectively) of the quality of service, the timely detection of these changes is an imperative requirement, hence, begging for the use of ARL-unbiased charts (Pignatiello et al., The performance of control charts for monitoring process dispersion. In: 4th industrial engineering research conference, pp 320–328, 1995), in the sense that they detect any shifts in the traffic intensity sooner than they trigger a false alarm.

In this chapter, we focus on the design of these type of charts for the traffic intensity of the three single server queues mentioned above.

Keywords Statistical process control · Dependent control statistics · ARL-unbiased charts

M. C. Morais (✉)
CEMAT & Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
e-mail: maj@math.ist.utl.pt

S. Knoth
Institute of Mathematics and Statistics, Department of Economics and Social Sciences, Helmut Schmidt University Hamburg, Hamburg, Germany
e-mail: knoth@hsu-hh.de

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_5

1 Introduction

The first contributions on queueing theory (QT) can be traced back to three pioneering chapters by A.K. Erlang (1878–1929). Erlang (1909, 1917, 1920) were in any case a response to concrete congestion problems arising in teletraffic.

Curiously, we have to leap to the late 1950s and 1960s for the earliest chapters referring to the statistical inference in QT: Clarke (1957) (resp. Beneš 1957) focused on the MLE for λ, μ and the traffic intensity of a $M/M/1$ (resp. $M/M/\infty$) system, $\rho = \lambda/\mu$; Cox (1965) and Lilliefors (1966) derived confidence intervals for the traffic intensity of a $M/M/1$ system.

In the following decade, the seminal work by Bhat and Rao was published and addressed the monitoring of the traffic intensity. Bhat and Rao (1972) proposed what we consider an *unusual* chart for the traffic intensity of the $M/G/1$ (resp. $GI/M/1$) queueing systems because:

- its rule to trigger a signal does not coincide with any of the ten sensitizing rules for Shewhart control charts in Montgomery (2009, p. 197, Table 5.1), such as the Western Electric run rules (Western Electric 1956); the traffic intensity is deemed out-of-control only if the control statistic exceeds (resp. does not exceed) the upper (resp. lower) control limit c_u (resp. c_l) longer than a pre-assigned number d_u (resp. d_l) of consecutive transitions;
- the run length (RL) is not considered as a performance measure and the control limits are not defined so as to achieve, for instance, a specific in-control average run length (ARL);
- the control limit c_u (resp. c_l) is the smallest (resp. largest) nonnegative integer for which the probability of having an observation above (resp. not above) c_u (resp. c_l) is at most α_u (resp. α_l), and the positive integer d_u (resp. d_l) is such that, when the control statistic has gone above (resp. not above) c_u (resp. c_l), it returns to a state $\leq c_u$ (resp. $> c_l$) in d_u (resp. d_l) or fewer steps with probability of at least $1 - \beta_u$ (resp. $1 - \beta_l$);
- the chart assumes that the system is observed under equilibrium or steady state conditions.

The thorough review on regulation techniques for the traffic intensity in Morais and Pacheco (2015, On control charts and the detection of increases in the traffic intensity, p. 44, unpublished manuscript) led Morais and Pacheco (2016) to add that the monitoring ρ can be basically divided in categories depending on:

- the control statistic being used, e.g.
  - the number of customers in the system at departure/arrival epochs (Bhat and Rao 1972; Rao et al. 1984; Shore 2000; Chen et al. 2011; Zobu and Sağlam 2013),
  - the number of arrivals while the n^{th} customer is being served, etc. (Jain and Templeton 1989);

- the statistical technique used to detect changes in the traffic intensity
  - a control chart (Bhat and Rao 1972; Shore 2000, 2006; Kim et al. 2007; Chen et al. 2011; Hung et al. 2012; Chen and Zhou 2015),
  - a sequential probability ratio test (Rao et al. 1984; Bhat 1987; Jain and Templeton 1989; Jain 2000; Zobu and Sağlam 2013).

1.1 Three Control Statistics: X_n, $\hat{X}_n$ and W_n

To monitor the traffic intensity of a single server queue and keep it at a target level ρ_0, Morais and Pacheco (2015, On control charts and the detection of increases in the traffic intensity, p. 44, unpublished manuscript; 2016) used the three following control statistics:

- X_n, the number of customers left behind in the $M/G/1$ system by the n^{th} departing customer;
- $\hat{X}_n$, the number of customers seen in the $GI/M/1$ system by the n^{th} arriving customer;
- W_n, the waiting time of the n^{th} arriving customer to the $GI/G/1$ system.

These three control statistics have been chosen by Morais and Pacheco (2016) for their simplicity, recursive and Markovian character. Their recursive is apparent if we note that these statistics can be rewritten as follows: where

System	Control statistic
$M/G/1$	$X_{n+1} = \max\{0,\ X_n - 1\} + Y_{n+1}$
$GI/M/1$	$\hat{X}_{n+1} = \max\{0, \hat{X}_n + 1 - \hat{Y}_{n+1}\}$
$GI/G/1$	$W_{n+1} = \max\{0,\ W_n + S_{n+1} - A_{n+1}\}$

- Y_{n+1} denotes the number of customers arriving during the service of the $(n+1)^{th}$ customer,
- $\hat{Y}_{n+1}$ represents the number of customers served between the arrivals of customers n and $(n+1)$,
- $S_{n+1} - A_{n+1}$ depends on the service time of the n^{th} customer, S_{n+1}, and on the time between the arrivals of customers n and $(n+1)$, A_{n+1},

for $n \in \mathbb{N}_0$.

1.2 X_n and the M/G/1 System

The reader should be reminded of some important facts: customers arrive to the $M/G/1$ queueing system according to a Poisson process with rate λ and are served

one at a time by the single server; the service times are independent and identically distributed (i.i.d.) positive random variables (r.v.), which are in turn independent of the interarrival times; S, $F_S(s)$ and $E(S) = \mu^{-1}$ stand from now on for the service time, its cumulative distribution function (c.d.f.) and expected value.

Kendall (1951, 1953) noted that $\{X_n,\ n \in \mathbb{N}\}$ forms a discrete time Markov chain (DTMC), termed the $M/G/1$ *embedded Markov chain*, with transition probability matrix (TPM)

$$\mathbf{P} = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \ddots \\ 0 & 0 & \alpha_0 & \alpha_1 & \ddots \\ 0 & 0 & 0 & \alpha_0 & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{1}$$

where α_i denotes the probability that exactly i customers arrive during a service time S. In addition,

$$\alpha_i = \int_0^{+\infty} e^{-\lambda s} \frac{(\lambda s)^i}{i!}\, dF_S(s), \quad i \in \mathbb{N}_0 \tag{2}$$

(Adan and Resing 2015, p. 63). Another revealing fact: $Y_n \overset{i.i.d.}{\sim} Y$, $n \in \mathbb{N}$, with common probability function (p.f.) given by $P_Y(i) = \alpha_i$, $i \in \mathbb{N}_0$.

1.3 $\hat{X}_n$ and the GI/M/1 System

The $GI/M/1$ queueing system is characterized by: interarrival times that are i.i.d. positive r.v. with common c.d.f. $F_A(a)$ and expected value $E(A) = \lambda^{-1}$; i.i.d. exponentially distributed service times, with expected value μ^{-1} and independent of the interarrival times.

Kendall (1951) established that $\{\hat{X}_n,\ n \in \mathbb{N}\}$ also forms a DTMC, the $GI/M/1$ *embedded Markov chain*, whose TPM is equal to

$$\hat{\mathbf{P}} = \begin{bmatrix} \hat{p}_{00} & \hat{\alpha}_0 & 0 & 0 & 0 & \cdots \\ \hat{p}_{10} & \hat{\alpha}_1 & \hat{\alpha}_0 & 0 & 0 & \cdots \\ \hat{p}_{20} & \hat{\alpha}_2 & \hat{\alpha}_1 & \hat{\alpha}_0 & 0 & \cdots \\ \hat{p}_{30} & \hat{\alpha}_3 & \hat{\alpha}_2 & \hat{\alpha}_1 & \hat{\alpha}_0 & \ddots \\ \vdots & \cdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{3}$$

where $\hat{\alpha}_i$ denotes the probability of serving i customers during an interarrival time U given that the server remains busy during this interval. Please note that

$$\hat{\alpha}_i = \int_0^{+\infty} e^{-\mu a} \frac{(\mu a)^i}{i!} dF_A(a), \quad i \in \mathbb{N}_0, \tag{4}$$

and $\hat{p}_{i0} = 1 - \sum_{j=0}^{i} \hat{\alpha}_j,\ i \in \mathbb{N}_0$ (Adan and Resing 2015, p. 82). Expectedly, $\hat{Y}_n \overset{i.i.d.}{\sim} \hat{Y},\ n \in \mathbb{N}$, with common p.f. $P_{\hat{Y}}(i) = \hat{\alpha}_i,\ i \in \mathbb{N}_0$.

1.4 W_n and the GI/G/1 System

This single-server queueing system is associated with: interarrival (resp. service) times that are i.i.d. positive r.v. with common c.d.f. $F_A(a)$ (resp. $F_S(s)$) and mean $E(A) = \lambda^{-1}$ (resp. $E(S) = \mu^{-1}$); service times are once more independent of the interarrival times.

$\{W_n,\ n \in \mathbb{N}_0\}$ also forms a DTMC (Kendall 1953) and $S_n - A_n \overset{i.i.d.}{\sim} S - A,\ n \in \mathbb{N}$.

Bear in mind that this DTMC has a continuous state space $\mathbb{R}_0^+$ if the interarrival or the service times are absolutely continuous r.v.

Following Morais and Pacheco (1998), we consider a discretized approximating DTMC with:

- state space $\mathbb{N}_0$;
- its first state corresponding to the singleton $\{0\}$;
- its state j associated with interval $((j-1)\Delta,\ j\Delta]$, for $j \in \mathbb{N}$, where Δ denotes the common range of all the intervals and is taken to be very small so as to improve the approximation;
- the interval $((j-1)\Delta,\ j\Delta]$ represented by point $(j - 1/2)\ \Delta$, for $j \in \mathbb{N}$.

The TPM of this approximating DTMC is given by

$$\tilde{\mathbf{P}} = \begin{bmatrix} F(0) & F(\Delta) - F(0) & F(2\Delta) - F(\Delta) & \cdots \\ F\left(-\frac{\Delta}{2}\right) & F\left(\frac{\Delta}{2}\right) - F\left(-\frac{\Delta}{2}\right) & F\left(\frac{3\Delta}{2}\right) - F\left(\frac{\Delta}{2}\right) & \cdots \\ F\left(-\frac{3\Delta}{2}\right) & F\left(-\frac{\Delta}{2}\right) - F\left(-\frac{3\Delta}{2}\right) & F\left(\frac{\Delta}{2}\right) - F\left(-\frac{\Delta}{2}\right) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{5}$$

where the c.d.f. $F \equiv F_{S-A}$. Note that its first row differs slightly from the one following Brook and Evans (1972) and used by Greenberg (1997) and Morais and Pacheco (2016), who considered that state j is associated with interval $((j - 1/2)\Delta,\ (j + 1/2)\Delta]$, for $j \in \mathbb{N}_0$, and that the interval $((j - 1/2)\Delta,\ (j + 1/2)\Delta]$ is represented by point $j\ \Delta$, for $j \in \mathbb{N}_0$.

We feel bound to point out that α_i, $\hat{\alpha}_i$, and $F_{S-A}(t)$ have fairly simple and closed expressions for some *typical* queueing systems with interarrival or service times with an Erlang distribution with k ($k \in \mathbb{N}$) phases, as shown in Appendix. This certainly proves to be convenient if we want to describe in detail the run length performance of the associated control charts.

A quick look at the expressions of α_i, $\hat{\alpha}_i$, and $F_{S-A}(t)$ in Appendix leads us to conclude that $F_{S-A}(x)$ depends upon both λ and μ, unlike α_i and $\hat{\alpha}_i$. Consequently, the entries of $\tilde{\mathbf{P}}$ will not depend exclusively on ρ like $\mathbf{P}$ and $\hat{\mathbf{P}}$.

1.5 On the Probability of Null Values of the Control Statistics

A closer look at the control statistics of the X_n-, $\hat{X}_n$- and W_n-charts suggests that they take null values quite frequently, as long as single server queueing systems are able to reach equilibrium, that is, if the traffic intensity is less than one.

Firstly, when it comes to the monitoring the traffic intensity of a $GI/G/1$ queueing system, we can certainly state that the "most frequent" value of W_n is zero because this statistic has an atom in that point and a continuous branch in $\mathbb{R}^+$.

Secondly, the limiting distribution of the number of customers seen in the $GI/M/1$ queueing system by the n^{th} arriving customer is geometric with parameter $(1-\sigma)$, where σ is the root in the interval $(0, 1)$ of the following equation involving the Laplace-Stieltjes transform of the common c.d.f. of the interarrival times:

$$\sigma = \tilde{F}_A[\mu(1-\sigma)] = \int_{a=0}^{+\infty} e^{-\mu(1-\sigma)\,a} dF_A(a) \tag{6}$$

(Kleinrock 1975, p. 251; Adan and Resing 2015, p. 83). Thus, zero is surely the most frequent value of the control statistic when the $GI/M/1$ system is in equilibrium.

Thirdly, the limiting probability generating function (p.g.f.) of the number of customers left behind in the $M/G/1$ queueing system by the n^{th} departing customer is equal to

$$E(z^X) = \frac{(1-\rho)\,\tilde{F}_S[\lambda(1-z)]\,(1-z)}{\tilde{F}_S[\lambda(1-z)] - z}, \; |z| \leq 1, \tag{7}$$

where $\tilde{F}_S(t) = \int_{s=0}^{+\infty} e^{-t\,s} dF_S(s)$ is the Laplace-Stieltjes transform of the common c.d.f. of the service times, according to Adan and Resing (2015, p. 65). Furthermore, Cohen (1982, p. 238) adds that the limiting probability of zero is equal to $(1-\rho)$; as a consequence the most frequent value of the control statistic X_n is surely zero if $\rho \leq 0.5$ while dealing with a $M/G/1$ queueing system. Adan and Resing (2015, p. 65) go on to say that inverting $E(z^X)$ is usually very difficult but, in case $\tilde{F}_S(s)$ is a quotient of polynomials in s (such as when the service times have an Erlang distribution), the limiting p.g.f. can be decomposed into partial fractions, and the

associated limiting p.f. can be easily determined. For instance, if S has an Erlang distribution with two phases and expected value $1/\mu$ then, after some algebraic manipulation, we obtain

$$E(z^X) = \frac{1-\rho}{(1-z/z_1)\,(1-z/z_2)},\ |z| \leq 1, \tag{8}$$

$$P(X=i) = (1-\rho)\left[\frac{z_1}{z_1-z_2}\left(\frac{1}{z_2}\right)^n - \frac{z_2}{z_1-z_2}\left(\frac{1}{z_1}\right)^n\right],\ i \in \mathbb{N}_0, \tag{9}$$

where $z_1 = (2/\rho + 1/2) + \sqrt{2/\rho + 1/4}$, $z_2 = (2/\rho + 1/2) - \sqrt{2/\rho + 1/4}$ and $z_1 z_2 = 4/\rho^2$. This specific limiting p.f. leads us to conclude that $P(X=0) > P(X=1)$ if $\rho^2 + 4\rho - 4 < 0$, that is, if $\rho < \sqrt{8} - 2$ for the $M/E_2/1$ queueing system in equilibrium.

Finally, the high frequency of zero when compared to other values of these three control statistics plays an important role in the design of the X_n-, $\hat{X}_n$- and W_n-charts. Indeed, if we are to set a chart to monitor the traffic intensity with a reasonably large in-control ARL, the LCL has to be equal to zero and the chart is inherently upper one-sided.

In the following section, we briefly describe three simple charts to monitor increases in the traffic intensity and go on to derive their ARL-unbiased versions whose in-control ARL is equal to a pre-specified value $ARL^\star$ and whose ARL curves attain a maximum when the traffic intensity is on target.

In Sect. 3, we present some instructive examples of ARL-unbiased charts for the traffic intensity of various single server queues, with small/medium/large target values, and compare competing control charts in terms of the RL under different out-of-control scenarios.

Section 4 wraps up the chapter with a few comments and recommendations for future work.

2 Detecting Upward and Downward Shifts in the Traffic Intensity

Since many production and service systems can be modelled as queueing systems (Chen and Zhou 2015), control charts can be used to efficiently monitor their traffic intensity. Keep in mind that downward (resp. upward) shifts in the traffic intensity can correspond to a decreasing (resp. increasing) interest in the offered services, thus calling for a timely detection.

The charts, whose performance we are going to describe at the end of this section, give protection to both increases and decreases in the traffic intensity, unlike the upper one-sided charts described by Chen and Zhou (2015) and Morais and Pacheco (2016) and designed to detect solely upward shifts in ρ.

2.1 Three Upper One-Sided Charts for the Traffic Intensity

The traffic intensity is deemed larger from its target level ρ_0 if the control statistic—be it X_n, $\hat{X}_n$ or W_n ($n \in \mathbb{N}$)—is above an upper control limit. Furthermore, if the monitoring of the traffic intensity started with an empty system, which is common practice (Chen et al. 2011), then the number of samples taken until a signal is triggered is given by

$$RL = \min\{n \in \mathbb{N} : Z_n > U \mid Z_0 = 0\}, \tag{10}$$

where:

- $Z_n \equiv X_n, \hat{X}_n, W_n$ is the control statistic we adopted to monitor ρ;
- $U \equiv U_Z$ is a positive integer (resp. real) upper control limit in case $Z_n = X_n, \hat{X}_n$ (resp. $Z_n = W_n$).

According to Morais and Pacheco (2016), *RL* denotes the identity of the first:

- departing (resp. arriving) customer who left behind (resp. found) in the $M/G/1$ (resp. $GI/M/1$) system a number of customers larger than U;
- arriving customer to the $GI/G/1$ system whose waiting time is above U.

In the X_n-chart case, the RL is related to the distribution of the time to absorption of a DTMC with transient states $\{0, \ldots, U\}$ and TPM represented in partitioned form

$$\begin{bmatrix} \mathbf{Q} & (\mathbf{I} - \mathbf{Q})\,\underline{\mathbf{1}} \\ \underline{\mathbf{0}}^\top & 1 \end{bmatrix}, \tag{11}$$

where: $\mathbf{Q} = [p_{ij}]_{i,j=0}^{U}$; $\mathbf{I}$ is the identity matrix with rank $(U+1)$; $\underline{\mathbf{1}}$ (resp. $\underline{\mathbf{0}}^\top$) is a column vector (resp. row vector) of $(U+1)$ ones (resp. zeros).

When we deal with the $\hat{X}_n$-chart we have to consider: the corresponding UCL, $U \equiv U_{\hat{X}}$; $\mathbf{Q} = [\hat{p}_{ij}]_{i,j=0}^{U}$.

Adopting the W_n-chart means the approximate distribution of the RL is related to the time to absorption of a DTMC, say $\{\tilde{W}_n,\ n \in \mathbb{N}_0\}$, with transient states $\{0, 1, \ldots, \tilde{y}-1, \tilde{y}\}$ corresponding to $\{0\} \cup \{((j-1)\Delta,\ j\Delta],\ j = 1, \ldots, \tilde{y}\}$, where: $U \equiv U_W = \tilde{y}\,\Delta$, that is to say U coincides with the upper limit of the last interval; $\tilde{y}$ is a pre-specified large positive integer leading to a very small range $\Delta = U/\tilde{y}$; $\mathbf{Q} = [\tilde{p}_{ij}]_{i,j=\tilde{x}}^{\tilde{y}}$. The resulting approximate run length is also denoted by *RL* for mere convenience.

The exact ARL of the X_n- and $\hat{X}_n$-charts and the approximate ARL of the W_n-chart can be written as

$$ARL^0 = \underline{\mathbf{e}}_0^\top \times (\mathbf{I} - \mathbf{Q})^{-1} \times \underline{\mathbf{1}}, \tag{12}$$

where $\underline{\mathbf{e}}_j$ represents the $(j+1)^{th}$ vector of the orthonormal basis of $\mathbb{R}^{U_X+1}$, $\mathbb{R}^{U_{\hat{X}}+1}$ and $\mathbb{R}^{\tilde{y}+1}$, when $Z_n = X_n, \hat{X}_n, \tilde{W}_n$.

2.2 A Brief Review of ARL-Unbiased Charts

The chart control limits should be set in a way that a peak of the ARL curve is produced at the in-control situation, while maintaining a pre-specified in-control ARL, say $ARL^{\star}$. A chart with the first feature was termed by Pignatiello et al. (1995) an *ARL-unbiased* chart.

As put by Morais (2016), considerable attention has been given to ARL-unbiased charts for parameters of absolutely continuous quality characteristics. Here is a partial list of works in chronological order: Uhlmann (1982, pp. 212–215), Krumbholz (1992), Pignatiello et al. (1995), Ramalhoto and Morais (1995, 1999), Acosta-Mejía and Pignatiello (2000), Huwang et al. (2010), Knoth (2010), Pascual (2010), Cheng and Chen (2011), Huang and Pascual (2011), Pascual (2012), Knoth and Morais (2013, 2015), Guo et al. (2014), and Guo and Wang (2015). The control statistics being used are in most cases independent, in contrast to the Markovian-type statistics X_n, $\hat{X}_n$ and W_n.

Existing ARL-unbiased designs involving discrete distributions are more recent and scarcer. Yang and Arnold (2015) propose an ARL-unbiased exponentially weighted moving average proportion chart to monitor the variance for process data with non-normal or unknown distributions. Paulino et al. (2016a) explore the notions of randomization of the emission of a signal and uniformly most powerful unbiased tests (UMPU) to eliminate the bias of the ARL function of the *c*-chart for i.i.d. Poisson counts and bring the in-control ARL exactly to a pre-specified value; this same technique was used by Morais (2016) to derive an ARL-unbiased *np*-chart, and by Morais (2017) to obtain ARL-unbiased counterparts of the geometric chart and the cumulative count of conforming chart under group inspection. Paulino et al. (2016b) derive an ARL-unbiased design to detect both increases and decreases in the mean of first-order integer-valued autoregressive (INAR(1)) Poisson counts.

As for regulation techniques for the traffic intensity, it is our impression that we did not stumble across any reference tackling the detection of both upward and downward shifts by using a control chart or a combination of two one-sided charts, SPRT or general likelihood procedures. Nonetheless, we ought to make a few comments before we proceed with the description of the ARL-unbiased charts to monitor the traffic intensity of single server queueing systems.

- Bhat and Rao (1972) do not use ARL as a performance measure and only provide two tables for the limits (c_u, c_l) and (d_u, d_l), for the queueing systems $M/E_k/1$ $(k = 1, 2, 3, 4, 5, 10, 15, \infty)$, $\rho_0 = 0.1, 0.2, \ldots, 0.8, 0.9$ (in short $\rho_0 = 0.1(0.1)0.9$ throughout the text), and $\alpha_l = \alpha_u = 0.01, 0.05, 0.1, 0.25$. One of the things that strikes us most forcibly is that this control chart had the

potential to detect increases and decreases in the traffic intensity and was not used with that particular purpose.
- Interestingly, Figure 6 of Chen and Zhou (2015), referring to the ARL comparison between a CUSUM chart and a generalized likelihood ratio (GLR) chart, has the ARL profiles of upper and lower one-sided charts for the traffic intensity. Their combined use could have led to the detection of both upward and downward shifts in the traffic intensity.

2.3 Deriving ARL-Unbiased Charts for the Traffic Intensity

In order to derive ARL-unbiased charts for the traffic intensity when the control statistic is X_n, we can capitalize on the ARL-unbiased c-chart proposed by Paulino et al. (2016b) for the mean of INAR(1) Poisson counts; after all the control statistic employed by those authors and X_n are governed by DTMC with discrete state spaces.

As a consequence, the ARL-unbiased chart used to monitor the traffic intensity of the $M/G/1$ queueing system should trigger a signal at the n^{th} departure with:

- probability one if the number of customers left behind by the n^{th} departing customer, x_n, is larger than the upper control limit U;
- probability γ_L (resp. γ_U) if x_n is equal to $L \equiv 0$ (resp. U).

As duly noted by Paulino et al. (2016b), randomizing the emission of a signal means considering the sub-stochastic matrix $\mathbf{Q} \equiv \mathbf{Q}(\gamma_L, \gamma_U)$ given by

$$\begin{bmatrix} p_{L\,L} \times (1-\gamma_L) & p_{L\,L+1} & \cdots & p_{L\,U-1} & p_{L\,U} \times (1-\gamma_U) \\ p_{L+1\,L} \times (1-\gamma_L) & p_{L+1\,L+1} & \cdots & p_{L+1\,U-1} & p_{L+1\,U} \times (1-\gamma_U) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{U-1\,L} \times (1-\gamma_L) & p_{U-1\,L+1} & \cdots & p_{U-1\,U-1} & p_{U-1\,U} \times (1-\gamma_U) \\ p_{U\,L} \times (1-\gamma_L) & p_{U\,L+1} & \cdots & p_{U\,U-1} & p_{U\,U} \times (1-\gamma_U) \end{bmatrix}. \tag{13}$$

Since $X_0 = 0$ the exact ARL is equal to $ARL^0 = \underline{\mathbf{e}}_0^\top \times [\mathbf{I} - \mathbf{Q}(\gamma_L, \gamma_U)]^{-1} \times \underline{\mathbf{1}}$. Even though $L \equiv 0$, we used the iterative search procedure thoroughly described by Paulino et al. (2016b) to obtain both control limits and the associated randomization probabilities—to bring the in-control ARL to $ARL^\star$ and to eliminate the bias of the ARL function. This search procedure is omitted to keep this chapter to a practical length.

Paulino et al. (2016a) note that the randomization of the emission of the signal can be done in practice by simply using a software to generate a pseudo-random number from a Bernoulli distribution with parameter γ_L (resp. γ_U) every time the control statistic equals L (resp. U).

Needless to say, the ARL-unbiased chart meant to control the traffic intensity of the $GI/M/1$ system can be obtained in a similar fashion.

Like the X_n- and $\hat{X}_n$-charts, the one meant to monitor the traffic intensity of a $GI/G/1$ queue relies on a control statistic governed by a DTMC. There similarity ends because we are now dealing with a nonnegative mixed control statistic. This fact begs for another change: there is no need to randomize the emission of a signal when $W_n = U$ because this event has zero probability.

Since we are supposed to trigger a signal with probability γ_L when $W_n = L \equiv 0$, the sub-stochastic matrix is equal to

$$\tilde{\mathbf{Q}}(\gamma_L) = \begin{bmatrix} \tilde{p}_{L\,L} \times (1-\gamma_L) & \tilde{p}_{L\,L+1} & \cdots & \tilde{p}_{L\,U-1} & \tilde{p}_{L\,U} \\ \tilde{p}_{L+1\,L} \times (1-\gamma_L) & \tilde{p}_{L+1\,L+1} & \cdots & \tilde{p}_{L+1\,U-1} & \tilde{p}_{L+1\,U} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{p}_{U-1\,L} \times (1-\gamma_L) & \tilde{p}_{U-1\,L+1} & \cdots & \tilde{p}_{U-1\,U-1} & \tilde{p}_{U-1\,U} \\ \tilde{p}_{U\,L} \times (1-\gamma_L) & \tilde{p}_{U\,L+1} & \cdots & \tilde{p}_{U\,U-1} & \tilde{p}_{U\,U} \end{bmatrix}, \tag{14}$$

and the ARL is given by $ARL^0 = \underline{\mathbf{e}}_0^\top \times [\mathbf{I} - \tilde{\mathbf{Q}}(\gamma_L)]^{-1} \times \underline{\mathbf{1}}$ because $W_0 = 0$.

Alternatively, we can obtain the ARL by solving an integral equation,[1] using the collocation method that leads to higher accuracy than currently established methods (Knoth 2005) such as the Markov chain approach. For more details on this alternative to the Markov chain approach, the reader is referred to Knoth (2005).

As for the search procedure responsible for the obtention of γ_L and U, it follows the same lines as the algorithm used by Knoth and Morais (2013, 2015) to obtain the control limits of the ARL-unbiased EWMA$-S^2$ chart for the variance of a normally distributed quality characteristic.

3 Preliminary Results

Several programs for the statistical software system R (R Core Team 2013) were used to obtain the ARL-unbiased designs and the corresponding ARL profiles.

Tables 1, 2, 3 and 4 summarize the control limits, the randomization probabilities, and the in-control and two out-of-control ARL values of the ARL-unbiased designs we obtained, by considering the target value of the traffic intensity and the pre-specified in-control ARL equal to $\rho_0 = 0.1(0.1)0.9$ and $ARL^\star = 500$. These ARL-unbiased designs were obtained using the Markov chain approach (in the case of the X_n- and $\hat{X}_n$-charts) and the collocation method (in the case of the W_n-chart) and refer to the control statistics (resp. queueing systems):

- X_n ($M/M/1$, $M/E_2/1$ and $M/E_{100}/1$);
- $\hat{X}_n$ ($M/M/1$, $E_2/M/1$ and $E_5/M/1$);
- W_n ($M/M/1$, $M/E_2/1$ and $E_2/M/1$, either with fixed arrival rate or with fixed service rate).

[1] $\mathcal{L}(z) = 1 + (1-\gamma_L) \times F_{S-A}(-z) \times \mathcal{L}(0) + \int_0^U f_{S-A}(y-z) \times \mathcal{L}(y)\, dy$, where $\mathcal{L}(z)$ represents the ARL of the W_n-chart when $W_0 = z$; the default value of z is zero.

Table 1 ARL-unbiased X_n-chart: control limits, randomization probabilities, in-control and out-of-control ARL values—$\rho_0 = 0.1(0.1)0.9$ and $ARL^\star = 500$

System	ρ_0	$[L, U]$	(γ_L, γ_U)	$ARL(0.95\,\rho_0)$	$ARL(\rho_0)$	$ARL(1.05\,\rho_0)$
$M/M/1$	0.1	[0, 4]	(0.002160, 0.629778)	499.816	500.000	499.805
	0.2	[0, 5]	(0.002377, 0.302403)	499.466	500.000	499.418
	0.3	[0, 6]	(0.002664, 0.080576)	498.884	500.000	498.748
	0.4	[0, 8]	(0.003044, 0.445146)	497.977	500.000	497.669
	0.5	[0, 10]	(0.003568, 0.609947)	496.526	500.000	495.881
	0.6	[0, 12]	(0.004332, 0.214732)	494.133	500.000	492.841
	0.7	[0, 15]	(0.005548, 0.016981)	489.888	500.000	487.347
	0.8	[0, 21]	(0.007769, 0.929236)	481.579	500.000	476.808
	0.9	[0, 30]	(0.013043, 0.709996)	462.258	500.000	455.964
$M/E_2/1$	0.1	[0, 3]	(0.002152, 0.068181)	499.838	500.000	499.829
	0.2	[0, 4]	(0.002370, 0.082010)	499.497	500.000	499.454
	0.3	[0, 5]	(0.002656, 0.073351)	498.923	500.000	498.793
	0.4	[0, 7]	(0.003041, 0.968999)	497.982	500.000	497.664
	0.5	[0, 8]	(0.003566, 0.320705)	496.497	500.000	495.810
	0.6	[0, 10]	(0.004342, 0.423160)	493.949	500.000	492.499
	0.7	[0, 13]	(0.005584, 0.929120)	489.339	500.000	486.316
	0.8	[0, 17]	(0.007876, 0.687065)	480.059	500.000	473.905
	0.9	[0, 24]	(0.013475, 0.066710)	457.401	500.000	447.720
$M/E_{100}/1$	0.1	[0, 3]	(0.002147, 0.328369)	499.855	500.000	499.848
	0.2	[0, 4]	(0.002365, 0.931684)	499.519	500.000	499.479
	0.3	[0, 4]	(0.002640, 0.085670)	498.998	500.000	498.880
	0.4	[0, 5]	(0.003024, 0.183917)	498.072	500.000	497.768
	0.5	[0, 6]	(0.003558, 0.170932)	496.514	500.000	495.797
	0.6	[0, 7]	(0.004350, 0.004998)	493.773	500.000	492.141
	0.7	[0, 9]	(0.005617, 0.027111)	488.750	500.000	485.099
	0.8	[0, 13]	(0.007995, 0.946832)	478.189	500.000	469.929
	0.9	[0, 19]	(0.014002, 0.943674)	450.843	500.000	434.972

By considering the $M/E_2/1$, $M/E_{100}/1$, $E_2/M/1$ and $E_5/M/1$ queueing systems, we cover different sorts of interarrival or service times, in particular with a coefficient of variation not larger than a unit.

The corresponding ARL-profiles can be found in Figs. 1, 2, 3 and 4, for $\rho_0 = 0.1, 0.5, 0.9$ and $ARL^\star = 500$. The profiles in Figs. 1 and 2 (resp. Figs. 3 and 4) were obtained using the Markov chain approach (resp. the collocation method).

The results in those tables and the plots in these figures suggest that we are indeed dealing with charts with:

- in-control ARL very close to the pre-stipulated value $ARL^\star = 500$;
- ARL curves with a maximum when the traffic intensity is equal to its target value ρ_0.

Table 2 ARL-unbiased $\hat{X}_n$-chart: control limits, randomization probabilities, in-control and out-of-control ARL values—$\rho_0 = 0.1(0.1)0.9$ and $ARL^\star = 500$

System	ρ_0	$[L, U]$	(γ_L, γ_U)	$ARL(0.95\,\rho_0)$	$ARL(\rho_0)$	$ARL(1.05\,\rho_0)$
$M/M/1$	0.1	[0, 4]	(0.002160, 0.634850)	499.816	500.000	499.805
	0.2	[0, 5]	(0.002377, 0.307742)	499.467	500.000	499.419
	0.3	[0, 6]	(0.002664, 0.086407)	498.888	500.000	498.753
	0.4	[0, 8]	(0.003043, 0.464904)	497.985	500.000	497.681
	0.5	[0, 10]	(0.003567, 0.651244)	496.545	500.000	495.914
	0.6	[0, 12]	(0.004329, 0.254463)	494.179	500.000	492.931
	0.7	[0, 15]	(0.005541, 0.068832)	490.009	500.000	487.605
	0.8	[0, 20]	(0.007746, 0.088495)	481.923	500.000	477.610
	0.9	[0, 29]	(0.012936, 0.221365)	463.558	500.000	458.852
$E_2/M/1$	0.1	[0, 3]	(0.002039, 0.876869)	499.898	500.000	499.886
	0.2	[0, 4]	(0.002148, 0.859637)	499.582	500.000	499.521
	0.3	[0, 5]	(0.002323, 0.731068)	499.016	500.000	498.846
	0.4	[0, 6]	(0.002580, 0.423089)	498.092	500.000	497.709
	0.5	[0, 7]	(0.002955, 0.065346)	496.559	500.000	495.751
	0.6	[0, 9]	(0.003520, 0.097782)	493.990	500.000	492.361
	0.7	[0, 12]	(0.004438, 0.304139)	489.378	500.000	486.143
	0.8	[0, 16]	(0.006149, 0.182911)	480.157	500.000	473.960
	0.9	[0, 24]	(0.010323, 0.532068)	458.093	500.000	449.697
$E_5/M/1$	0.1	[0, 2]	(0.002004, 0.238163)	499.973	500.000	499.967
	0.2	[0, 3]	(0.002046, 0.652609)	499.731	500.000	499.668
	0.3	[0, 4]	(0.002148, 0.925662)	499.178	500.000	498.977
	0.4	[0, 5]	(0.002324, 0.825244)	498.234	500.000	497.768
	0.5	[0, 6]	(0.002600, 0.408281)	496.673	500.000	495.704
	0.6	[0, 8]	(0.003040, 0.976148)	493.932	500.000	491.935
	0.7	[0, 10]	(0.003771, 0.442419)	489.010	500.000	484.988
	0.8	[0, 13]	(0.005166, 0.020108)	478.917	500.000	470.893
	0.9	[0, 20]	(0.008666, 0.133624)	453.910	500.000	441.644

Interestingly enough, some additional results lead us to believe that, even though taking $X_0 = \hat{X}_0 = W_0 = 0$ could be considered giving a head-start to the three ARL-unbiased charts (especially when $\rho_0 = 0.9$), these designs seem to give proper protection to false alarms. In fact for $\rho_0 = 0.9$, $P[RL^0(\rho_0) = 1]$ (resp. $P[RL^0(\rho_0) \leq 10]$) do not exceed $0.0072 = 3.6 \times (ARL^\star)^{-1}$ (resp. 0.042), for the $M/M/1$, $M/E_2/1$ and $E_2/M/1$ queueing systems.

3.1 M/G/1 Queueing System

Before we continue to comment on the results, we should remind the reader of a known property of the $M/G/1$ queueing systems in equilibrium.

Table 3 ARL-unbiased W_n-chart, FIXED SERVICE RATE: upper control limit, randomization probability, in-control and out-of-control ARL values—$\rho_0 = 0.1(0.1)0.9$ and $ARL^\star = 500$

System	ρ_0	U	γ_L	$ARL(0.95\,\rho_0)$	$ARL(\rho_0)$	$ARL(1.05\,\rho_0)$
$M/M/1$	0.1	7.077585	0.002068	499.949	500.000	499.948
	0.2	8.010018	0.002235	499.784	500.000	499.775
	0.3	9.071026	0.002487	499.457	500.000	499.419
	0.4	10.335393	0.002839	498.866	500.000	498.753
	0.5	11.912665	0.003335	497.823	500.000	497.533
	0.6	13.984468	0.004065	495.951	500.000	495.260
	0.7	16.890639	0.005227	492.431	500.000	490.858
	0.8	21.370674	0.007344	485.203	500.000	481.863
	0.9	29.461491	0.012315	467.940	500.000	463.249
$M/E_2/1$	0.1	4.738168	0.002095	499.925	500.000	499.923
	0.2	5.563324	0.002287	499.711	500.000	499.695
	0.3	6.471000	0.002556	499.307	500.000	499.248
	0.4	7.538379	0.002921	498.599	500.000	498.434
	0.5	8.863481	0.003433	497.372	500.000	496.960
	0.6	10.604398	0.004186	495.194	500.000	494.228
	0.7	13.059001	0.005393	491.102	500.000	488.900
	0.8	16.890368	0.007621	482.601	500.000	477.780
	0.9	24.001438	0.013020	461.564	500.000	453.861
$E_2/M/1$	0.1	6.423954	0.002006	499.989	500.000	499.988
	0.2	6.924320	0.002055	499.896	500.000	499.888
	0.3	7.634301	0.002180	499.644	500.000	499.608
	0.4	8.557287	0.002399	499.124	500.000	499.014
	0.5	9.757652	0.002741	498.137	500.000	497.836
	0.6	11.375280	0.003272	496.264	500.000	495.500
	0.7	13.693807	0.004144	492.567	500.000	490.692
	0.8	17.357775	0.005776	484.579	500.000	480.155
	0.9	24.234798	0.009750	464.162	500.000	456.334

The expected number of customers left behind by a departing customer can be obtained by using the Pollaczek-Khinchin mean-value formula (Kleinrock 1975, p. 187), it is equal to $\rho + [(1+k)/(2k)] \times \rho^2/(1-\rho)$ when we are dealing with E_k service times, and, thus, it is not severely affected by k, in particular for small values of the traffic intensity.

We believe that this last property is in part responsible for the apparent similarity of ARL profiles in Fig. 1, for the $M/M/1$, M/E_2 and $M/E_{100}/1$ systems and a fixed target value ρ_0, namely when $\rho = 0.1$.

The ARL results in Table 1 and the plots in Fig. 1 also suggest that the larger the target value ρ_0 the quicker is the average detection time of small upward and downward shifts in the traffic intensity by the X_n-chart.

Table 4 ARL-unbiased W_n-chart, FIXED ARRIVAL RATE: upper control limit, randomization probability, in-control and out-of-control ARL values—$\rho_0 = 0.1(0.1)0.9$ and $ARL^{\star} = 500$

System	ρ_0	U	γ_L	$ARL(0.95\,\rho_0)$	$ARL(\rho_0)$	$ARL(1.05\,\rho_0)$
$M/M/1$	0.1	0.911543	0.002198	499.505	500.000	499.400
	0.2	1.979006	0.002441	498.848	500.000	498.588
	0.3	3.253009	0.002750	497.952	500.000	497.459
	0.4	4.810239	0.003154	496.688	500.000	495.834
	0.5	6.773410	0.003706	494.831	500.000	493.402
	0.6	9.353440	0.004506	491.943	500.000	489.572
	0.7	12.950170	0.005774	487.093	500.000	483.168
	0.8	18.438763	0.008086	477.988	500.000	471.750
	0.9	28.254818	0.013579	457.637	500.000	451.070
$M/E_2/1$	0.1	0.590423	0.002200	499.450	500.000	499.313
	0.2	1.333646	0.002447	498.719	500.000	498.380
	0.3	2.263006	0.002760	497.722	500.000	497.076
	0.4	3.437682	0.003171	496.316	500.000	495.197
	0.5	4.957663	0.003733	494.249	500.000	492.373
	0.6	6.999475	0.004552	491.027	500.000	487.895
	0.7	9.904859	0.005855	485.578	500.000	480.306
	0.8	14.440527	0.008257	475.177	500.000	466.442
	0.9	22.822927	0.014126	451.037	500.000	440.365
$E_2/M/1$	0.1	0.807874	0.002049	499.785	500.000	499.733
	0.2	1.705269	0.002171	499.279	500.000	499.099
	0.3	2.744663	0.002360	498.488	500.000	498.091
	0.4	3.993490	0.002631	497.298	500.000	496.538
	0.5	5.554374	0.003021	495.474	500.000	494.093
	0.6	7.602153	0.003604	492.537	500.000	490.065
	0.7	10.471464	0.004549	487.439	500.000	482.987
	0.8	14.911901	0.006309	477.480	500.000	469.575
	0.9	23.099001	0.010632	453.865	500.000	443.251

It is interesting to confirm that all the LCL we obtained are equal to zero, unlike the LCL of the ARL-unbiased charts with discrete control statistics derived so far by Paulino et al. (2016a,b) and Morais (2016, 2017).

Another striking feature of the ARL-unbiased X_n-chart: the values of $\gamma_{L\equiv 0}$ tend to be much smaller than the ones of γ_U. As a result, this chart is more prone to trigger a signal when the control statistic is equal to the UCL than when the control statistic takes a zero value. This follows from the need to achieve a fixed and fairly large in-control ARL in the presence of very frequent zero values of the control statistic.

We can also add that larger target values of the traffic intensity require, expectedly, larger upper control limits to achieve the same pre-specified in-control ARL and give proper protection to early false alarms.

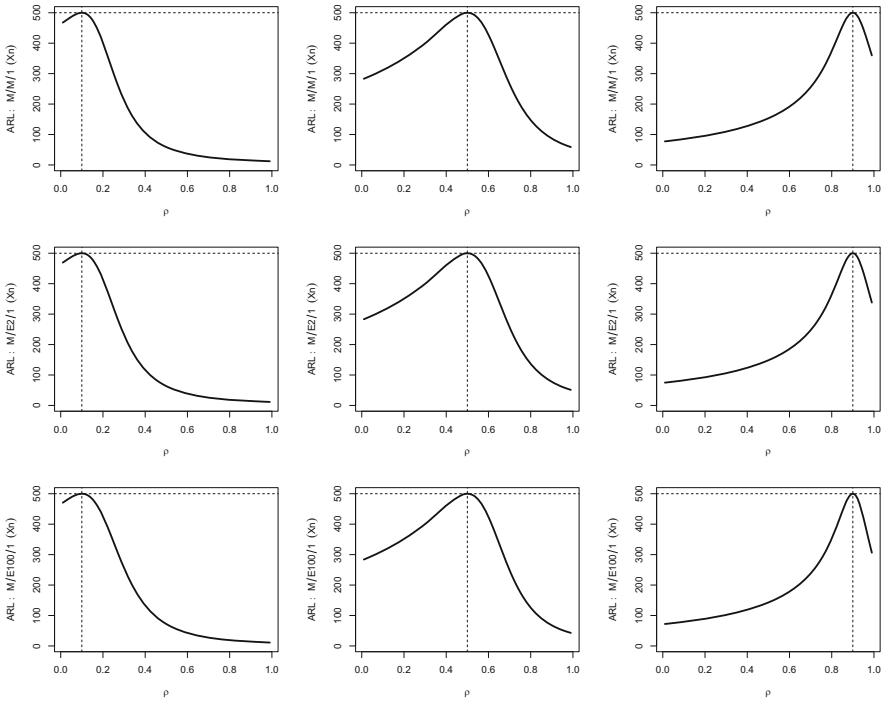


Fig. 1 ARL profiles of the ARL-unbiased X_n-chart—$M/M/1$, $M/E_2/1$ and $M/E_{100}/1$ systems with $\rho_0 = 0.1, 0.5, 0.9$

3.2 GI/M/1 Queueing System

When it comes to the $\hat{X}_n$-chart for the traffic intensity of the $M/M/1$, $E_2/M/1$ and $E_5/M/1$ systems, though comparable for a fixed ρ_0 and different interarrival time distributions, the ARL profiles are dissimilar for distinct target values ρ_0, as illustrated by Fig. 2.

In addition, as the coefficients of variation k^{-1} ($k = 1, 2, 5$) of the interarrival times become smaller and the times between consecutive arrivals become more *regular* for a fixed target value ρ_0, the smaller (resp. larger) is the detection speed of the $\hat{X}_n$-chart in the presence of small and medium (resp. small) size upward and downward shifts in the traffic intensity, as illustrated by the ARL profiles in Fig. 2 (resp. the out-of-control values in Table 2).

The ARL-unbiased design is also associated with null LCL in all cases and small randomization probabilities γ_L, and therefore agrees with what has been previously said and with the results referring to the $M/G/1$ queueing system.

We ought to note that the $\hat{X}_n$- and X_n-charts have similar performances when it comes to the monitoring of the traffic intensity of the $M/M/1$ system, judging by the corresponding ARL profiles in Figs. 1 and 2.

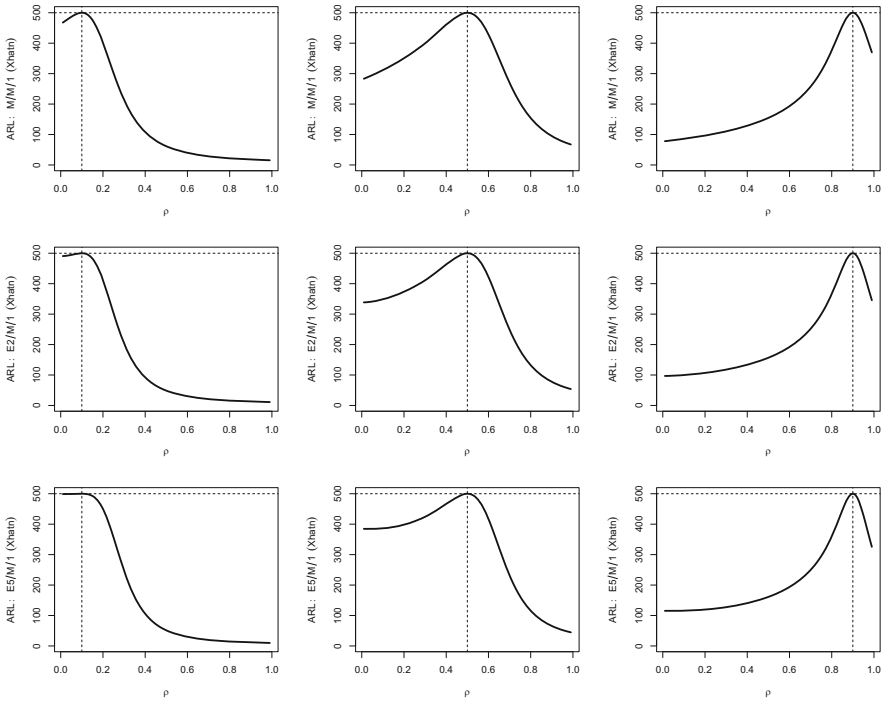


Fig. 2 ARL profiles of the ARL-unbiased $\hat{X}_n$-chart — $M/M/1$, $E_2/M/1$ and $E_5/M/1$ systems with $\rho_0 = 0.1, 0.5, 0.9$

3.3 GI/G/1 Queueing System

Since the RL of the W_n-chart explicitly depends upon the arrival and service rates, the discussion of the results refers now to two scenarios:

- the traffic intensity changes due to change in λ, while the service rate μ is fixed;
- ρ is off-target as a result of a change in μ, whereas the arrival rate λ remains the same.

In both scenarios the probability of triggering a signal when $W_n = L \equiv 0$ does not exceed 1.5% for any of the queueing systems we have considered, like the X_n- and $\hat{X}_n$-charts. The importance of this small randomization probability γ_L lies in its ability to transform these three upper one-sided charts into monitoring schemes that are capable of also detecting decreases in the traffic intensity.

The detection speed of the W_n-chart becomes all the more clearer by looking at the ARL profiles in Figs. 3 and 4:

- the ARL profiles change considerably with the target value ρ_0, as they did for the X_n- and $\hat{X}_n$-charts;

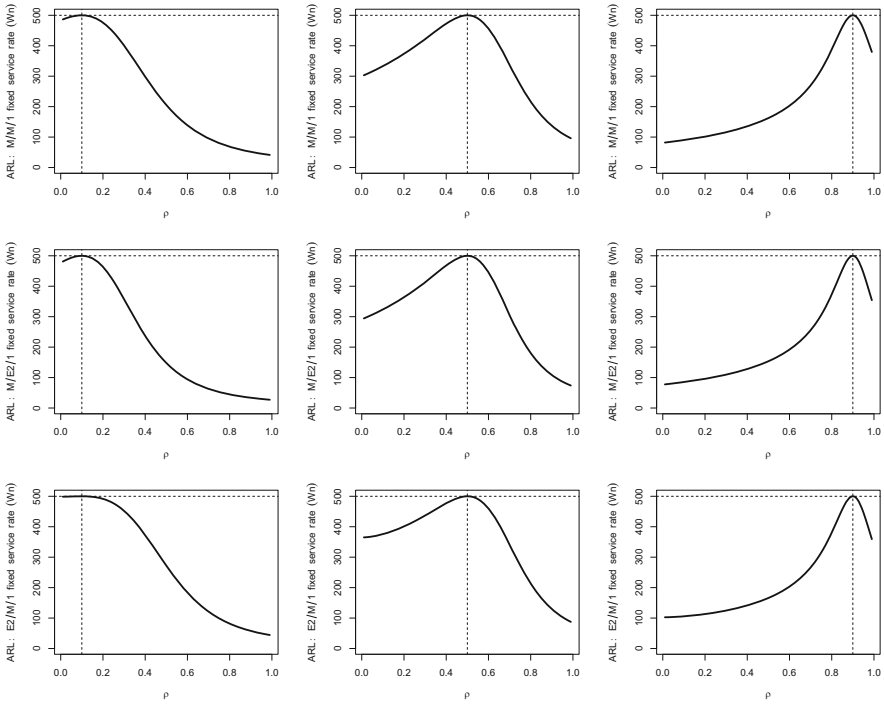


Fig. 3 ARL profiles of the ARL-unbiased W_n-chart, FIXED SERVICE RATE—$M/M/1$, $M/E_2/1$ and $E_2/M/1$ systems with $\rho_0 = 0.1, 0.5, 0.9$

- when the service rate μ is fixed, a change in ρ is due to an increase or decrease of the arrival rate and it seems to be more easily detected if we are monitoring the traffic intensity of the $M/M/1$ and $M/E_2/1$ systems than the traffic intensity of a $E_2/M/1$ queueing system, judging by the corresponding plots in Fig. 3;
- when λ is fixed, the ARL profiles, in Fig. 4, associated with the $M/M/1$ and $M/E_2/1$ queueing systems are very similar for the same target value ρ_0, as we have previously mentioned in the discussion of the results concerning the X_n-chart;
- it is also apparent from Fig. 4 that the W_n-chart seems to take longer to detect decreases in the traffic intensity of the $E_2/M/1$ system with a fixed arrival rate than in the one of the $M/M/1$ and $M/E_2/1$ queueing systems;
- by comparing the ARL profiles in Figs. 3 and 4, we can conclude that a small change in the traffic intensity seems to be detected more swiftly by the W_n-chart if that decrease (resp. increase) in ρ is due to an increase (resp. a decrease) in the service rate than to a decrease (resp. an increase) in the arrival rate, regardless of the queueing system and the target value ρ_0.

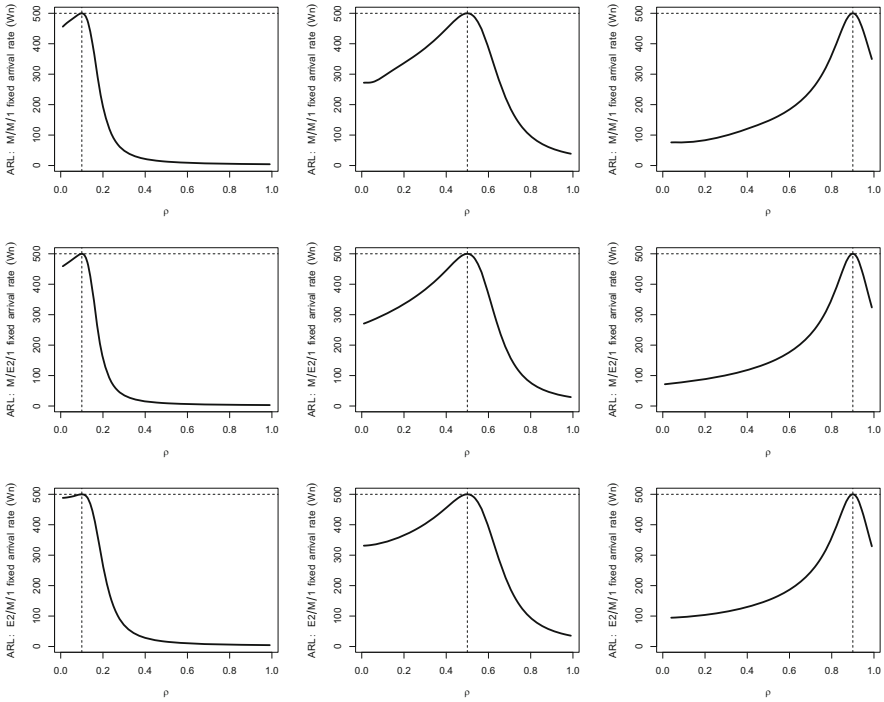


Fig. 4 ARL profiles of the ARL-unbiased W_n-chart, FIXED ARRIVAL RATE—$M/M/1$, $M/E_2/1$ and $E_2/M/1$ systems with $\rho_0 = 0.1, 0.5, 0.9$

3.4 Mixed vs. Discrete Control Statistics

We end this section with a brief discussion on whether or not the ARL-unbiased W_n-chart leads, in average, to swifter detections than its discrete counterparts, the ARL-unbiased X_n- and $\hat{X}_n$-charts, which require less *bookkeeping* and are computationally less demanding as far as their design is concerned.

We limit the confrontations to the X_n- (resp. $\hat{X}_n$-) and W_n-charts meant to control the traffic intensity of the $M/M/1$ and $M/E_2/1$ (resp. $E_2/M/1$) queueing systems.

Programs for *Mathematica* (Wolfram Research, Inc. 2015) were used to produce Fig. 5 (resp. 6), where we can find the plots of the percentage reduction in ARL,

$$\left[1 - \frac{ARL_{W_n}(\rho)}{ARL_{X_n}(\rho)}\right] \times 100\% \qquad (\text{resp. } [1 - ARL_{W_n}(\rho)/ARL_{\hat{X}_n}(\rho)] \times 100\%)$$

when the X_n-chart (resp. $\hat{X}_n$-chart) is replaced with the W_n-chart. The curves were drawn resorting to the Markov chain approach with $(250 + 1)$ transient states.

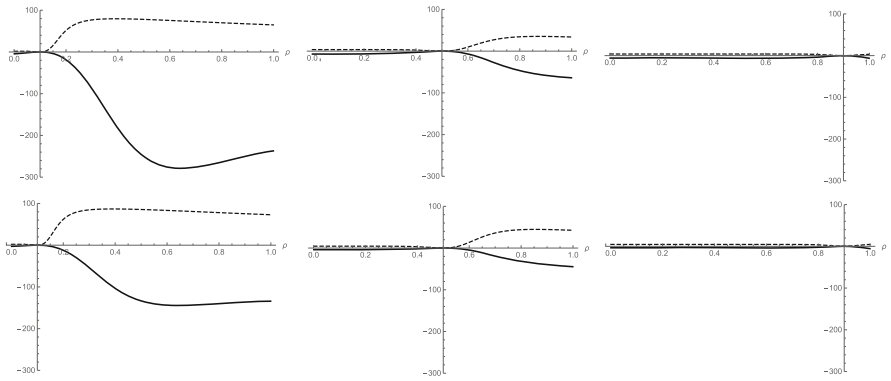


Fig. 5 Plots of the relative ARL reduction, $[ARL_{W_n}(\rho)/ARL_{X_n}(\rho)-1]\times 100\%$—$M/M/1$ (top) and $M/E_2/1$ (bottom) systems with $\rho_0 = 0.1, 0.5, 0.9$ and $ARL^{\star} = 500$; fixed service (resp. arrival) rate corresponds to the solid (resp. dashed) lines

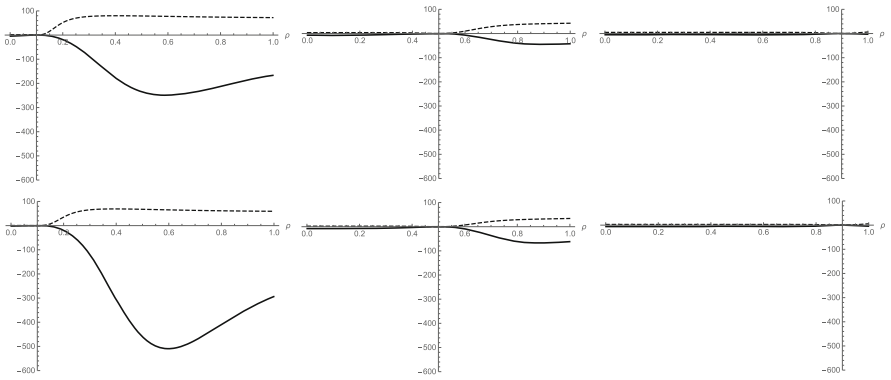


Fig. 6 Plots of the relative ARL reduction, $[ARL_{W_n}(\rho)/ARL_{\hat{X}_n}(\rho)-1]\times 100\%$—$M/M/1$ (top) and $E_2/M/1$ (bottom) systems with $\rho_0 = 0.1, 0.5, 0.9$ and $ARL^{\star} = 500$; fixed service (resp. arrival) rate corresponds to the solid (resp. dashed) lines

Figures 5 and 6 suggest that the ARL profiles of both charts with discrete control statistics compare unfavourably to the one of the W_n-chart, as noted by Morais and Pacheco (2016), when the arrival rate has been fixed (dashed line).

It is also very interesting to see that the smaller the target value of the traffic intensity, the larger seems to be the relative reduction in ARL due to the adoption of the W_n-chart. Thus, extra bookkeeping makes a worthwhile improvement to the detection of shifts in the traffic intensity due to changes in the service rate when $\rho_0 = 0.1$.

The solid lines in these two figures suggest that replacing the X_n- and $\hat{X}_n$-charts with a W_n-chart does not pay-off in terms of ARL performance, when the service rate has been fixed. Strictly speaking, relying on the number of customers seen in the

queueing system by the departing or arriving customer seems to be more beneficial than the waiting time of an arriving customer, when the shifts in the traffic intensity are due entirely on changes in the arrival rate.

For instance, when the traffic intensity of a $E_2/M/1$ queueing system shifts from its target value $\rho_0 = 0.1$ to $\rho = 0.6$, then we would expect to see the first arriving customer, who would have:

- to see at least three customers in upon arrival, to be approximately arrival number 30;
- to wait longer than $U = 6.423954$ time units until being served, to be roughly arrival number 184.

This corresponds to a weighty 509% relative increase in the ARL of the $\hat{X}_n$-chart.

The reader should be aware that in Santos (2016) there is also evidence that using the upper one-sided W_n-chart, to monitor exclusively increases in the traffic intensity when the arrival (resp. service) rate is unaltered, does (resp. does not) improve the detection speed of charts based on the discrete control statistics X_n and $\hat{X}_n$.

4 Conclusion

The aim of this chapter is twofold.

On the one hand, we intend to draw the attention of quality practitioners and operation researchers alike to the use of control charts to monitor the traffic intensity of (single-server) queueing systems.

On the other hand, we make a point of deriving three *ARL-unbiased* charts associated with two discrete-valued and one mixed-valued control statistics. These charts can be easily implemented and are designed in such way that:

- their in-control ARL take a pre-stipulated value $ARL^{\star}$;
- the associated ARL curves attain a maximum when the traffic intensity is on target, thus it takes us less time (in average) to be alerted to any increase or decrease of the traffic intensity than to run into a false alarm.

By relying on the randomization probabilities (resp. probability) γ_L and γ_U (resp. γ_L) to trigger a signal when the control statistic is equal to the LCL or the UCL (resp. LCL), the ARL-unbiased X_n- and $\hat{X}_n$-charts (resp. W_n-chart) for the traffic intensity can definitively handle the *curse* of the null values of the control statistics and still detect decreases in ρ in a timely fashion.

The preliminary results we obtained so far should be complemented with:

- further ARL-unbiased designs, namely referring to other interarrival time distributions such as the hyperexponential and hypoexponential, commonly used in QT and in practice;
- additional comparisons between the two charts with discrete control statistics X_n and $\hat{X}_n$ and the one that makes use of the waiting time W_n, in a scenario suggested by Santos (2016) where the traffic intensity shifts from its target value ρ_0 to

a different value ρ_1 because the arrival and service rates change proportionally from their target values λ_0 and μ_0 to $\lambda_1 = \sqrt{\rho_1/\rho_0}\,\lambda_0$ and $\mu_1 = \sqrt{\rho_0/\rho_1}\,\mu_0$, respectively; these comparisons should rely not only on ARL but also on the RL percentage points and its standard deviation (SDRL).

A direction of future research comprises the derivation of ARL-unbiased versions of the *WZ*, the *nL* and the sophisticated CUSUM charts proposed by Bhat and Rao (1972), Chen et al. (2011) and Chen and Zhou (2015) (respectively), in order to detect not only increases and but also decreases in the traffic intensity of (single-server) queueing systems in an expedient manner.

Acknowledgements The first author gratefully acknowledges: the financial support received from CEMAT (Center for Computational and Stochastic Mathematics) to attend the XIIth International Workshop on Intelligent Statistical Quality Control, Hamburg, Germany, August 16–19, 2016; the partial support given by FCT (Fundação para a Ciência e a Tecnologia) through projects UID/Multi/04621/2013, PEst-OE/MAT/UI0822/2014 and PEst-OE/MAT/UI4080/2014.

We are greatly indebted to: Prof. António Pacheco, for drawing our attention to the potential of the application of SPC in the monitoring of the traffic intensity of queueing systems; Prof. Christian Weiss, for having alerted us to the publication of Chen and Zhou (2015); Marta Santos, for the stimulating discussions during the preparation of her M.Sc. thesis (Santos 2016); Profs. Peter-Theodor Wilrich, William H. Woodall, Eugénio K. Epprecht and Murat C. Testik, and Dr. Detlef Steuer, for the encouraging words and the valuable feedback following our presentation in the IWISQC2016.

Appendix

If the service times of an $M/G/1$ queueing system have an Erlang distribution with k ($k \in \mathbb{N}$) phases and probability density function (p.d.f.) given by

$$f_S(s) = (k\mu)^k\, s^{k-1}\, e^{-k\mu s}/(k-1)!, \quad s \geq 0,$$

then

$$\alpha_i = \binom{k+i-1}{k-1} \left(\frac{\rho}{k+\rho}\right)^i \left(\frac{k}{k+\rho}\right)^k, \quad i \in \mathbb{N}_0 \tag{15}$$

(Feller 1971, p. 57). In other words, Y has a negative binomial distribution with parameters k and $k(k+\rho)^{-1}$, when we are dealing with the $M/E_k/1$ queueing system.

If the $GI/M/1$ queueing system is associated with interarrival times with an Erlang distribution with density

$$f_A(a) = (k\lambda)^k\, a^{k-1}\, e^{-k\lambda a}/(k-1)!, \quad a \geq 0,$$

Morais and Pacheco (2016) adds that

$$\hat{\alpha}_i = \binom{k+i-1}{k-1} \left(\frac{k^{-1}}{k^{-1}+\rho}\right)^i \left(\frac{\rho}{k^{-1}+\rho}\right)^k, \quad i \in \mathbb{N}_0. \tag{16}$$

This is to say that Y has a negative binomial distribution with parameters k and $\rho\,(k^{-1}+\rho)^{-1}$, for the $E_k/M/1$ queue.

When it comes to the $GI/G/1$ queueing system, the results derived by Nadarajah and Kotz (2005), for the c.d.f. and p.d.f. of a linear combination $(\alpha X + \beta Y)$ of exponential (X) and gamma (Y) independent r.v. (with $\alpha > 0$), come in handy.

For the $M/M/1$ queueing system with arrival rate $\lambda = 1/E(A)$ and service rate $\mu = 1/E(S)$, Morais and Pacheco (2016) wrote

$$F_{S-A}(x) = \begin{cases} \frac{\mu\, e^{\lambda x}}{\lambda+\mu}, & x \leq 0 \\ 1 - \frac{\lambda\, e^{-\mu x}}{\lambda+\mu}, & x > 0. \end{cases} \tag{17}$$

Similar calculations led Morais and Pacheco (2016) to conclude that:

$$F_{S-A}(x) = \begin{cases} e^{\lambda x} \left(\frac{k\mu}{k\mu+\lambda}\right)^k, & x \leq 0 \\ F_{Gamma(k,k\mu)}(x) + e^{\lambda x} \left(\frac{k\mu}{k\mu+\lambda}\right)^k \bar{F}_{Gamma(k,k\mu+\lambda)}(x), & x > 0, \end{cases} \tag{18}$$

for the $M/E_k/1$ queueing system; and

$$F_{S-A}(x) = \begin{cases} \bar{F}_{Gamma(k,k\lambda)}(-x) \\ \qquad -e^{-\mu x} \left(\frac{k\lambda}{k\lambda+\mu}\right)^k \bar{F}_{Gamma(k,k\lambda+\mu)}(-x), & x \leq 0 \\ 1 - e^{-\mu x} \left(\frac{k\lambda}{k\lambda+\mu}\right)^k, & x > 0, \end{cases} \tag{19}$$

for the $E_k/M/1$ system.

References

Acosta-Mejía, C. A., & Pignatiello, J. J. Jr. (2000). Monitoring process dispersion without subgrouping. *Journal of Quality Technology, 32*, 89–102.

Adan, I., & Resing, J. (2015). *Queueing Theory*. Department of Mathematics and Computing Science, Eindhoven University of Technology. http://www.win.tue.nl/~iadan/queueing.pdf. Accessed 27 May 2016.

Beneš, V. E. (1957). A sufficient set of statistics for a simple telephone exchange model. *Bell System Technical Journal, 36*, 939–964.

Bhat, U. N. (1987). A statistical technique for the control of traffic intensity in Markovian queue. *Annals of Operations Research, 8*, 151–164.

Bhat, U. N., & Rao, S. S. (1972). A statistical technique for the control of traffic intensity in the queuing systems M/G/1 and GI/M/1. *Operations Research, 20*, 955–966.

Brook, D., & Evans, D. A. (1972). An approach to the probability distribution of CUSUM run length. *Biometrika, 59*, 539–549.

Chen, N., Yuan, Y., & Zhou, S. (2011). Performance analysis of queue length monitoring of M/G/1 systems. *Naval Research Logistics, 58*, 782–794.

Chen, N., & Zhou, S. (2015). CUSUM statistical monitoring of M/M/1 queues and extensions. *Technometrics, 57*, 245–256.

Cheng, C.-S., & Chen, P.-W. (2011). An ARL-unbiased design of time-between-events control charts with runs rules. *Journal of Statistical Computation and Simulation, 81*, 857–871.

Clarke, A. B. (1957). Maximum likelihood estimates in a simple queue. *The Annals of Mathematical Statistics, 28*, 1036–1040.

Cohen, J. W. (1982). *The single server queue* (revised edition). Amsterdam: North-Holland.

Cox, D. R. (1965). Some problems of statistical analysis connected with congestion. In W. L. Smith & W. E. Wilkinson (Eds.), *Proceedings of the Symposium on Congestion Theory* (pp. 289–316). Chapel Hill: University of North Carolina Press.

Erlang, A. K. (1909). Sandsynlighedsregning og Telefonsamtaler. *Nyt Tidsskrift for Matematik B (Copenhagen), 20*, 33–41; Translation: The theory of probabilities and telephone conversations. In: Brockmeyer, Halstrøm & Jensen (1948, pp. 131–137).

Erlang, A. K. (1917). Løsning af nogle Problemer fra Sandsynlighedsregningen af Betydning for de automatiske Telefoncentraler. *Elektrotkeknikeren (Copenhagen), 13*, 5–13; Translation: Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. In: Brockmeyer, Halstrøm & Jensen (1948, pp. 138–155).

Erlang, A. K. (1920). Telefon-Ventetider. Et Stykke Sandsynlighedsregning. *Matematisk Tidsskrift B (Copenhagen), 31*, 25–42; Translation: Telephon waiting times: an example of probability calculus. In: Brockmeyer, Halstrøm & Jensen (1948, pp. 156–171).

Feller, W. (1971). *An introduction to probability theory and its applications* (2nd ed.). New York: John Wiley & Sons.

Greenberg, I. (1997). Markov chain approximation methods in a class of level-crossing problems. *Operations Research Letters, 21*, 153–158.

Guo, B., & Wang, B. X. (2015). The design of the ARL-unbiased S^2 chart when the in-control variance is estimated. *Quality and Reliability Engineering International, 31*, 501–511.

Guo, B., Wang, B. X., & Xie, M. (2014). ARL-unbiased control charts for the monitoring of exponentially distributed characteristics based on type-II censored samples. *Journal of Statistical Computation and Simulation, 84*, 2734–2747.

Huang, X., & Pascual, F. (2011). ARL-unbiased control charts with alarm and warning lines for monitoring Weibull percentiles using the first-order statistic. *Journal of Statistical Computation and Simulation, 81*, 1677–1696.

Hung, Y.-C., Michailidis, G., & Chuang, S.-C. (2012). Estimation and monitoring of traffic intensities with application to control of stochastic systems. *Applied Stochastic Models in Business and Industry, 30*, 200–217.

Huwang, L., Huang, C.-J., & Wang, Y.-H. T. (2010). New EWMA control charts for monitoring process dispersion. *Computational Statistics and Data Analysis, 54*, 2328–2342.

Jain, S. (2000). An autoregressive process and its application to queueing model. *Metron-International Journal of Statistics, 58*, 131–138.

Jain, S., & Templeton, J. G. C. (1989). Problem of statistical inference to control the traffic intensity. *Sequential Analysis, 8*, 135–146.

Kendall, D. G. (1951). Some problems in the theory of queues. *Journal of the Royal Statistical Society, Series B (Methodological), 13*, 151–185.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics, 24*, 338–354.

Kim, S.-H., Alexopoulos, C., Tsui, K.-L., & Wilson, J. R. (2007). A distribution-free tabular CUSUM chart for autocorrelated data. *IIE Transactions, 39*, 317–330.

Kleinrock, L. (1975). *Queueing systems, volume I: Theory*. New York: John Wiley & Sons.
Knoth, S. (2005). Accurate ARL computation for EWMA-S^2 control charts. *Statistics and Computing, 15*, 341–352.
Knoth, S. (2010). Control charting normal variance–reflections, curiosities, and recommendations. In H.-J. Lenz & P.-T. Wilrich (Eds.), *Frontiers in statistical quality control* (Vol. 9, pp. 3–18). Heidelberg: Physica.
Knoth, S., & Morais, M. C. (2013). On ARL-unbiased control charts. In S. Knoth, W. Schmid, & R. Sparks (Eds.), *Proceedings of the XIth International Workshop on Intelligent Statistical Quality Control* (pp. 31–50), Sydney, Australia, 20–23 August 2013.
Knoth, S., & Morais, M. C. (2015). On ARL-unbiased control charts. In S. Knoth, & W. Schmid (Eds.), *Frontiers in Statistical Quality Control* (Vol. 11, pp. 95–117). Switzerland: Springer.
Krumbholz, W. (1992). Unbiased control charts based on the range. *Österreichische Zeitschrift für Statistik und Informatik, 22*, 207–218.
Lilliefors, H. W. (1966). Some confidence intervals for queues. *Operations Research, 14*, 723–727.
Montgomery, D. C. (2009). *Introduction to statistical quality control* (6th ed.). New York: John Wiley & Sons.
Morais, M. C. (2016). An ARL-unbiased np-chart. *Economic Quality Control, 31*, 11–21.
Morais, M. C. (2017). ARL-unbiased geometric and CCC_G control charts. *Sequential Analysis, 36*, 513–527.
Morais, M. C., & Pacheco, A. (1998). Comparing first passage times of Markovian processes. *Proceedings of the Second International Symposium on Semi-Markov Models: Theory and Applications*. Compiègne, France, December 9–11, 1998.
Morais, M. C., & Pacheco, A. (2016). On stochastic ordering and control charts for the traffic intensity. *Sequential Analysis, 35*, 536–559.
Nadarajah, S., & Kotz, S. (2005). On the linear combination of exponential and gamma random variables. *Entropy, 7*, 161–171.
Pascual, F. (2010). EWMA charts for the Weibull shape parameter. *Journal of Quality Techonology, 42*, 400–416.
Pascual, F. (2012). *Individual and Moving Ratio Charts for Weibull Processes*. Technical Report (#2012-3), Department of Mathematics, Washington State University.
Paulino, S., Morais, M. C., & Knoth, S. (2016a). An ARL-unbiased c-chart. *Quality and Reliability Engineering International, 32*, 2847–2858.
Paulino, S., Morais, M. C. & Knoth, S. (2016b). On ARL-unbiased c-charts for INAR(1) Poisson counts. *Statistical Papers*. http://rdcu.be/nGs8
Pignatiello, J. J. Jr., Acosta-Mejía, C. A., & Rao, B. V. (1995). The performance of control charts for monitoring process dispersion. In: *4th Industrial Engineering Research Conference* (pp. 320–328).
R Core Team (2013).*R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org
Ramalhoto, M. F., & Morais, M. (1995). Cartas de controlo para o parâmetro de escala da população Weibull tri-paramétrica (Control charts for the scale parameter of the Weibull population). In: *Actas do II Congresso Anual da Sociedade Portuguesa de Estatística (Proceedings of the Annual Congress of the Portuguese Statistical Society)* (pp. 345–371).
Ramalhoto, M. F., & Morais, M. (1999). Shewhart control charts for the scale parameter of a Weibull control variable with fixed and variable sampling intervals. *Journal of Applied Statistics, 26*, 129–160.
Rao, S. S., Bhat, U. N., & Harishchandra, K. (1984). Control of traffic intensity in a queue–a method based on SPRT. *Opsearch, 21*, 63–80.
Santos, M. D. M. (2016). *On Control Charts and the Detection of Increases in the Traffic Intensity of Queueing Systems*. M. Sc. thesis, Instituto Superior Técnico, Universidade de Lisboa.
Shore, H. (2000). General control charts for attributes. *IIE Transactions, 32*, 1149–1160.
Shore, H. (2006). Control charts for the queue length in a G/G/s system. *IIE Transactions, 38*, 1117–1130.
Uhlmann, W. (1982). *Statistische Qualitätskontrolle (2. Aufl.)*. Stuttgart: Teubner.

Western Electrical (1956). *Statistical Quality Control Handbook*. Indianopolis: Western Electrical Corporation.

Wolfram Research, Inc. (2015). *Mathematica, Version 10.3*, Champaign, IL. http://reference.wolfram.com/language/. Accessed 31 March 2016.

Yang, S.-F., & Arnold, B. C. (2015). Monitoring process variance using an ARL-unbiased EWMA-p control chart. *Quality and Reliability Engineering International, 32*, 1227–1235.

Zobu, M., & Sağlam, V. (2013). Control of traffic intensity in hyperexponential and mixed Erlang queueing systems with a method based on SPRT. *Mathematical Problems in Engineering*. Article ID 241241, 9 pages. http://www.hindawi.com/journals/mpe/2013/241241/. Accessed 11 November 2015.

Risk-Adjusted Exponentially Weighted Moving Average Charting Procedure Based on Multi-Responses

Xu Tang and Fah Fatt Gan

Abstract Quality control charting procedures like cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) charting procedures are traditionally used for monitoring the quality of manufactured products. Unlike a manufacturing process where the raw material is usually reasonably homogeneous, patients' risks of various surgical outcomes are usually quite different. The risks will have to be taken into consideration when monitoring surgical performances. A risk-adjusted CUSUM charting procedure for monitoring surgical performances has already been developed in the literature. In this chapter, we develop a risk-adjusted EWMA charting procedure based on two or more outcomes. The properties of this procedure are studied. It is also compared with the risk-adjusted CUSUM procedure using a real surgical data set. Our study shows that the risk-adjusted EWMA procedure is an attractive alternative because of its performance and ease of interpretation.

Keywords Cumulative sum charting procedure · Odds ratio · Parsonnet scores · Patient mix · Proportional odds logistic regression model · Quality monitoring · Surgical outcomes

1 Introduction

The need for effective monitoring of surgical performances has gained much attention in recent years after the public was alerted to a high profile case of professional misconduct over the quality of heart surgeries (BRI Inquiry Panel 2001). Treasure et al. (1997), Waldie (1998) and Treasure et al. (2004) have also highlighted several other critical cases. The importance of effective online

X. Tang · F. F. Gan (✉)
Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore
e-mail: staganff@nus.edu.sg

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_6

monitoring procedures cannot be understated because such procedures allow prompt detection of any deterioration in surgical performance, and hence investigations of possible causes and eventually reduction in undesirable outcomes.

In a manufacturing process, raw material fed into the process is often quite homogeneous. The added complexity in monitoring surgical performances is that patients usually have different health conditions which affect the surgical outcomes directly. If the heterogeneity of patients is not taken into consideration, then monitoring procedures could lead to misleading inferences (Steiner et al. 2000). To estimate the risk of death from a cardiac operation, Parsonnet (1989) proposed an additive scoring system based on a patient's health condition like age, blood pressure, existence of certain disease such as diabetes, morbid obesity etc. This score is commonly known as the Parsonnet score. Steiner et al. (2000) for example, fitted a binary logistic regression model using the Parsonnet score as the explanatory variable to estimate the probability of death from a cardiac operation. The Euroscore which was developed by Roques et al. (1999) for estimating the probability of death was also obtained by fitting a binary logistic regression model. Their model is based on 19,030 cardiac surgeries, using various measures of health condition as explanatory variables. For three or more surgical outcomes, Tang et al. (2015) fitted a proportional odds logistic regression model using the Parsonnet score as the explanatory variable to estimate the probabilities of various surgical outcomes.

The earliest risk-adjusted monitoring procedure was developed by Lovegrove et al. (1997, 1999) and Poloniecki et al. (1998). Their simple risk-adjustment is done using the difference between the surgical outcome (0 for survival within 30 days and 1 for death) and the estimated probability of death. The main disadvantage of this procedure is the lack of a proper signaling rule. The risk-adjusted cumulative sum (CUSUM) charting procedure developed by Steiner et al. (2000) is based on accumulating the log likelihood ratio derived from testing the odds ratio that a patient dies. This chart is also based on the same binary outcomes. A more general risk-adjusted CUSUM procedure obtained by testing the probability of death was given by Gan et al. (2012). In order to improve the effectiveness of this procedure, Tang et al. (2015) developed a risk-adjusted CUSUM procedure based on two or more outcomes: death and different grades of survival. Grigg and Spiegelhalter (2007) developed a risk-adjusted exponentially weighted moving average (EWMA) chart for exponential family data. Steiner and Jones (2010) developed a risk-adjusted EWMA chart based on survival time. Their EWMA chart is only feasible for monitoring surgical performances with two outcomes. However, more effective procedures can be obtained by classifying the surgical outcomes into more than two outcomes as explained in Tang et al. (2015). None of the binary risk-adjusted EWMA procedures is a special case of our proposed procedure. A recent review chapter on monitoring surgical outcomes can be found in Woodall et al. (2015).

In this chapter, we will develop a risk-adjusted EWMA chart based on two or more outcomes. In Sect. 2, a proportional odds logistic regression model is used to estimate the probabilities of various surgical outcomes. We then develop a risk-adjusted statistic based on the likelihood ratio approach. The properties of this statistic are investigated and conditions are derived for it to be a reasonable

monitoring statistic. In Sect. 3, we develop a risk-adjusted EWMA charting procedure based on this statistic. The risk-adjusted EWMA and CUSUM procedures are used to study the performance of three surgeons based on a real data set in Sect. 4. Similarities and differences between these two procedures are compared. Conclusions are given in Sect. 5.

2 Proportional Odds Logistic Regression Model and Log Likelihood Ratio Statistic

The Parsonnet score S measures the mortality risk of a patient undergoing a cardiac surgery. The outcome is usually determined after 30 days of an operation and it can be represented by a discrete random variable Y which takes a value from 0 to J. Let $Y = 0$ when a patient has a fully recovery, $Y = 1, 2, \ldots, J-1$ denote various states of partial recovery, with a smaller number associated with a better state of recovery and $Y = J$ when a patient dies.

We will follow the notation used by Tang et al. (2015). Conditional on a patient's risk score $S = s$, the distribution Y is denoted as

$$P(Y = k|S = s) = \pi_k(s),\ k = 0, 1, \ldots, J.$$

The cumulative logit is defined as

$$\text{logit}[P(Y \le k|S = s)] = \log\left[\frac{P(Y \le k|S = s)}{1 - P(Y \le k|S = s)}\right] = \log\left[\frac{\pi_0(s) + \cdots + \pi_k(s)}{\pi_{k+1}(s) + \cdots + \pi_J(s)}\right],$$

where $k = 0, \ldots, J-1$. The cumulative distribution function of Y can be estimated using the proportional odds logistic regression model (McCullagh 1980) as

$$\text{logit}[P(Y \le k|S = s)] = \alpha_k + \beta s,\ k = 0, \ldots, J-1, \tag{1}$$

based on a historical data set of patients' risk scores and surgical outcomes. The model is based on the assumption that the cumulative logits share the same slope β but with different intercepts, α_k's. The assumption of parallel logit surfaces is known as the proportional odds assumption. For this application, the parameter α_k is increasing in k because the probability $P(Y \le k|S = s)$ increases in k for all s and the logit is an increasing function of this probability. Also, the cumulative probability $P(Y \le k|S = s)$ decreases with increasing risk score s and hence the parameter β is negative.

Following the notation used by Tang et al. (2015), we let the probability density function (pdf) of the risk score of a patient be $f(s)$. The joint density of (S, Y) is then given as $f(s, y) = \pi_y(s)f(s)$, $y = 0, \ldots, J$. We consider testing the null hypothesis $H_0 : f_0(s, y)$ against the alternative hypothesis $H_A : f_A(s, y)$ where $(\pi_0(s), \ldots, \pi_J(s)) = (\pi_0^0(s), \ldots, \pi_J^0(s))$ under the null hypothesis and $(\pi_0(s), \ldots, \pi_J(s)) = (\pi_0^A(s), \ldots, \pi_J^A(s))$ under the alternative hypothesis.

The nth log likelihood ratio statistic is given by

$$W_n = \log(f_A(S_n, Y_n)/f_0(S_n, Y_n)).$$

The statistic W_n is hence obtained by risk-adjusting Y_n using S_n. The joint pdf's under the null and alternative hypotheses are given by $f_0(s_n, y_n) = \pi^0_{y_n}(s_n)f(s_n)$ and $f_A(s_n, y_n) = \pi^A_{y_n}(s_n)f(s_n)$ respectively, hence,

$$W_n = \log(\pi^A_{Y_n}(S_n)/\pi^0_{Y_n}(S_n)). \tag{2}$$

The statistic W_n does not contain $f(s_n)$ because the risk distribution is assumed to be the same for both hypotheses.

Based on the multi-response proportional odds logistic regression model, a natural way of defining performance of a surgeon is to use the one based on cumulative probabilities,

$$\frac{\sum_{i=0}^{k} \pi_i^*(s)}{1 - \sum_{i=0}^{k} \pi_i^*(s)} = R_k \frac{\sum_{i=0}^{k} \pi_i(s)}{1 - \sum_{i=0}^{k} \pi_i(s)}, \tag{3}$$

$k = 0, \cdots, J-1$ where R_k is the odds ratio of cumulative probabilities of recovery. In order for the probabilities $\pi_k^*(s), k = 0, \cdots, J$ to be in $[0, 1]$, Tang et al. (2015) showed that the odds ratios must satisfy the condition

$$\alpha_0 + \log(R_0) \le \alpha_1 + \log(R_1) \le \cdots \le \alpha_{J-1} + \log(R_{J-1}). \tag{4}$$

In practice, we may assume that $R_0 = \ldots = R_{J-1} = 1$ under the null hypothesis which means that the performance under the null hypothesis is characterized by the fitted logistic regression model. The values of R_k's can then be set to be greater than 1 for detecting improvement and less than 1 for detecting deterioration. Once an alternative hypothesis is chosen, the monitoring statistic $W(Y, S)$ is then defined by Eq. (2).

Let the target alternative performance be $\pi_Y^+(S)$ based on odds ratios $R_0^+, \cdots, R_{J-1}^+$ for detecting improvement and $\pi_Y^-(S)$ based on $R_0^-, \cdots, R_{J-1}^-$ for detecting deterioration. Then, the statistics for detecting improvement and deterioration can be determined using Eq. (2) as

$$W^+(Y, S) = \log(\pi_Y^+(S)/\pi_Y^0(S)),$$

and

$$W^-(Y, S) = \log(\pi_Y^-(S)/\pi_Y^0(S)),$$

respectively. One could use a charting procedure based on $W^+(Y, S)$ for monitoring improvement and another procedure based on $W^-(Y, S)$ for monitoring deterioration but this would involve two procedures. We propose the statistic

$$W_a(Y, S) = W^+(Y, S) - W^-(Y, S). \tag{5}$$

as the monitoring statistic. This statistic has some attractive properties. The statistic can be expressed as

$$W_a(Y, S) = \log(\pi_Y^+(S)/\pi_Y^-(S)).$$

It is the log likelihood ratio of the probability of an outcome Y given a risk score S assuming a surgeon performing better than average to that of a surgeon performing worst than average. This provides mathematical support for the use of this statistic for monitoring.

The statistic $W_a(Y, S)$ can also be viewed meaningfully as a penalty-reward score for monitoring. In general, a reward score is given for a successful operation and a penalty score is given for a failed operation. The penalty-reward score is a positive number if it is a reward, and a negative number if it is a penalty. Given a particular outcome $Y = k,\ k = 0, \cdots, J$, the penalty-reward score should increase as the risk score increases. This means that for detecting deterioration, a surgeon should be given a lower penalty score for a higher-risk patient given the same outcome. Also, for detecting an improvement, a surgeon should be given a higher reward score for a higher-risk patient given the same outcome. For a given risk score s, we also require $W_a(0, s) > W_a(1, s) > \cdots > W_a(J, s)$ to be satisfied. This defines a proper ordering of the penalty-reward score for all the outcomes.

In order for $W_a(Y, S)$ to satisfy the property that $W_a(y, s)$ is an increasing function of s and a decreasing function of y, it only requires the condition in Theorem 1 to be true. The proof of this theorem is given in Appendix 1.

Theorem 1 *Assume Eqs. (1), (2) and (3) hold. Suppose* $R_0^+, \cdots, R_{J-1}^+$ *define* $\pi_Y^+(S)$*, and* $R_0^-, \cdots, R_{J-1}^-$ *define* $\pi_Y^-(S)$*. Then the condition*

$$R_0^+/R_0^- = \cdots = R_{J-1}^+/R_{J-1}^- > 1,$$

is necessary and sufficient for $W_a(y, s)$ *to be (i) an increasing function of s given y, and (ii) a decreasing function of y given s.*

Additional properties of the statistic are given in Theorem 2. The proof of this theorem is given in Appendix 2.

Theorem 2 *Assume Eqs. (1), (2) and (3) hold. If*

$$R_0^+/R_0^- = \cdots = R_{J-1}^+/R_{J-1}^- = R^+/R^- > 1,$$

then $W_a(y, s)$ satisfies the following condition:

1. $W_a(0, s) > 0$.
2. $W_a(J, s) < 0$.
3. $W_a(0, s) \to 0$ *when* $s \to -\infty$, $W_a(J, s) \to 0$ *when* $s \to \infty$.
4. *For* $y \in \{0, \cdots, J-1\}$, $W_a(y, s) \to log(R^+/R^-)$ *when* $s \to \infty$.
5. *For* $y \in \{1, \cdots, J\}$, $W_a(y, s) \to -log(R^+/R^-)$, *when* $s \to -\infty$.

Note that Theorems 1 and 2 do not require $R_0^+ = \cdots = R_{J-1}^+ = R^+$ and $R_0^- = \cdots = R_{J-1}^- = R^-$. They only require $R_0^+/R_0^-, \cdots, R_{J-1}^+/R_{J-1}^-$ to be the same as the ratio R^+/R^-. The condition $R_0^+ = \cdots = R_{J-1}^+ = R^+$ and $R_0^- = \cdots = R_{J-1}^- = R^-$ is just a special case and the more natural one to use. Hence, in this chapter, we will be using $R_0^+ = \cdots = R_{J-1}^+ = R^+$ and $R_0^- = \cdots = R_{J-1}^- = R^-$.

The properties of $W_a(y, s)$ as stated in Theorems 1 and 2 can be explained further using Fig. 1 which shows a plot of $W_a(y, s)$ against s for $y = 0$, 1 and 2. This figure shows that as reward, the score $W_a(y, s)$ is positive and as penalty, negative. Results 1 and 2 of Theorem 2 show that the penalty-reward score $W_a(0, s)$ is positive for full recovery and negative for death. For partial recovery, results 4 and 5 of Theorem 2 show that $W_a(k, s) < 0$ for s less than some s^* and $W_a(k, s) > 0$ for s greater than s^*. Thus, for a patient with a risk s less than s^*, the penalty-reward score is negative (a penalty) if the patient makes a partial recovery. On the other hand, for a patient with a risk s greater than s^*, the penalty-reward is positive (a reward) if the patient makes a partial recovery. This is reasonable because if a high-risk patient makes even a partial recovery, this is considered a desirable outcome, whereas if a low-risk patient who is more likely to make a full recovery, makes only a partial recovery, this is not considered a desirable outcome. Note that the score $W_a(0, s)$ is always

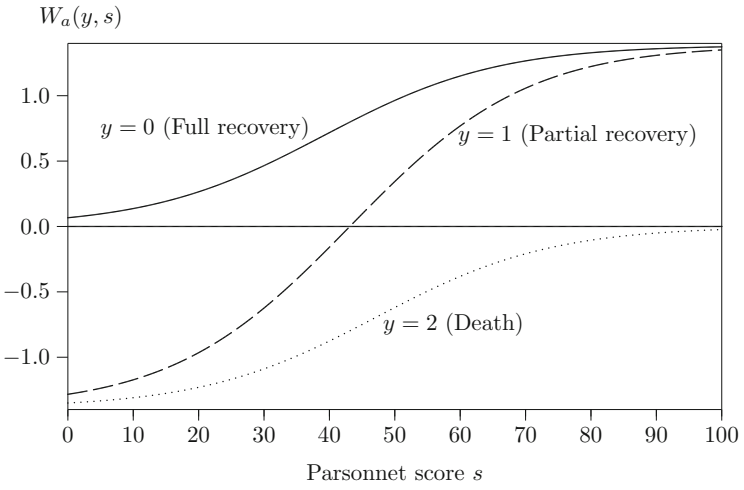


Fig. 1 Plots of $W_a(y, s)$ against the Parsonnet score s for $y = 0, 1, 2$ when $J = 2$, $R^+ = 2$ and $R^- = 0.5$

a reward and $W_a(J, s)$ is always a penalty. The score $W_a(k, s), k = 1, \ldots, J-1$ can be viewed either as a penalty or a reward depending on the Parsonnet score of a patient and the state of partial recovery. Furthermore, it can be seen from result 4 of Theorem 2 that for a very high risk patient who makes a partial recovery, the reward given is very close to that of a full recovery. This means that any state of partial recovery is considered almost as good as a full recovery for a very high risk patient. Similarly, result 5 of Theorem 2 implies that for a very low risk patient, the penalty given for any partial recovery is very close to that of a patient who dies. This means that any state of partial recovery is considered almost as bad as dead for a very low risk patient. For a very low-risk patient, the only desirable outcome is a full recovery.

3 Risk-Adjusted Exponentially Weighted Moving Average Charting Procedure

Suppose X_n is the monitoring statistic based on the nth sample obtained. Let the mean and variance of X_n be μ and σ^2 respectively. The EWMA chart is obtained by plotting

$$Z_n = (1-\lambda)Z_{n-1} + \lambda X_n,$$

against the sample number n where λ is a smoothing parameter such that $0 < \lambda \leq 1$. The starting value Z_0 is usually taken to be $Z_0 = \mu$. The statistic Z_n can also be expressed as

$$Z_n = \lambda \sum_{i=0}^{n-1} (1-\lambda)^i X_{n-i} + (1-\lambda)^n Z_0.$$

It can be shown that if $Z_0 = \mu$, then $E(Z_n) = \mu$. The variance of Z_n can be shown to be $Var(Z_n) = \sigma^2\lambda[1-(1-\lambda)^{2n}]/(2-\lambda)$ and hence the asymptotic variance of Z_n is given as $\sigma^2\lambda/(2-\lambda)$. The upper and lower control limits for the EWMA chart are typically set as

$$UCL = \mu + L_1\sqrt{\frac{\lambda}{2-\lambda}}\sigma = H,$$

and

$$LCL = \mu - L_2\sqrt{\frac{\lambda}{2-\lambda}}\sigma = h,$$

respectively where L_1 and L_2 are some constants. The constants L_1 and L_2 are usually chosen to achieve a specific in-control average run length (ARL). If the risk distribution can be determined, the ARL can be approximated using the collocation procedure presented by Knoth (2005) based on the integral equation derived by Crowder (1987). The details are described in Appendix 3.

We can now summarize the procedure of constructing a risk-adjusted EWMA chart for monitoring surgical performances.

Step 1. Fit a proportional odds logistic regression model (1) using some past surgical data to estimate the probabilities of various outcomes $\pi_k(s)$, $k = 0, \ldots, J$, given a Parsonnet score s.

Step 2. Set the alternative hypothesis $H^+ : R_0 = R_1 = \cdots = R_{J-1} = R^+ > 1$ for detecting improvement and the alternative hypothesis $H^- : R_0 = R_1 = \cdots = R_{J-1} = R^- < 1$ for detecting deterioration. The probabilities of various outcomes $\pi_k^*(s)$, $k = 0, \ldots, J$, given a Parsonnet score s, assuming the odds ratios $R_0, R_1, \ldots, R_{J-1}$ for a surgeon can be determined using Eq. (3). The penalty-reward score $W_a(y, s)$ can then be calculated using Eq. (5).

Step 3. Plot $Z_n = \lambda W_a(S_n, Y_n) + (1 - \lambda)Z_{n-1}$ against n and signal if $Z_n > H$ or $Z_n < h$.

4 Evaluation of the Performances of Three Surgeons

In this section, we will construct risk-adjusted EWMA and CUSUM charts of three surgeons and compare their performances. These three surgeons are among a group of seven surgeons who performed heart bypass operations on 6449 patients. A patient is considered to have died ($Y = 2$) if the patient dies within 30 days of the surgery. A patient is considered to have a partial recovery ($Y = 1$) if the patient survives more than 30 days but died later before the study concluded. A patient who survives throughout the entire period of study is considered a full recovery ($Y = 0$). Our classification of the three outcomes is only approximate and quite likely not the best possible classification. A surgeon should be able make a more appropriate classification. For the three-outcome data, we first fit a proportional odds logistic regression model using the data set as

$$\begin{aligned} \log\left[\frac{\pi_0(s)}{\pi_1(s) + \pi_2(s)}\right] &= \alpha_0 + \beta s, \\ \log\left[\frac{\pi_0(s) + \pi_1(s)}{\pi_2(s)}\right] &= \alpha_1 + \beta s, \end{aligned} \tag{6}$$

where $\alpha_0 = 3.057$, $\alpha_1 = 3.691$ and $\beta = -0.078$. A score test performed for the proportional odds assumption gives a p-value of 0.36 which is not significant, thus it is reasonable to use the proportional odds logistic regression model. The

probabilities of the three outcomes can be obtained using Eq. (6) as $\pi_0(s) = \exp(3.057-0.078s)/[1+\exp(3.057-0.078s)], \pi_2(s) = 1/[1+\exp(3.691-0.078s)]$, $\pi_1(s) = 1 - \pi_0(s) - \pi_2(s)$. These probabilities assume the average performance of surgeons in the entire data set. For a surgeon whose performance is characterized by R_0 and R_1, these probabilities can be calculated using Eq. (3). Note that if the risk distribution of a surgeon is different from this data set, the in-control run length performance will be affected. The simulation-based method developed by Zhang and Woodall (2015) can be used to ensure a desired ARL.

We will highlight the performances of three surgeons. The risk-adjusted EWMA charts constructed for surgeons A, B and C are displayed in Figs. 2, 3 and 4 respectively. The smoothing constant λ for these charts is set to be 0.01. A very small λ is used here because of the large variability of the penalty-reward score. The large variability is natural for this type of data. The chart limits are chosen such that the in-control ARL is about 100. Unlike an industrial process, the in-control ARL for this application should ideally be chosen to be small so that it will signal earlier should there be any deterioration in surgical performance. Surgeon A operated on 986 patients. Figure 2 shows that his performance remained stable for about the first 700 patients and then started to improve steadily after that. Surgeon B operated on 1654 patients. Figure 3 shows that his performance deteriorated for approximately the first 550 patients but turned around after that and continued to improve for the rest of the patients. Surgeon C operated on 568 patients. Figure 4 shows that his performance is stable for approximately the first half of the patients but deteriorated for the rest of the patients.

The risk-adjusted CUSUM charts for the three surgeons are displayed in Figs. 5, 6 and 7 respectively. The upper-sided CUSUM chart is designed to be

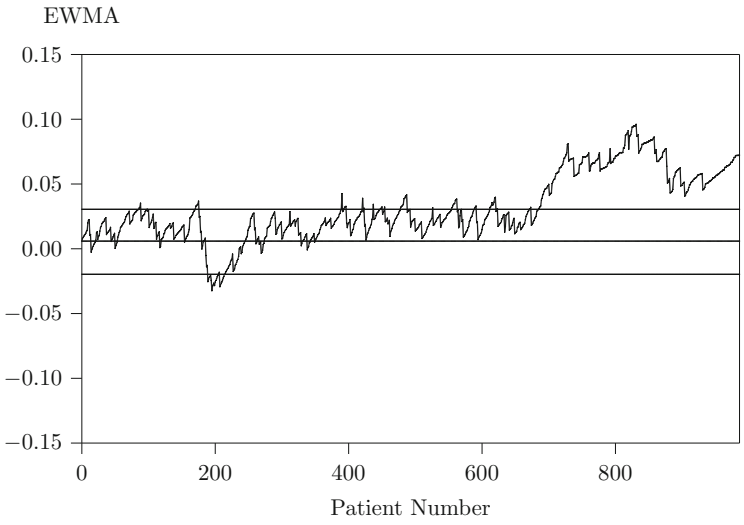


Fig. 2 Plot of risk-adjusted EWMA chart for Surgeon A

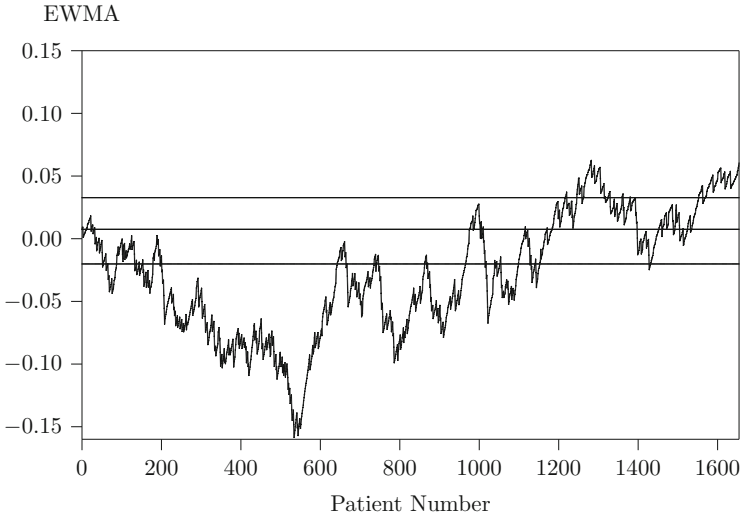


Fig. 3 Plot of risk-adjusted EWMA chart for Surgeon B

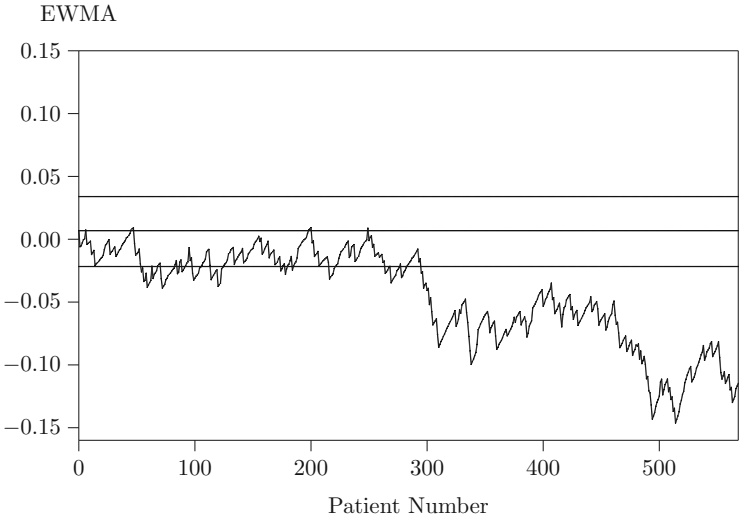


Fig. 4 Plot of risk-adjusted EWMA chart for Surgeon C

optimal in detecting $R = 2$ and the lower-sided CUSUM chart is designed to be optimal in detecting $R = 0.5$. The inferences drawn from these CUSUM charts are similar to those drawn from the EWMA charts. The EWMA chart has the advantage of ease of interpretation.

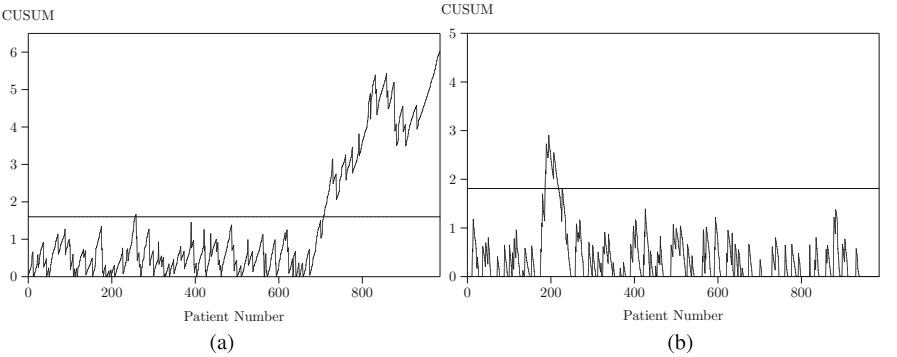


Fig. 5 Risk-adjusted CUSUM charts for detecting improvement ($R = 2$) and deterioration ($R = 0.5$) for Surgeon A. (**a**) Detecting improvement ($R = 2$). (**b**) Detecting deterioration ($R = 0.5$)

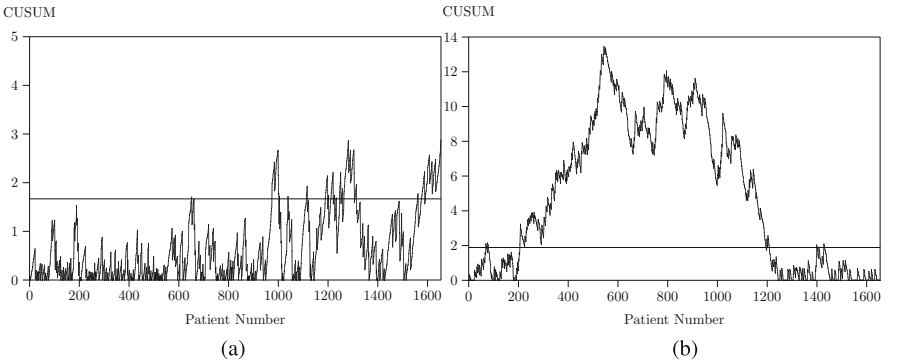


Fig. 6 Risk-adjusted CUSUM charts for detecting improvement ($R = 2$) and deterioration ($R = 0.5$) for Surgeon B. (**a**) Detecting improvement ($R = 2$). (**b**) Detecting deterioration ($R = 0.5$)

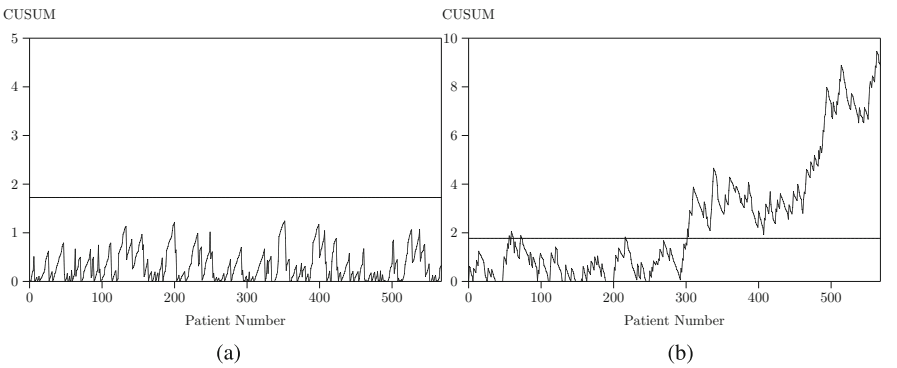


Fig. 7 Risk-adjusted CUSUM charts for detecting improvement ($R = 2$) and deterioration ($R = 0.5$) for Surgeon C. (**a**) Detecting improvement ($R = 2$). (**b**) Detecting deterioration ($R = 0.5$)

5 Conclusions

Steiner et al. (2000) developed a risk-adjusted CUSUM charting procedure for monitoring surgical performances based on binary outcomes: survival or death. However, for a patient who survives an operation, there can be many different grades of survival. In order to improve the effectiveness of the CUSUM procedure, Tang et al. (2015) developed a risk-adjusted CUSUM procedure based on three or more outcomes. The EWMA procedure is known to have run length properties similar to the CUSUM procedure but with the advantage of ease of interpretation. In this chapter, we develop a risk-adjusted EWMA procedure based on two or more outcomes. The monitoring statistic is a statistic obtained by combining the log likelihood ratio statistics for detecting improvement and deterioration. The properties of this statistic are studied and conditions are established to ensure that there is a proper ordering according to the severities of surgical outcomes. We compare the performances of these two competing charting procedures by analysing three surgeons' surgical data. A more comprehensive comparison can be done using ARL. The performances of the two procedures were found to be similar. The EWMA procedure is an attractive alternative with the advantage of ease of interpretation.

Acknowledgements The first author is supported by Academic Research Fund Tier 1 (R-155-000-159-112), Ministry of Education, Singapore. We are grateful to Dr Zhang Lingyun and Dr Stefan Steiner for the data set.

Appendix 1: Proof of Theorem 1

To prove sufficiency, let $R_0^+/R_0^- = \cdots = R_{J-1}^+/R_{J-1}^- = R^+/R^- \geq 1$. From the proportional odds logistic regression model

$$\text{logit}[P(Y \leq k|S = s)] = \alpha_k + \beta s, k = 0, \ldots, J-1,$$

we can obtain the conditional probability

$$\begin{aligned}\pi_k(s) &= P(Y \leq k|S = s) - P(Y \leq k-1|S = s)\\ &= \frac{\exp(\alpha_k + \beta s)}{1 + \exp(\alpha_k + \beta s)} - \frac{\exp(\alpha_{k-1} + \beta s)}{1 + \exp(\alpha_{k-1} + \beta s)}\\ &= \frac{\exp(\beta s)[\exp(\alpha_k) - \exp(\alpha_{k-1})]}{[1 + \exp(\alpha_k + \beta s)][1 + \exp(\alpha_{k-1} + \beta s)]},\end{aligned}$$

where $k = 0, \cdots, J$, $\alpha_{-1} = -\infty$ and $\alpha_J = \infty$. From Eq. (3), we have

$$\log\left(\sum_{i=0}^{k} \pi_i^+(s)/\left[1 - \sum_{i=0}^{k} \pi_i^+(s)\right]\right) = \log(R_k^+) + \alpha_k + \beta s.$$

Then, we have

$$\pi_k^+(s) = \frac{\exp(\beta s)[\exp(\alpha_k + \log(R_k^+)) - \exp(\alpha_{k-1} + \log(R_{k-1}^+))]}{[1 + \exp(\alpha_k + \log(R_k^+) + \beta s)][1 + \exp(\alpha_{k-1} + \log(R_{k-1}^+) + \beta s)]}$$

where $k = 0, \cdots, J$, and $R_{-1}^+ = R_J^+ = 1$. Similarly,

$$\pi_k^-(s) = \frac{\exp(\beta s)[\exp(\alpha_k + \log(R_k^-)) - \exp(\alpha_{k-1} + \log(R_{k-1}^-))]}{[1 + \exp(\alpha_k + \log(R_k^-) + \beta s)][1 + \exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)]}$$

where $k = 0, \cdots, J$, and $R_{-1}^- = R_J^- = 1$. Hence,

$$\begin{aligned} W_a(k, s) &= \log[\pi_k^+(s)/\pi_k^-(s)] \\ &= \mathrm{D}_k + \log(1 + \exp(\alpha_k + \log(R_k^-) + \beta s)) + \log(1 + \exp(\alpha_{k-1} \\ &\quad + \log(R_{k-1}^-) + \beta s)) - \log(1 + \exp(\alpha_k + \log(R_k^+) + \beta s)) \\ &\quad - \log[1 + \exp(\alpha_{k-1} + \log(R_{k-1}^+) + \beta s)], \end{aligned}$$

where

$$\begin{aligned} D_k = &\ \log[\exp(\alpha_k + \log(R_k^+)) - \exp(\alpha_{k-1} + \log(R_{k-1}^+))] \\ &- \log[\exp(\alpha_k + \log(R_k^-)) - \exp(\alpha_{k-1} + \log(R_{k-1}^-))] = \log(R^+/R^-). \end{aligned}$$

Taking the first derivative with respect to s, we obtain

$$\begin{aligned} \frac{\partial W_a(k, s)}{\partial s} &= \beta\Big[\frac{\exp(\alpha_k + \log(R_k^-) + \beta s)}{1 + \exp(\alpha_k + \log(R_k^-) + \beta s)} + \frac{\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)}{1 + \exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)} \\ &\quad - \frac{\exp(\alpha_k + \log(R_k^+) + \beta s)}{1 + \exp(\alpha_k + \log(R_k^+) + \beta s)} - \frac{\exp(\alpha_{k-1} + \log(R_{k-1}^+) + \beta s)}{1 + \exp(\alpha_{k-1} + \log(R_{k-1}^+) + \beta s)}\Big] \\ &= \beta\Big[\frac{1}{1 + \exp(\alpha_k + \log(R_k^+) + \beta s)} + \frac{1}{1 + \exp(\alpha_{k-1} + \log(R_{k-1}^+) + \beta s)} \\ &\quad - \frac{1}{1 + \exp(\alpha_k + \log(R_k^-) + \beta s)} - \frac{1}{1 + \exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)}\Big] \\ &= \beta E, \end{aligned}$$

where

$$E = \frac{1}{1+\exp(\alpha_k + \log(R_k^+) + \beta s)} - \frac{1}{1+\exp(\alpha_k + \log(R_k^-) + \beta s)} + \frac{1}{1+\exp(\alpha_{k-1} + \log(R_{k-1}^+) + \beta s)} - \frac{1}{1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)}.$$

Note that $R_0^+/R_0^- = \cdots = R_{J-1}^+/R_{J-1}^- = R^+/R^- \geq 1$, this implies $E \leq 0$. In addition, note that $\beta < 0$ from earlier discussion. Thus, $\partial W_a(k,s)/\partial s \geq 0$. It follows that $W_a(y,s)$ is an increasing function of s given y.

In addition, let $\Delta = \log(R^+/R^-) \geq 0$. Define $g_k(\Delta) = W_a(k+1,s) - W_a(k,s)$. Then

$$\begin{aligned} g_k(\Delta) = &\log(1+\exp(\alpha_{k+1} + \log(R_{k+1}^-) + \beta s)) \\ &-\log(1+\exp(\alpha_{k+1} + \log(R_{k+1}^-) + \Delta + \beta s)) \\ &-\log(1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)) \\ &+\log(1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \Delta + \beta s)). \end{aligned}$$

It is clear that $g_k(0) = 0$. Take the first derivative of $g_k(\Delta)$,

$$\begin{aligned} g_k'(\Delta) = &-\frac{\exp(\alpha_{k+1} + \log(R_{k+1}^-) + \Delta + \beta s)}{1+\exp(\alpha_{k+1} + \log(R_{k+1}^-) + \Delta + \beta s)} \\ &+\frac{\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \Delta + \beta s)}{1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \Delta + \beta s)} \\ = &\frac{1}{1+\exp(\alpha_{k+1} + \log(R_{k+1}^-) + \Delta + \beta s)} \\ &-\frac{1}{1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \Delta + \beta s)}. \end{aligned}$$

Note that $\alpha_{k+1} + \log(R_{k+1}^-) \geq \alpha_{k-1} + \log(R_{k-1}^-)$, thus $g_k'(\Delta) \leq 0$. Hence, $g_k(\Delta) \leq 0$ for $\Delta \geq 0$. Thus, $W_a(k+1,s) \leq W_a(k,s)$. In other words, $W_a(y,s)$ is a decreasing function of y conditional on s. This proves the sufficiency.

Note that $\pi_Y^+(S)$ and $\pi_Y^-(S)$ are defined as:

$$\frac{\sum_{i=0}^{k} \pi_i^+(s)}{1-\sum_{i=0}^{k} \pi_i^+(s)} = R_k^+ \frac{\sum_{i=0}^{k} \pi_i(s)}{1-\sum_{i=0}^{k} \pi_i(s)},$$

and

$$\frac{\sum_{i=0}^{k}\pi_i^-(s)}{1-\sum_{i=0}^{k}\pi_i^-(s)} = R_k^- \frac{\sum_{i=0}^{k}\pi_i(s)}{1-\sum_{i=0}^{k}\pi_i(s)}.$$

It follows that

$$\frac{\sum_{i=0}^{k}\pi_i^+(s)}{1-\sum_{i=0}^{k}\pi_i^+(s)} = \frac{R_k^+}{R_k^-} \frac{\sum_{i=0}^{k}\pi_i^-(s)}{1-\sum_{i=0}^{k}\pi_i^-(s)}. \tag{7}$$

In other words, the odds ratios of $\pi_Y^+(S)$ related to $\pi_Y^-(S)$ is given by R_k^+/R_k^-.

To prove necessity, assume $W_a(y, s)$ is a decreasing function of y conditional on s: $W_a(0, s) \geq W_a(1, s) \geq \cdots \geq W_a(J, s)$. Equivalently,

$$\frac{\pi_0^+(s)}{\pi_0^-(s)} \geq \frac{\pi_1^+(s)}{\pi_1^-(s)} \geq \cdots \geq \frac{\pi_J^+(s)}{\pi_J^-(s)}.$$

Then, we have

$$\frac{\pi_0^+(s)}{\pi_0^-(s)} \geq \frac{\pi_0^+(s)+\pi_1^+(s)}{\pi_0^-(s)+\pi_1^-(s)} \geq \cdots \geq \frac{\sum_{i=0}^{J-1}\pi_i^+(s)}{\sum_{i=0}^{J-1}\pi_i^-(s)} \geq \frac{\sum_{i=0}^{J}\pi_i^+(s)}{\sum_{i=0}^{J}\pi_i^-(s)} = 1, \tag{8}$$

$$\frac{\pi_J^+(s)}{\pi_J^-(s)} \leq \frac{\pi_{J-1}^+(s)+\pi_J^+}{\pi_{J-1}^-(s)+\pi_J^-(s)} \leq \cdots \leq \frac{\sum_{i=1}^{J}\pi_i^+(s)}{\sum_{i=1}^{J}\pi_i^-(s)} \leq \frac{\sum_{i=0}^{J}\pi_i^+(s)}{\sum_{i=0}^{J}\pi_i^-(s)} = 1. \tag{9}$$

Based on the odds ratio of cumulative probabilities defined in Eq. (7), we obtain

$$\frac{\sum_{i=0}^{k}\pi_i^+(s)}{\sum_{i=0}^{k}\pi_i^-(s)} = \frac{R_k^+/R_k^-}{1-\sum_{i=0}^{k}\pi_i^-(s)+R_k^+/R_k^-\sum_{i=0}^{k}\pi_i^-(s)}, \ k = 0, \cdots, J-1, \tag{10}$$

$$\frac{\sum_{i=k+1}^{J}\pi_i^+(s)}{\sum_{i=k+1}^{J}\pi_i^-(s)} = \frac{1}{1-\sum_{i=0}^{k}\pi_i^-(s)+R_k^+/R_k^-\sum_{i=0}^{k}\pi_i^-(s)}, \ k = 0, \cdots, J-1. \tag{11}$$

Substitute (10) into (8) and (11) into (9), we get

$$\frac{R_0^+/R_0^-}{1-\pi_0^-(s)+R_0^+/R_0^-\cdot\pi_0^-(s)} \geq \frac{R_1^+/R_1^-}{1-\sum_{i=0}^{1}\pi_i^-(s)+R_1^+/R_1^-\sum_{i=0}^{1}\pi_i^-(s)} \geq \cdots$$

$$\geq \frac{R_{J-1}^+/R_{J-1}^-}{1-\sum_{i=0}^{J-1}\pi_i^-(s)+R_{J-1}^+/R_{J-1}^-\sum_{i=0}^{J-1}\pi_i^-(s)} \geq 1, \tag{12}$$

and

$$\frac{1}{1-\sum_{i=0}^{J-1}\pi_i^-(s)+R_{J-1}^+/R_{J-1}^-\sum_{i=0}^{J-1}\pi_i^-(s)} \leq \frac{1}{1-\sum_{i=0}^{J-2}\pi_i^-(s)+R_{J-2}^+/R_{J-2}^-\sum_{i=0}^{J-2}\pi_i^-(s)}$$

$$\leq \cdots \leq \frac{1}{1-\pi_0^-(s)+R_0^+/R_0^-\cdot\pi_0^-(s)} \leq 1. \tag{13}$$

From the definition of risk score, if $s \to \infty$, $\pi_J^-(s) \to 1$, thus $\sum_{i=0}^{k}\pi_i^-(s) \to 0$ for $k = 0, \cdots, J-1$ and we obtain the following from Eq. (12)

$$R_0^+/R_0^- \geq R_1^+/R_1^- \geq \cdots \geq R_{J-1}^+/R_{J-1}^- \geq 1. \tag{14}$$

Similarly, if $s \to -\infty$, $\pi_0^-(s) \to 1$, thus $\sum_{i=0}^{k}\pi_i^-(s) \to 1$ for $k = 0, \cdots, J-1$ and we obtain the following from Eq. (13)

$$1 \geq R_0^-/R_0^+ \geq R_1^-/R_1^+ \geq \cdots \geq R_{J-1}^-/R_{J-1}^+. \tag{15}$$

(14) and (15) imply

$$R_0^+/R_0^- = R_1^+/R_1^- = \cdots = R_{J-1}^+/R_{J-1}^- \geq 1.$$

Appendix 2: Proof of Theorem 2

Let $\Delta = \log(R^+/R^-)$. Note that $\Delta > 0$.

$$\begin{aligned} W_a(k,s) = \Delta &+ \log(1+\exp(\alpha_k + \log(R_k^-) + \beta s)) \\ &+ \log(1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)) \\ &- \log(1+\exp(\alpha_k + \log(R_k^+) + \beta s)) \\ &- \log[1+\exp(\alpha_{k-1} + \log(R_{k-1}^+) + \beta s)]. \end{aligned}$$

1. For $Y = 0$ and $\alpha_{-1} = -\infty$, then

$$\begin{aligned} W_a(0, s) &= \Delta + \log(1 + \exp(\alpha_0 + \log(R_0^-) + \beta s)) \\ &\quad -\log(1 + \exp(\alpha_0 + \log(R_0^-) + \Delta + \beta s)) \\ &= \log(1 + \exp(\alpha_0 + \log(R_0^-) + \beta s)) \\ &\quad -\log(\exp(-\Delta) + \exp(\alpha_0 + \log(R_0^-) + \beta s)). \end{aligned}$$

   Since $\Delta > 0$, $\exp(-\Delta) < 1$ and hence $W_a(0, s) > 0$.
2. For $Y = J$ and $\alpha_J = \infty$, then

$$\begin{aligned} W_a(J, s) &= \log(1 + \exp(\alpha_{J-1} + \log(R_{J-1}^-) + \beta s)) \\ &\quad -\log(1 + \exp(\alpha_{J-1} + \log(R_{J-1}^-) + \Delta + \beta s)) \end{aligned}$$

   Since $\Delta > 0$, $W_a(J, s) < 0$.
3. This is clear from the functions of $W_a(0, s)$ and $W_a(J, s)$ given in parts 1 and 2.

4 and 5. For $Y = k \in \{1, \cdots, J-1\}$,

$$\begin{aligned} W_a(k, s) &= \Delta + \log(1 + \exp(\alpha_k + \log(R_k^-) + \beta s)) \\ &\quad -\log(1 + \exp(\alpha_k + \log(R_k^-) + \Delta + \beta s)) \\ &\quad -\log(1 + \exp(\alpha_{k-1} + \log(R_{k-1}^-) + \Delta + \beta s)). \end{aligned}$$

Note that $\beta < 0$ for our logistic model. Let $s \to \infty$, then $W_a(k, s) \to \Delta < 0$. Let $s \to -\infty$, then $W_a(k, s) \to -\Delta > 0$.

For $Y = 0$, from the function $W_a(0, s)$ obtained in part 1, let $s \to \infty$, then $W_a(0, s) \to \Delta$. For $Y = J$, from the function $W_a(J, s)$ obtained in part 2, we can show that

$$\begin{aligned} W_a(J, s) &= \log(1 + \exp(\alpha_{J-1} + \log(R_{k-1}^-) + \beta s)) \\ &\quad -\log(1 + \exp(\alpha_{J-1} + \log(R_{k-1}^-) + \Delta + \beta s)) \\ &= -\Delta + \log(1 + \exp(\alpha_{J-1} + \log(R_{k-1}^-) + \beta s)) \\ &\quad -\log(\exp(-\Delta) + \exp(\alpha_{J-1} + \log(R_{k-1}^-) + \beta s)). \end{aligned}$$

Let $s \to -\infty$, then $W_a(J, s) \to -\Delta > 0$.

Appendix 3: Average Run Length of EWMA Chart

Page (1954) introduced an integral equation method for evaluating the ARL of a CUSUM chart, and Crowder (1987) derived a similar integral equation for the EWMA chart. Let $L(u)$ denote the ARL of the EWMA chart that starts at $Z_0 = u$, then the integral equation for the ARL can be shown as

$$L(u) = 1 + \frac{1}{\lambda}\int_h^H L(x) f_a\Big(\frac{x-(1-\lambda)u}{\lambda}\Big)dx,$$

where $f_a(\cdot)$ is the pdf of $W_a(Y, S)$. The function $L(u)$ can be approximated numerically by using the collocation method (Knoth 2005).

References

BRI Inquiry Panel (2001). Learning from Bistol: The report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984–1995. London: The Stationery Office. Available from https://www.bristol-inquiry.org.uk/final_report.

Crowder, S. V. (1987). A simple method for studying run-length distributions of exponentially weighted moving average charts. *Technometrics, 29*, 401–407.

Gan, F. F, Lin, L., & Loke, C. K. (2012). Risk-adjusted cumulative sum charting procedures. In H. J. Lenz, P. Th. Wilrich, & W. Schmid (Eds.), *Frontiers in statistical quality control* (Vol. 10, pp. 207–225). Berlin: Springer.

Grigg, O., & Spiegelhalter, D. (2007). A simple risk-adjusted exponentially weighted moving average. *Journal of the American Statistical Association, 102*, 140–152.

Knoth, S. (2005). Accurate ARL computation for EWMA-S^2 control charts. *Statistics and Computing, 15*, 341–352.

Lovegrove, J., Sherlaw-Johnson, C., Valencia, O., Treasure, T., & Gallivan, S. (1999). Monitoring the performance of cardiac surgeons. *The Journal of the Operational Research Society, 50*, 684–689.

Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C., & Gallivan, S. (1997). Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet, 350*, 1128–1130.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, 42*, 109–142.

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

Parsonnet, V., Dean, D., & Bernstein, A. D. (1989). A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation, 79*, I-3–I-12.

Poloniecki, J., Valencia, O., & Littlejohns, P. (1998). Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *British Medical Journal, 315*, 1697–1700.

Roques, F., Nshef, S. A., Michel, P., Gauducheau, E., de Vincentiis, C., Baudet, E., et al. (1999). Risk factors and outcome in European cardiac surgery: Analysis of the EuroSCORE multinational database of 19,030 patients. *European Journal of Cardio-Thoracic Surgery, 15*, 816–822; discussion 822–823.

Steiner, S. H., Cook, R. J., Farewell, V. T., & Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics, 1*(4), 441–452.

Steiner, S. H., & Jones, M. (2010). Risk adjusted survival time monitoring with an updating exponentially weighted moving average (EWMA) control chart. *Statistics in Medicine, 29*, 444–454.

Tang, X., Gan, F. F., & Zhang, L. Y. (2015). Risk-adjusted cumulative sum charting procedure based on multiple responses. *Journal of the American Statistical Association, 110*, 16–26.

Treasure, T., Gallivan, S., & Sherlaw-Johnson, C. (2004). Monitoring cardiac surgical performance: a commentary. *The Journal of Thoracic and Cardiovascular Surgery, 128*, 823–825.

Treasure, T., Taylor, K., & Black, N. (1997). *Independent Review of Adult Cardiac Surgery-United Bristol*. Bristol: Health Care Trust, March.

Waldie, P. (1998). Crisis in the cardiac unit. *The Globe and Mail*, Canada's National Newspaper, Oct. 27; Sect. A:3(col. 1).

Woodall, W. H., Fogel, S. L., & Steiner, S. H. (2015). The monitoring and improvement of surgical outcome quality. *Journal of Quality Technology, 47*, 383–399.

Zhang, X., & Woodall, W. H. (2015). Dynamic probability control limits for risk adjusted Bernoulli CUSUM charts. *Statistics in Medicine, 34*, 3336–3348.

A Primer on SPC and Web Data

Erwin Saniga, Darwin Davis, and James M. Lucas

Abstract In this chapter we compare the website visitor data generated by a variety of commercially available analytics packages and discuss issues of data accuracy, consistency and unavailability of some important measures. We also discuss some common and perhaps new SPC methods for monitoring website effectiveness using this data.

Keywords SPC · Web Data · Markov Chains

1 Introduction

In this chapter we investigate the use of statistical process control tools in monitoring web site visitor data generated by a variety of commercially available analytics packages. In doing this study we implemented several analytics packages on two web sites currently in use. One has less than one hundred visitors per month while the other has several thousand visitors per month.

We find there may be issues with data quality on particular analytics software and outline possible reasons for this shortcoming. We provide a comparative table of the software we employ based upon various characteristics that may be necessary to provide information required to employ particular SPC monitoring tools. We also show that useful information may be difficult to obtain on the analytics software we employ. While our investigation is limited to a few popular analytics software packages, some general conclusions may be drawn.

E. Saniga (✉) · D. Davis
Department of Business Administration, Alfred Lerner College of Business & Economics, University of Delaware, Newark, DE, USA
e-mail: saniga@udel.edu; dd@udel.edu

J. M. Lucas
James Lucas and Associates, Newark, NJ, USA
e-mail: James.Lucas@verizon.net

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_7

Given the data available, some common process monitoring tools are described. In addition, we investigate the use of Markov chains as a model for the flow of a visitor through the website and discuss the value of monitoring this Markov chain for changes over time or for determining the effectiveness of website interventions. We also discuss a statistical tool for monitoring this variable.

2 The Study

We implemented several popular analytics packages on two websites. The first was a personal site experiencing less than 100 visits per month. The second was a subset of a large commercial site experiencing more than 30,000 visits per month.

Our first interest in this study is to compare the results of the various analytics software in terms of the accuracy of their reporting of the actual number of visitors and the path they traversed through the site.

Table 1 compares the results of these counts for various analytics software, for the small website, for a 4 week period.

Note that there are substantial differences in the count data of visitors between the four analytics software packages. Comparing total users (new + returning) over the 7 week period shows an average of 24.89 visits per week (new plus returning users) across all software but averages of 26.85, 24.57, 22 and 25.5 for the four respective sites Google Analytics,[1] w3 counter,[2] statcounter[3] and WP SLimstat[4] Analytics.

We implemented three software analytics packages on the large commercial site (Google Analytics, Clicky[5] and Statcounter) over a 3 month period and obtained the results depicted in Table 2. Note again the disparity between the numbers of visitors reported by the three software packages. If we apply a c chart to this data we can find UCL $= 36,758$ and LCL $= 35,617$. One can see that two of the three outcomes are outside the three sigma control limits. (We calculated the average count as $\bar{c} = (35291 + 36676 + 35967)/3$. Recall that for a c chart, the variance is $\bar{c}$. We use three-sigma limits.)

What are the reasons for this disparity? One might be where the tracking code is installed on the site. One might investigate whether it is installed correctly. For example, if it is installed near the bottom of the page of HTML code and the page does not completely load then that particular visit may not be logged. On the other hand if it is installed in the top of the page and the page does not completely load then that may be counted as a visit.

[1] www.google.com/analytics/.
[2] https://www.w3counter.com/.
[3] https://statcounter.com/.
[4] http://www.wp-slimstat.com/.
[5] https://clicky.com/.

Table 1 Visits to a small website as recorded by various analytics packages

		Google analytics	w3counter	Statscounter	WPSlimstat analytics
March 13–20	New users	9	11	11	Installed on 21st March
	Returning users	7	0	3	Installed on 21st March
March 21–27	New users	26	28	18	15
	Returning users	4	0	5	1
March 28–April 3	New users	17	16	13	11
	Returning users	3	0	3	2
April 4–10	New users	40	29	30	5
	Returning users	0	7	9	39
April 11–17	New users	36	23	16	6
	Returning users	3	3	3	28
April 18–24	New users	12	21	13	20
	Returning users	6	3	6	3
April 25–May 1	New users	24	30	23	22
	Returning users	1	1	1	1

Table 2 Visits to a large commercial website as recorded by various analytics packages

	Google analytics	Clicky	StatCounter
Sessions/Visits April 14–May 5, 2016	35,291	36,676	35,967

A second reason for disparity might be that there are IP (Internet Protocol) addresses being blocked for bots (automated computer programs that enter the site). For example Google Analytics filters out "known" bot traffic by default. It is a complex process that is described by Sharif (2014). Further reading on the issue of bots is described by Zeifman (2015).

A practical solution to this problem is to host one's website on an internal server. Then one can run one's own counter of visits and other desired measures to ensure that tracking was accurate. On one of our sites this code was implemented along with Google Analytics and it was found that the latter software reported roughly 30–40% less traffic than the internal server logs would report. Nonetheless, the problem remains messy. As Chimphlee et al. (2006, p. 372) note: "Web log files contain a large amount of erroneous, misleading and incomplete information", and they recommend the elimination of items that are not requested by the user, in particular, graphics.

One interest in this study is to analyze the capability of the analytics packages in terms of their ability to provide the data in a form one can use in the process control analyses one might find effective in monitoring websites.

One analysis we discuss is the use of a Markov Chain model to model the flow of a visitor through the website. To build this model one needs to generate the flow for each user and combine these for all users for a particular time period. Table 3 shows the transition matrices for several consecutive weeks for the small website. This data was generated by the inefficient method of individually tracking each user's flow (where each user is identified by their IP address) and combining these for each week's data, a time consuming process. Of the four software packages we investigated only Google Analytics and Statcounter enabled us to find this users flow through the website. Nevertheless, we found that there are some problems with the data obtained from Google Analytics.

One problem is that the users flow for the 5 week period as reported on the users flow link in Google analytics was reported as 100% drop off by visitors after they reached the home page of the small site. Since the data we report in Table 3 is generated by tabulating the flow of each user as identified by their IP address we can argue that Table 3 data is correct insofar as users flow is concerned. (Although Table 1 does show disparity between the count of visitors by software.)

On the large site we have found that the user flow data on Google Analytics does show visitor tracks throughout the website. We have observed, however, that this data is incomplete. For example, on the large site the transition counts from one page to the next are incomplete, being reported simply as some number of "other pages visited" and is exhaustive when listed.

Table 3 Raw Data and Transition Probability Matrix, respectively, for different time spans

(a) April 25 – May 1, 2016

Future State

Current State	A	B	C	D	E	F	G	H	I	J	Drop Off
A	0	6	4	3	2	3	5	1	1	2	11
B	0	0	2	1	0	0	0	2	0	11	0
C	2	1	0	0	7	0	0	0	0	12	1
D	1	1	1	0	0	0	0	0	0	6	0
E	1	2	1	1	0	0	1	0	0	14	0
F	0	0	1	0	0	0	0	0	0	0	2
G	2	0	1	1	0	0	0	0	1	0	1
H	0	0	0	0	1	0	0	0	0	3	0
I	1	0	0	0	0	0	0	0	0	0	1
J	9	5	14	2	8	0	0	1	0	0	9
Drop Off	0	0	0	0	0	0	0	0	0	0	1

(b) April 25 – May 1, 2016

Future State

Current State	A	B	C	D	E	F	G	H	I	J	Drop Off
A	0	0.158	0.105	0.079	0.053	0.079	0.132	0.026	0.026	0.053	0.289
B	0	0	0.125	0.063	0	0	0	0.125	0	0.688	0
C	0.087	0.043	0	0	0.304	0	0	0	0	0.522	0.043
D	0.111	0.111	0.111	0	0	0	0	0	0	0.667	0
E	0.050	0.100	0.050	0.050	0	0	0.050	0	0	0.700	0
F	0	0	0.333	0	0	0	0	0	0	0	0.667
G	0.333	0	0.167	0.167	0	0	0	0	0.167	0	0.167
H	0	0	0	0	0.250	0	0	0	0	0.750	0
I	0.500	0	0	0	0	0	0	0	0	0	0.500
J	0.188	0.104	0.292	0.042	0.167	0	0	0.021	0	0	0.188
Drop Off	0	0	0	0	0	0	0	0	0	0	1.000

(c) May 2 – May 8, 2016

Future State

Current State	A	B	C	D	E	F	G	H	I	J	Drop Off
A	0	0.111	0.111		0.222		0.222	0.111	0	0	0.222
B	0	0	0.143		0		0	0	0	0.857	0
C	0	0	0		0		0	0	0	1.000	0
D											
E	0	0	0		0		0	0	0	1.000	0
F											
G	0	0.500	0		0		0	0	0.500	0	0
H	0	0.667	0		0		0	0	0	0.333	0
I	0	0	0		1.000		0	0	0	0	0
J	0.095	0.143	0.190		0.286		0	0	0	0	0.286
Drop Off	0	0	0		0		0	0	0	0	1.000

= There was no activity on these pages so the row probabilites cannot be estimated.
A transition probability matrix can be created by removing the shaded rows and columns.

(d) May 9 – May 15, 2016

Future State

Current State	A	B	C	D	E	F	G	H	I	J	Drop Off
A	0	0.125	0.125	0.063	0.063	0.063	0.125			0.125	0.313
B	0	0	0	1.000	0	0	0			0	0
C	0.222	0	0	0	0	0	0			0.667	0.111
D	0	0	0.400	0	0	0	0			0.600	0
E	0	0	0	0	0	0	0			1.000	0
F	0.333	0	0	0	0	0	0.333			0	0.333
G	0.333	0	0.333	0.333	0	0	0			0	0
H											
I											
J	0.250	0	0.333	0.083	0	0.083	0			0	0.250
Drop Off	0	0	0	0	0	0	0			0	1.000

(e) May 16 – May 22, 2016

Future State

Current State	A	B	C	D	E	F	G	H	I	J	Drop Off
A	0	0.091	0.182	0.061	0.182	0.091	0	0.182	0.061	0	0.152
B	0	0	0.071	0	0.143	0	0	0	0	0.714	0.071
C	0	0.040	0	0.040	0.040	0	0	0	0	0.880	0
D	0	0.125	0	0	0	0	0	0	0	0.875	0
E	0	0.053	0.053	0	0	0.026	0.053	0	0	0.789	0.026
F	0	0	0	0.286	0	0	0.571	0	0.143	0	0
G	0	0	0.167	0	0	0.333	0	0	0	0.167	0.333
H	0	0.125	0.125	0	0.125	0	0	0	0	0.500	0.125
I	0	0	0	0	0	0.333	0	0	0	0.333	0.333
J	0.173	0.080	0.187	0.040	0.360	0	0	0.013	0	0	0.147
Drop Off	0	0	0	0	0	0	0	0	0	0	1.000

= There was no activity on these pages so the row probabilites cannot be estimated.
A transition probability matrix can be created by removing the shaded rows and columns.

In summary, we advise caution when using the data generated by the various software analytics packages. We have found a disparity between the results generated by these packages and in one case an incomplete reporting of the results. While we have identified some possible reasons why this disparity exists we wish to emphasize that in practice one might wish to implement their own analytics code and methodology to generate Web analytics data. Nonetheless, there is a data wrangling issue that must be addressed when employing data for SPC from the commercially available software we investigate.

3 Monitoring Web Data

Software such as Google Analytics presents many different measures of users actions on a particular Website. Consider as an example one of the common measures-new visitors to a Website. Many sites would find interest in monitoring this variable as it indicates significant shifts in the public's interest in their site The count of new visitors is a count variable and can be monitored using a c chart (see, e.g., Montgomery 2005) or a CUSUM chart for counts (see Lucas 1985).

Alternatively, variables such as bounce rate may be important when monitoring a commercial Website where purchases may be made. There, managers would be interested in the proportion of people that travel to a particular product page and "bounce" out before clicking on a purchase request. Obviously, a smaller bounce rate here would be preferred and additionally, monitoring this bounce rate over time would be advantageous as well. Another application would be to find if an intervention to improve bounce rate was effective. Bounce rate is measured by a proportion and thus can be monitored by a p chart or a binomial CUSUM chart. See, e.g. Montgomery (2005) or Hawkins and Olwell (1998).

In addition to signaling the occurrence of an event over time that is out of control or statistically significant in this context, we have found that the use of CUSUM plots for these discrete variables can be of importance in identifying regimes where lower or higher rates of counts or proportions occur. Saniga et al. (2009) illustrate the use of these plots in an actual business setting. This reliance on visual information is of great value in that long term regimes of higher or lower counts may be deemed important to the user even though these regimes are not significant. In addition, this visual presentation allows the communication of results to be done at a much higher level than reporting that a shift in a CUSUM chart is significant, say.

One interesting type of monitoring not usually addressed is the monitoring of the transition probability matrix of traffic through a site. Researchers have addressed the issue of modeling traffic using Markov chains but little has been done on monitoring these chains in an SPC sense. Some examples of modeling traffic research are the use of a Markov model to predict where a user will visit on the site given a sequence of pages the user has already visited. Chimphlee et al. (2006) summarize some of

this work and discuss prediction using higher order Markov models other than the usual first order model.

Marques and Belo (2011) use Markov Chains to help identify usage profiles (i.e., understand how users are using the web resources provided by teachers). They do not show a way to track changes in usage patterns or give statistical methods for determining changes in website effectiveness.

Huang et al. (2004) study the use of continuous time models requiring the estimation of both the transition probabilities and the expected transition rates, assuming the time spent in a state follows an exponential distribution. The focus of their research is building a model to make the following predictions:

- What page will a user visit next, and when will they transition to that page.
- The transition count from one web page to another.
- How many people visit a web page within some period of time.

They do not, however, provide any tools for tracking changes in web site performance.

Zhu et al. (2002a,b) use an m-order Markov Model (assumes the users next step is only dependent on the last m pages visited) to make link predictions that assist new users as they navigate an adaptive web site. These m-order models lead to very large, sparse transition matrices. A clustering algorithm is used to identify groups of web pages with similar transition behaviors, which is then used with a compression algorithm to create a smaller transition probability matrix that is denser that the original transition matrix.

A key difference between our focus and what we see in much of the above literature is as follows. Many articles are focused on prediction, such as which page will a visitor will go to next. Researchers have built models for such predictions, some based only on the page the user is now on, and some based on a longer history of pages visited by the user. Our focus is not on prediction, but on monitoring website quality/effectiveness. Tools developed for prediction do not seem to be of use for monitoring quality and signaling changes. An essential element of monitoring for quality/effectiveness is for the site owner to define the purpose of the site and how effectiveness can best be measured.

For example, in our focus on SPC, one can use a Phase I approach to determine the longer term average transition probabilities. These can be used to study and also predict typical user flow through a site and use marketing methods, say, to take advantage of this knowledge. One can use the method of Chatfield (1973) to test the suitability of a kth order Markov chain as an appropriate model which will aid in this process.

One can also use the resulting Markov chain in a Phase II sense. That is, it would be of value to monitor the typical users flow through a site to determine when change has occurred. This would be of value in Web redesign or in many other applications one can envision. Tests that would be valuable in this context are discussed by Anderson and Goodman (1957) who present methods to test if the transition probabilities of a first order Markov chain are constant and are specified numbers, and a test that the process is an *i*th order chain versus the alternative that it is a *j*th order chain. They also find maximum likelihood estimates of the transition probabilities. Agresti (2013) also presents inference methods for Markov chains.

For the small site on our study we present some weekly data illustrating the transition probability matrices derived from the Statcounter analytics package. These are presented for illustration purposes. In practice the determination of the sampling interval (here it is a week) would be an important decision that would have to be made in Phase I or Phase II studies. Generally, we would expect the sampling interval would be long if no interventions to the Website are made. If an intervention were to be made to redirect the flow of the users, the inference methods of Anderson and Goodman (1957) could be used to see if the intervention was effective in redirecting user flow through the system. Note that if one monitors a probability with a *p* chart, these charts may not be homogeneous when the process is in-control because of shorter and longer seasonal factors including time of day, day of week, week of month, month, or even longer seasons. Sparks (2017) has shown the efficacy of using adaptive EWMA charts to handle this problem.

Most well-designed websites have the purpose of ensuring that the user takes the quickest route to a designated page; e.g. a commercial site would like the customer to get to the "purchase" page as quickly as possible. One could measure departures from this route to measure the effectiveness of the design of the site and continuously monitor this over time to see if there is a need for redesign. Many different types of measures of user visits to a website are presented in the various analytics packages we tested in this chapter. A summary of some of the common ones are presented in Table 4, which presents the variable of interest, the measure of that variable, the type of control method recommended for monitoring, and the reference regarding design of that control method for the advanced user. Most of these are self-explanatory, except for the one labeled engagement, which is a frequency distribution of the number of sessions classified by session duration and the number of page views classified by session duration.

Table 4 Monitoring methods for web analytics data

Variable	Measure	Control method	Reference
New visitors	Count	CUSUM for counts	Lucas (1985)
Returning visitors	Count	CUSUM for counts	Lucas (1985)
Bounce rate	Count	CUSUM for counts	Lucas (1985)
Bounce rate	Proportion	CUSUM for binomial	Hawkins and Olwell (1998)
Users flow	Transition probability matrix	Markov chain	Anderson and Goodman (1957)
Country of origin	Multinomial	Multinomial	Topiladou and Psarakis (2009)
Sex of visitor	Proportion	Binomial CUSUM	Hawkins and Olwell (1998)
Duration of visit	Mean	CUSUM for a mean	Hawkins and Olwell (1998)
Engagement	Multinomial	Multinomial	Topiladou and Psarakis (2009)
Visit length	Multinomial	Multinomial	Topiladou and Psarakis (2009)
Visit length	Mean	CUSUM for a mean	Hawkins and Olwell (1998)
Browsers	Multinomial	Multinomial	Topiladou and Psarakis (2009)

4 Conclusions

We have employed several commercially available Web Analytics packages on two websites and presented some data representing user visits to these sites as well as users flow for one of the sites. Our observations are that some disparity exists between the data generated by this software and that a data wrangling issue does exist in this context.

We have also addressed the use of SPC tools for Phase I and II studies including the use of Markov chain models to monitor website effectiveness.

Acknowledgements The authors thank Adam Sexton, Digital Media Specialist at the University of Delaware for his contributions to this chapter. We thank the referee for helpful comments.

References

Agresti, A. (2013). *Categorical Data Analysis*. New York: Wiley.

Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics, 28*(1), 89–110.

Chatfield, C. (1973). Statistical inference regarding Markov chain models. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 22*(1), 7–20.

Chimphlee, S., Salim, N., Ngadiman, M., & Chimphlee, W. (2006). In T. Sabh, & K. Elleithy (Eds.), *Advances in Systems, Computing Sciences and Software Engineering* (pp. 371–376).

Hawkins, D., & Olwell, D. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer.

Huang, Q., Yang, Q., Huang, J. Z., & Ng, M. K. (2004). Mining of web-page visiting patterns with continuous-time markov models. In *Advances in Knowledge Discovery and Data Mining* (pp. 549–558). Berlin: Springer.

Lucas, J. (1985). Counted data cusums. *Technometrics, 27*(2), 129–144.

Marques, A., & Belo, O. (2011). Discovering student web usage profiles using Markov chains. *Electronic Journal of e-Learning, 9*(1), 63–74.

Montgomery, D. (2005). *Introduction to Statistical Quality Control*, 5th ed. New York: Wiley.

Saniga, E., Davis, D., & Lucas, J. (2009). Using Shewhart and CUSUM charts for diagnosis with count data in a vendor certification study. *Journal of Quality Technology, 41*(3), 217–227.

Sharif, S. (2014), Understanding bot and spider filtering from google analytics. http://www.lunametrics.com/blog/2014/08/07/bot-spider-filtering-google-analytics/.

Sparks, R. (2017). Linking EWMA p charts and risk adjustment control charts. *Quality and Reliablity Engineering International, 33*, 617–636.

Topiladou, E., & Psarakis, S. (2009). Review of multiattribute and multinomial control charts. *Quality and Reliability Engineering International, 25*, 773–809.

Zeifman, I. (2015). 2015 Bot Traffic Report: Humans Take Back the Web, Bad Bots Not Giving Any Ground. https://www.incapsula.com/blog/bot-traffic-report-2015.html

Zhu, J., Hong, J., & Hughes, J. G. (2002a). Using Markov chains for link prediction in adaptive web sites. In *Soft-Ware 2002: Computing in an Imperfect World* (pp. 60–73). Berlin: Springer.

Zhu, J., Hong, J., & Hughes, J. G. (2002b). Using Markov models for web site link prediction. In *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia* (pp. 169–170). New York: ACM.

The Variable-Dimension Approach in Multivariate SPC

Eugenio K. Epprecht, Francisco Aparisi, and Omar Ruiz

Abstract With multivariate processes, it may happen that some quality variables are more expensive and/or difficult to measure than the other ones, or they may demand much more time to measure. Their measurement may even be destructive. For monitoring such processes, the *variable dimension* approach was recently proposed. The idea is to measure always (at each sampling time) the "non-expensive" variables and to measure the expensive ones only when the values of the non-expensive variables give some level of evidence that the process may be out of control. The procedure bears much similarity with the one of *variable parameters* (or *adaptive*) control charts, but differs in that it is not the sample size or sampling interval or control limits that are made dynamically variable, but rather the very variables being measured (thus the name "variable dimension"). We review and compare the several variants of the approach, the last one being an EWMA version. The approach may lead to significant savings in sampling costs (the savings depending, of course, on the ratio between the costs of measuring the "expensive" and the "inexpensive" variables). Also, in many cases, contradicting the intuition, the variable dimension control charts may detect special causes even faster than their fixed (full) dimension counterparts.

E. K. Epprecht (✉)
Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: eke@puc-rio.br

F. Aparisi
Departamento de Estadística e I.O. Aplicadas y Calidad, Universidad Politécnica de Valencia, Valencia, Spain
e-mail: faparisi@eio.upv.es

O. Ruiz
ESPOL, Polytechnic University, Escuela Superior Politécnica del Litoral, Facultad de Ciencias de la Vida, Guayaquil, Ecuador
e-mail: oruiz@espol.edu.ec

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_8

Keywords Variable dimension control charts · Multivariate statistical process control · Sampling costs · Multivariate control charts · Adaptive control charts · Genetic algorithms · EWMA · T^2 control chart · Control chart design · Optimization

1 Introduction

With multivariate processes, it may happen that some quality variables are more expensive and/or difficult to measure than the other ones, or they may demand much more time to measure. Their measurement may even be destructive. Aparisi et al. (2012) give as an example a process of producing an electronic component, whose quality variables are two easily measured voltages and a third voltage which is the voltage that will burn it.

For monitoring such processes, the *variable dimension* approach was recently proposed. The idea is to measure always (at each sampling time) the "non-expensive" variables and to measure the expensive ones only when the values of the non-expensive variables give some level of evidence that the process may be out of control. An underlying assumption (validity condition) of the approach is that the variables are correlated and, although the inexpensive variables provide some information about the state of the process, the measurements of the expensive variables add more statistical evidence, increasing the probability of a signal when the process is out of control.

The approach can lead to significant savings in sampling costs (the gain depending, of course, on the ratio between the costs of measuring the "expensive" and the "inexpensive" variables). Also, in many cases, contradicting the intuition, the incorporation of the variable dimension approach to a control chart may even increase its speed in detecting special causes.

The general principle of the approach has been formalized concretely in a number of process control charts proposals, which differ in their specific forms. The purpose of this chapter is to review and compare the several variants of the approach. The first one to appear was the *variable-dimension* T^2 (*VDT2*) chart (Aparisi et al. 2012).

The procedure bears much similarity with the one of *variable parameters* (or *adaptive*) control charts, pioneered by Reynolds et al. (1988); other examples, far from being exhaustive, are Costa (1999) and, regarding the T^2 chart, Aparisi (1996) and Aparisi and Haro (2001). In these, the sample size and/or the sampling interval and/or other parameter of the control chart (such as the control limits or the smoothing parameter in EWMA schemes) are made variable according to the most recent sample information. The variable dimension approach differs though from the variable parameters approach in that it is not the sample size or sampling interval or control limits that are made dynamically variable, but rather the very variables being measured (thus the name "variable dimension").

Note that there is a difference between the variable dimension approach and all previous approaches that aim to reduce the dimensionality of the variable space,

such as principal components (Jackson 1980, 2003), latent variables or PLS methods, which are mostly used in the chemical industry (Kourti and MacGregor 1996; Nomikos and MacGregor 1995; Ferrer 2007, 2014), the U^2 chart (Runger 1996) and other similar approaches (Bodnar and Schmid 2005). Namely, all approaches cited, although reducing the dimension of the space considered for process control, require nevertheless measuring all variables (in the original high dimension space) prior to the transformation that leads to the dimensionality reduction. The variable dimension approach aims to reduce the number of variables actually measured. The goals are different, as are the underlying assumptions or context. The motivation of the previous approaches cited is the difficulty in interpreting and/or analyzing a huge number of variables (whereas there may be no problem in measuring them; for instance, the PLS approach is typically applied in data-rich environments in which sensors easily provide measurements of many variables with a high frequency). On the other hand, the variable-dimension approach is devised for situations in which, even if the number of variables may be small, some variables are much more costly to measure than the other ones.

An approach whose motivation is closer to the one of the variable-dimension approach is the variable selection method proposed by González and Sánchez (2010); with this, however, the dimensionality reduction is permanent: some variables are never measured. In the variable-dimension approach the number of variables measured is, as its name says, variable, in an adaptive way—that is, according to the information provided by the last sample statistic.

In the univariate case, Steiner (2000) has a similar motivation of reducing the cost of measurements through an adaptive procedure. His procedure differs from the variable dimension approach not only in being univariate but also in that what is made variable is the measurement system or device: he proposes alternating between "a fast but relatively inaccurate measurement system (...) and a more accurate and expensive, and possibly slower, alternative measurement device". The work deserves being mentioned here because of the similar idea of alternating measures in order to reduce the general cost, even if the particular context and concrete procedure are quite different from the ones of the variable dimension approach.

Four process control charts based on the variable-dimension approach have been developed. They are described, in chronological order, in the next four sections. The final section summarizes the main points.

2 The Variable-Dimension T^2 (VDT2) Control Chart

The VDT2 chart, developed by Aparisi et al. (2012), is one-sided. In its most general version, it has a pair of upper control limits (CL_1 and CL_2) and a pair of warning limits (w_1 and w_2), where the subscript "1" refers to the samples that have only the p_1 variables that are cheap and easy to measure and the subscript "2" refers to the samples that have all the p variables. When the sample has only p_1 variables, the T^2 statistic is computed only with the corresponding covariance submatrix.

When the sampling point (T_i^2, $i = 1, 2$), exceeds the corresponding control limit, the process is declared out of control; when $w_i \leq T_i^2 \leq C_i$ the next sample is taken with all p variables, and when $T_i^2 < w_i$, the next sample is taken with only the p_1 "inexpensive" variables.

The analysis of a large spectrum of cases in the chapter showed that the deterioration in performance was negligible when the warning limits were made equal ($w_1 = w_2 = w$); also, CL_1 could be made equal to infinity without significant effect on the chart performance. Making CL_1 equal to infinity is equivalent to have no control limit for samples with p_1 variables, and implies that a signal cannot occur with a sample with the p_1 variables. The performance is not impaired though, because this enables tightening the control limit (relatively to the CL_2 of the chart with two control limits) since a false alarm cannot occur with p_1 variables. On average, this compensates for the delay imposed by the need of a sample with p variables to have a signal: the resulting average run length is practically not reduced. The result of having only one control limit and one warning limit is a simpler control chart to operate.

For the details, the reader is referred to Aparisi et al. (2012).

The analysis showed that the VDT2 chart can considerably reduce the sampling costs and, quite surprisingly, even reduce the out-of-control ARL. This apparently paradoxical result can be ascribed to the aforementioned tightening of the control limit; this bears some analogy with the greater efficiency of adaptive control charts relative to fixed parameter charts, which comes from a better allocation of sampling effort. Another way of viewing this, as suggested by the editor of the journal in which the chapter was published,[1] is that the chart has some kind of memory, since another sample point is needed for a signal when a T^2 value from a sample with p_1 variables exceeds the warning limit. This constitutes a sort of "run rule".

For preserving space, we do not reproduce here the three pages of tables of the given reference, but the results were that in most cases analyzed the VDT2 chart exhibited an out-of-control ARL shorter than even the ARL of the T^2 chart on all p variables, together with a significant reduction in the sampling cost (the p variables having to be measured only part of the times). This refers to optimized designs. A computer program running in Windows and with a user-friendly interface was made available for such optimization. The percentage of times all variables are measured, $\%p$, is thus a result of the optimization, and depends on the shifts in the mean vector used for optimization. Only for very small shifts (for which the T^2 chart is quite inefficient though) this percentage is high as 70 or 80%; for large shifts it can be as low as 5%. The savings are quite relevant ($\%p$ ranges from 10 to 50%) for moderate shifts.

For moderate shifts, the reduction in the ARL provided by the variable-dimension approach is substantial; only for large shifts (that are quickly signalled even by the T^2 chart) there is no reduction or there is even a small increase, but this also results in small ARLs, of the order of 2 or less. On the other hand, these are cases where samples with all the variables are taken less than 20% of the times, and often less

[1] Daniel Apley.

than 10%. In addition, a sensitivity analysis has shown a considerable robustness of the optimal solutions with respect to the choice of the shifts for which to perform the optimization.

3 The Double-Dimension T^2 (DDT2) Control Chart

An idea that naturally comes to the mind is "When the sampling point exceeds the warning limit, why to wait for the next sampling time to measure the costly variables? Why not to measure them immediately?"

This idea has an intuitive appeal, by the analogy it bears (in operational terms) with double-sampling procedures (although with a distinction that is similar to the one between the VDT2 chart and variable-parameter control charts, namely that what is being increased is the number of variables rather than the sample size). At each sampling time, a sample is initially taken with p_1 variables only and the corresponding T^2 statistic ($T^2_{p_1}$) is calculated; if this is not sufficient to make a decision on the state of the process, then the "expensive" $p - p_1$ remaining variables are measured, the overall T^2 statistic based on the p dimensions (T^2_p) is calculated and compared with another control limit. The performance of the so-called *double-dimension* T^2 (*DDT2*) chart was investigated by Epprecht et al. (2013).

The DDT2 chart has, as double-sampling plans and double-sampling control charts (Croasdale (1974), Daudin (1992), Steiner (1999), Costa and De Magalhães (2005), Rodrigues et al. (2011); and specifically for the T^2 chart, Champ and Aparisi (2008)), a pair of thresholds for $T^2_{p_1}$ obtained from the initial sample with p_1 variables (in the case, a warning limit w and a control limit UCL_{p_1}) and a control limit for the statistic T^2_p obtained from the full dimension sample. The expensive variables are only measured when $w \leq T^2_{p_1} < UCL_{p_1}$.

The mathematical model for obtaining the ARLs of the DDT2 chart is conceptually more involved than the one for obtaining the ARLs of the VDT2 chart since it requires as an intermediate step the distribution of the difference $T^2_p - T^2_{p_1}$.

As with the VDT2 chart in Aparisi et al. (2012), a user-friendly program was also made available for optimization of the design of the DDT2 chart, and used for performance and sensitivity analyses. The analyses have shown that, however appealing the idea of not waiting for the next sampling time to measure the costly variables could be, the DDT2 chart did not reveal itself more efficient than the VDT2 chart: it presented in general ARLs similar to or larger than the ones of the VDT2 chart for the same shifts. Only in a very few cases the DDT2 chart ARLs were smaller, but not significantly. We will not linger on the DDT2 chart, for this reason.

Given the good results of the variable dimension approach (proven reduction in sampling costs, often accompanied by reduction in the out-of-control ARLs), a natural follow-up to the work on the VDT2 and DDT2 charts would be the investigation of more efficient versions of them. In particular, their performance, although good and even superior to the one of the T^2 chart on all variables, is poor for small shifts. Since the VDT2 chart exhibited equal or better performance than

the DDT2 chart, two extensions have been proposed to it: a variable sample size version of it, the VSSVDT2 chart (Aparisi et al. 2014) and an EWMA version of it, the VDEWMA-T2 chart (Epprecht et al. 2018). These are described next.

4 The Variable-Sample-Size Variable-Dimension T^2 (VSSVDT2) Control Chart

The VSSVDT2 control chart (Aparisi et al. 2014) combines, as its name indicates, the variable-dimension approach with the variable-sample-size (VSS) procedure proposed by Prabhu et al. (1993) and by Costa (1994). Several other VSS charts were proposed thereafter, being of particular interest in our context the VSST2 chart by Aparisi (1996).

The idea underlying the VSSVDT2 chart is the same of adaptive charts in general: to intensify inspection when there is more evidence that the process may be out of control (and to reduce it otherwise, in order not to increase the average inspection effort). For this purpose, the chart is constructed with two control limits, CL_{p_1} and CL_p, and a (single) warning limit, w. When the T^2 statistic of a sample exceeds the warning limit (but not the respective control limit), the next sample is taken with all p variables *and* sample size n_2; when it does not exceed w, the next sample is taken with only the p_1 "non-expensive" variables and sample size n_1. When using only p_1 variables and sample size n_1, the control limit to be considered is CL_{p_1} and, when using all p variables and sample size n_2, the control limit to be considered is CL_p. The very first sample, for the beginning of the monitoring or for resuming it after an alarm and intervention in the process, can be taken with p_1 variables and sample size n_1 or with all p variables and sample size n_2; this is an operational decision. In the chapter cited, the authors considered that this first sample is of small dimension and size.

The chart is illustrated in Fig. 1.

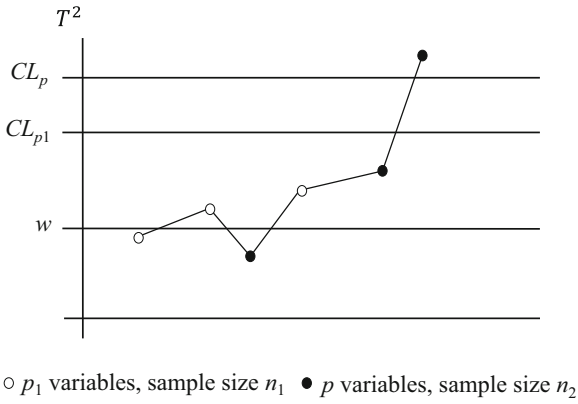


Fig. 1 VSSVDT2 chart

Similarly to the VDT2 chart, the performance analysis revealed that very often the control limit for samples with p_1 variables can be eliminated without any effect of practical significance on the performance of the VSSVDT2 chart. This makes the chart operationally simpler.

The optimization of the design of the chart is more complex (or more computationally intensive) than the ones of the VDT2 and DDT2 charts, because the number of decision variables is larger, since the chart has four or five design parameters: n_1, n_2, CL_p, (and CL_{p1} for the chart with two control limits), w. And to the constraint on the ARL_0, constraints are added on the average sample size (which should equal a specified value n_0) and on the maximum value acceptable for the larger sample size n_2. A program has also been developed, using a Markov chain model for the calculations and genetic algorithms for the optimization.

In contrast with the VDT2 and the DDT2 control charts, in which the economy in sampling costs is a straightforward function of the *proportion* of samples with p variables (so that this proportion can be used as a measure of the gain in sampling cost), with the VSSVDT2 chart, the relationship between the gain and this proportion is less direct, because the samples with p variables have larger size. The expected (or average) cost of a sample is given by

$$ACS = \frac{\%p}{100} \cdot C_{p_1}(a \cdot n_2 - n_1) + n_1 \cdot C_{p_1}$$

where a is the ratio between the costs C_p, of measuring p variables, and C_{p_1}, of measuring p_1 variables. Therefore, denoting by $\%p$ the percentage of times (samples) with p variables, the percent economy in sampling cost (relative to the T^2 chart) achieved with the VSSVDT2 chart can be straightforwardly derived as

$$\frac{\%p \cdot (a \cdot n_2 - n_1) + 100n_1}{n_0 \cdot a}$$

This ratio tends to the lower bound $\%p \cdot n_2/n_0$ when a tends to infinity.

The ACS of the VSSVDT2 chart with average sample size n_0 is higher than the ACS of the VDT2 chart with (fixed) sample size n_0. The ratio between them is

$$\frac{\%p \cdot (a \cdot n_2 - n_1) + 100n_1}{\%p \cdot n_0(a-1) + 100n_0}$$

which tends to n_2/n_0 when a tends to infinity.

These costs should be taken into account when deciding between using or not a VSSVDT2 chart. The performance analysis has shown that the VSSVDT2 chart provides great improvement in the ARL performance of the (fixed sample size) VDT2 chart: depending on the shifts considered, the ARLs can be reduced in 44–83%. This benefit should be balanced against the costs, which vary according to n_1, n_2 and a.

Again, a complete and more concrete picture of the performance of the VSSVDT2 chart would require a large number of tables, which are not pertinent

here, but are available in Aparisi et al. (2014). We just summarize below a couple of additional conclusions of the performance analysis in that chapter.

The ARL performance of the VSSVDT2 chart can never match the ARL performance of the VSST^2 chart (Aparisi 1996) on all p variables (in contrast with the VDT2 chart, which outperforms the T^2 chart on all p variables). But the cost of the VSST^2 chart on all p variables is larger, and the ARL differences are small. So, the VSSVDT2 chart remains an interesting option when a is large.

For large process shifts the VDT2 chart shows better, equal or very close performance to the one of the VSSVDT2 chart and becomes then the best choice, given its smaller sampling cost.

The higher cost of the VSSVDT2 chart relative to the VDT2 chart motivates investigating other enhancements to the VDT2 chart that do not increase its sampling cost. The EWMA procedure is one of the approaches known to speed up the detection of small to moderate shifts, with no increase in the cost of sampling (for a same value of $\%p$) and is operationally simpler than adaptive procedures (such as the VSS one). An EWMA version of the VDT2 chart is the subject of the next section.

5 The Variable-Dimension EWMA T^2 (VDEWMA-T2) Control Chart

The traditional multivariate EWMA chart is the MEWMA chart by Lowry et al. (1992). In this chart, at every sampling time, first the measures of all variables are smoothed separately, yielding (or rather updating) as many EWMA statistics as different variables, and then these EWMA statistics are combined into a single T^2 statistic. In that chapter, the choice of proceeding to the smoothing first was justified by the performance analysis, carried out by those authors, of this procedure and of the alternative procedure of smoothing the T^2 statistics of the successive samples, that would be computed for each sample prior to being entered into a single EWMA recursive expression. The analysis had shown that smoothing the data first led to faster detection of shifts in the process mean.

With the variable-dimension approach, however, it wouldn't make sense to smooth the successive values of the costly variables that would have been measured at irregular time intervals (skipping different numbers of sampling intervals), and, moreover, to compute T^2 statistics combining the EWMA values obtained this way (as if they were meaningful) with EWMA values of variables that would have been measured at regular time intervals. For this reason, the VDEWMA-T2 chart (Epprecht et al. 2018) computes the T^2 values first and next smooths them.

A difficulty remains, nevertheless: how to combine T^2 values from samples of different dimensions (T^2 values with different degrees of freedom) in a single EWMA statistic? The solution found was to scale these statistics, or to reduce them to a same measurement unit, so that they become comparable. Namely, a probability integral transformation is made, which is simply to compute the value of the cdf of the T^2 value of each sample, that is, to compute $F_{T^2_{p_1}}(T^2)$ in the case of the samples

with p_1 variables and $F_{T_p^2}(T^2)$ in the case of the samples with p variables, where $F_{T_{p_1}^2}(\cdot)$ and $F_{T_p^2}(\cdot)$ denote the cdfs of the in-control T^2 statistic from samples with p_1 and with p variables, respectively. These are measures of the statistical evidence that the process might be off-target. This way, at each sampling time, a T^2 statistic is calculated and its cdf value is obtained. Next, to make easier the operation of the chart, the cumulative probability thus obtained is converted to a Z score, by use of the inverse cumulative standard normal distribution. The normal distribution was chosen just for convenience; the point is that the result is a value of the $N(0, 1)$ distribution that has the same exceedance probability as the T^2 value obtained from the sample, regardless of the number of variables in it. These Z values can then be smoothed in an EWMA statistic.

The VDEWMA-T2 chart has one control limit and one warning line. Also, a reflecting boundary (lower bound for the EWMA statistic) was added to make the chart more sensitive to shifts in the process mean. The use of such a bound for one-sided EWMA charts was proven effective by Gan (1993) and adopted since by other authors.

A VDEWMA-T2 chart is depicted in Fig. 2.

The chart operation is as follows: at every sampling time, a sample is taken. It will consist of measures of only the subset of p_1 "inexpensive" variables if the previous point fell below the warning line; and it will consist of measures of all the variables if the previous point fell between the warning line and the control limit. A point above the control limit is a signal; the first sample after a signal (after investigation for special causes and resuming the monitoring) may consist of measures of only the subset of p_1 variables or of measures of all the variables; this is up to the user, a decision of practical nature. The performance analysis in Epprecht et al. (2018) considered that it would consist only of measures of the subset of p_1 variables, for economy and because after the intervention it is more likely that the process is in control.

Taken the sample, the T^2 statistic is computed, either with $p_1 - 1$ or with $p - 1$ degrees of freedom (according to the sample dimension) and the cumulative

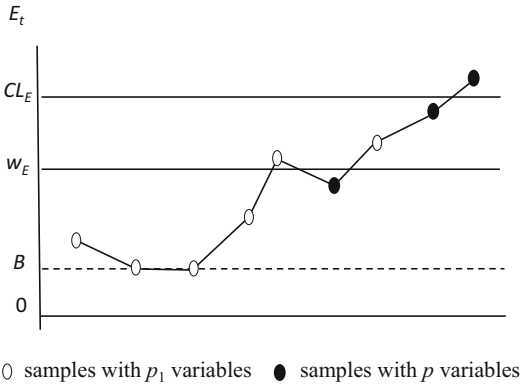


Fig. 2 VDEWMA-T2 chart

probability of that T^2 value is converted to a Z score by:

$$z_t = \Phi^{-1}\left(F_{\chi^2_v}\left(T^2_v(t)\right)\right)$$

It is the Z score which is smoothed into an EWMA statistic:

$$E_t = \max\left\{B, rz_t + (1-r)E_{t-1}\right\}$$

where r is the smoothing constant and $E_0 = B$.

After a signal and intervention, for resuming the monitoring, the initial value of the EWMA is returned to $E_0 = B$.

In contrast with the VDT2 chart, whose control limit is active only with samples of p variables, the single control limit of the VDEWMA-T2 chart is always active. This makes sense because it applies to an EWMA value that combines data from several samples, of both dimensions (p_1 and p), and which had been put to a same "scale" through the probability integral transformation (which yields the corresponding cumulative probability) and computation of the Z score.

Just for the record, the authors had analyzed another EWMA scheme, consisting of two charts: a VDT2 chart (with only one control limit and one warning line) combined with an EWMA chart on the Z score, computed the same way as indicated above. The differences are that the decision for switching from p_1 to p variables (and vice-versa) is based on the T^2 value in the VDT2 chart, and that this chart can also signal.

The performance analysis has been carried out using Markov chain models for computing the ARLs. These models were also used by computer programs for optimization of the charts design. The programs, also running in Windows and with user-friendly interfaces, take as entries the desired ARL_0 and the shift for which the (steady-state) out-of-control ARL (ARL_1) should be minimized. The decision variables are the charts limits, the reflecting boundary and the smoothing constant.

The analysis has shown that the two versions of EWMA schemes (the VDEWMA-T2 chart and the joint VDT2 and EWMA charts) performed quite similarly. Then the VDEWMA-T2 chart was the only retained and described in detail in the chapter, because it is operationally simpler. The Markov chain model of the joint scheme is much more involved, too, and its optimization is more time-consuming in processing time.

An interesting result is that the optimization based on ARL_1 minimization leads almost always to solutions where p variables are measured in all samples or in a quite large (over 95%) proportion of samples. That is, the variable-dimension procedure degenerates into a fixed-dimension one. This should be intuitively expected, weren't it the fact that with the VDT2 chart the same ARL_1 minimization criterion leads to solutions in which the p variables are measured only a small proportion of the times. This contrasting behavior of the optimal solutions for the VDEWMA-T2 chart is not fully understood; maybe (this is only a conjecture) the reason is that, unlike the VDT2 chart, the VDEWMA-T2 chart cannot benefit from the non-existence of a control limit for samples with p_1 variables to reduce the

control limit for samples with p variables, and, as the EWMA statistic "drifts" slowly (in contrast with the serial independence of the T^2 values in the VDT2 chart), taking the samples with all variables will make the VDEWMA-T2 chart signal out-of-control conditions faster.

This observation showed the need to introduce a constraint on the percentage of times that all variables are sampled (denoted by $\%p$) in the optimization problem. The program admits this as an input from the user. The solutions satisfying this constraint still have smaller out-of-control ARLs than the ones of the VDT2 chart.

The user can then set $\%p$ at any desired value, say 50% or 30%. They can also try different values to choose a solution based on cost-benefit analysis. The average cost of one sample is $ACS = \left(1 + a \cdot \frac{\%p}{100}\right) C_{p_1}$, where C_{p_1} is the cost of a sample with only p_1 variables and the cost of a sample with all variables is aC_{p_1}. With a sampling interval of h, the sampling cost per time equals ACS/h. The benefit is the detection speed, which is the reciprocal of the average detection delay $AATS = (ARL_1 - 0.5)h$. The product

$$(ACS/h) \cdot AATS = \left(1 + a \cdot \frac{\%p}{100}\right) C_{p_1}(ARL_1 - 0.5)$$

(note how h cancels out in the right-hand side number) corresponds to *cost per time* over *detection speed*. It can be used as an objective function. The user can then try different values of $\%p$, get the solutions, calculate the quantity above and the solution that minimizes it is the most efficient. Then, h can be determined according to a maximum feasible/tolerated sampling cost per time ACS/h (and the AATS will be minimized according to this constraint). Alternatively, one can determine h according to a constraint on the AATS (and the sampling cost per time will be minimized).

The reader is referred to Epprecht et al. (2018) for more details and extensive tables of results, but in general, for small and moderate shifts in the process mean, with constraints of $\%p = 30\%$ and 50%, the reductions in the ARL with respect to the VDT2 chart range from 30% to 50%, approximately (larger $\%p$ leading to larger reductions, naturally).

The VDEWMA-T2 chart presents two additional advantages over the VSSVDT2 chart: namely, it always outperforms the VDT2 chart (while the VSSVDT2 chart not always does), and it can be used with samples of small size and even size 1 (the VSSVDT2 chart requires a pair of sample sizes, a smaller and a larger one).

6 Summary

In multivariate process control, when some of the quality variables are much more costly to measure than the other ones, the variable-dimension approach can lead to substantial reduction in the sampling costs, being still very effective in

signalling out-of-control situations. We reviewed the existing charts using this approach. Surprisingly, the variable-dimension T^2 chart (VDT2 chart) can signal mean shifts even faster than its fixed-dimension counterpart, requiring measuring all variables only a limited proportion of the times. The double dimension T^2 chart (DDT2) chart exhibits equivalent behavior. The variable-dimension EWMA-T^2 chart (VDEWMA-T2 chart) is still faster than them. The variable-sample-size VDT2 chart (VSSVDT2 chart) is another enhancement to the VDT2 chart. User-friendly software was developed for every one of these charts, for automatically performing the optimization of the chart design, thus making the techniques applicable in practice. For a more detailed presentation and analysis of each of these charts, the reader is referred to the original chapters.

Acknowledgements The first author was partly supported by the CNPq (Brazilian Council for the Scientific and Technological Development).

References

Aparisi, F. (1996). Hotelling's T^2 control chart with adaptive sample sizes. *International Journal of Production Research, 34*, 2853–2862.

Aparisi, F., Epprecht, E. K., Carrión, A., & Ruiz, O. (2014). The variable sample size variable dimension T^2 control chart. *International Journal of Production Research, 52*(2), 368–383.

Aparisi, F., Epprecht, E. K., & Ruiz, O. (2012). T^2 control charts with variable dimension. *Journal of Quality Technology, 44*(4), 375–393.

Aparisi, F., & Haro, C. (2001). Hotelling's T^2 control chart with variable sampling intervals. *International Journal of Production Research, 39*(14), 3127–3140.

Bodnar, O., & Schmid, W. (2005). Multivariate control charts based on a projection approach. *Allgemeines Statistisches Archiv, 89*, 75–93.

Champ, C. W., & Aparisi, F. (2008). Double sampling Hotelling's T^2 charts. *Quality and Reliability Engineering International, 24*, 153–166.

Costa, A. F. B. (1994). $\bar{X}$ charts with variable sample size. *Journal of Quality Technology, 26*(3), 155–163.

Costa, A. F. B. (1999). $\bar{X}$ charts with variable parameters. *Journal of Quality Technology, 31*(4), 408–416.

Costa, A. F. B., & De Magalhães, M. S. (2005). Economic design of two-stage $\bar{X}$ charts: The Markov-chain approach. *International Journal of Production Economics, 95*(1), 9–20.

Croasdale, R. (1974). Control charts for a double-sampling scheme based on average production run length. *International Journal of Production Research, 12*, 585–592.

Daudin, J. J. (1992). Double sampling charts. *Journal of Quality Technology, 24*(2), 78–87.

Epprecht, E. K., Aparisi, F., & Ruiz, O. (2018). Optimum variable-dimension EWMA chart for multivariate statistical process control. *Quality Engineering, 30*(2), 268–282. http://doi.org/10.1080/08982112.2017.1358367

Epprecht, E. K., Aparisi, F., Ruiz, O., & Veiga, A. (2013). Reducing sampling costs in multivariate SPC with a double-dimension T^2 control chart. *International Journal of Production Economics, 144*(1), 90–104.

Ferrer, A. (2007). Multivariate statistical process control based on principal component analysis (MSPC-PCA): Some reflections and a case study in an autobody assembly process. *Quality Engineering, 19*, 311–325.

Ferrer, A. (2014). Latent structures-based multivariate statistical process control: A paradigm shift. *Quality Engineering, 26*, 72–91.

Gan, F. F. (1993). Exponentially weighted moving-average control charts with reflecting boundaries. *Journal of Statistical Computation and Simulation, 46*(1–2), 45–67.

González, I., & Sánchez, I. (2010). Variable selection for multivariate statistical process control. *Journal of Quality Technology, 42*(3), 242–259.

Jackson, J. E. (1980). Principal components and factor analysis: Part I—principal components. *Journal of Quality Technology, 12*(4), 201–213.

Jackson, J. E. (2003). *A User's Guide to Principal Components*. Hoboken: Wiley.

Kourti, T., & MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology, 28*(4), 409–428.

Lowry, C. A., Woodall, W. H., Champ, C. W., & Rigdon, S. E. (1992). A multivariate exponentially weighted moving average chart. *Technometrics, 34*, 46–53.

Nomikos, P., & MacGregor, J. F. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics, 37*(1), 41–59.

Prabhu, S. S., Runger, G. C., & Keats, J. B. (1993). An adaptive sample size chart. *International Journal of Production Research, 31*, 2895–2909.

Reynolds, M. R. Jr., Amin, R. W., Arnold, J. C., & Nachlas, J. A. (1988). $\bar{X}$ charts with variable sampling intervals. *Technometrics, 30*(2), 181–192.

Rodrigues, A. A. A., Epprecht, E. K., & De Magalhães, M. S. (2011). Double sampling control charts for attributes. *Journal of Applied Statistics, 38*(1), 87–112.

Runger, G. C. (1996). Projections and the U-squared multivariate control chart. *Journal of Quality Technology, 28*, 313–319.

Steiner, S. H. (1999). Confirmation sample control charts. *International Journal of Production Research, 37*(4), 737–748.

Steiner, S. H. (2000). Statistical process control using two measurement systems. *Technometrics, 42*, 178–187.

Distribution-Free Bivariate Monitoring of Dispersion

Ross Sparks and Subha Chakraborti

Abstract This chapter focuses on evaluating practical approaches to monitoring the dispersion for a wide range of positively distributed and correlated bivariate data. It provides good practical advice regarding monitoring the dispersion of variables with skewed distributions.

Keywords Asymmetric distributions · Statistical process control · Variance

1 Introduction

Sewerage treatment plants (STP) deal with volatile and noisy inputs (e.g., see Hamed et al. 2004) and, therefore, need to regulate their treatment processes accordingly (e.g., Choi and Park 2001) to have effluent output that will do as little harm as possible when discharged to the surrounding environment. STPs typically monitor Biological Oxygen Demand, Chemical Oxygen Demand, Total Organic Carbon and Total Suspended Solids (TSS) as well as Total Nitrogen, Ammonium Nitrogen, Nitrate, Phosphorus, Temperature and pH. In addition, it provides information on the out-going effluent quality and treatment efficiency. The volatility in these variables often provides us with information about the underlying control process of the STP. In the STP application in this chapter, the two variables we have near complete data on are TSS and Total Residual Chlorine(TRC), and so we are going to use these variables to demonstrate processes for monitoring of

R. Sparks (✉)
CSIRO Australia, Data61, Sydney, Australia
e-mail: Ross.Sparks@csiro.au

S. Chakraborti
Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa, AL, USA
e-mail: schakrab@cba.ua.edu

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_9

bivariate volatility as an assessment of control process of the STP (e.g, see Saby et al. 2002). Although this example does not solve the problem of monitoring the volatility of full STP process, it does demonstrate it for the bivariate case before moving onto the more difficult multivariate case. This bivariate case will be covered in the application section later after developing the monitoring methodology.

Non-parametric (distribution-free) charts are growing in popularity in the literature because control measures often have asymmetric distributions and the distribution of the control variables are generally unknown. In particular, environmental measures such as e-coli, chlorophyll, nutrient loads, turbidity, etc. are all positive right-skewed measures. Often the log-normal distribution is assumed for these measures as a matter of convenience (e.g., Sukumar et al. 2002). The assumption that the variables are log-normally distributed is inappropriate at times (e.g., see Dodds et al. 1998). The monitoring of log-normal distribution data is handled in two ways, firstly on the log-scale (which separates the mean and dispersion parameters, e.g., Morrison (1958) and Joffe and Sichel (1968)), and secondly on the untransformed scale (Ferrell 1958; Cheng and Xie 2000). The option of transforming the data using a Box-Cox (Box and Cox 1964) transformation and then applying the S-chart has been demonstrated as unreliable, particularly for flagging changes in dispersion. Therefore, alternatives need to be investigated that are more reliable.

This chapter focuses on evaluating practical approaches to monitoring the dispersion for a wide range of positively distributed bivariate data. It plans to provide good practical advice to those monitoring variables which typically follow an unknown skewed distribution. In this chapter we explore Liu 's (1990)'s data depth function in R as a means of assessing outliers or out-of-control situations. As an alternative to this methodology we explore regressing ordered statistics against their expected values conditional on the data being in-control. Assume that we have a rational subgroup of twenty observations, then we order these from smallest to largest value and compare these to their expected values when the observations are drawn from an in-control distribution. Let the ordered rational subgroup of size n for the kth time period be denoted

$$x_k^{(1)} \leq x_k^{(2)} \leq \ldots x_k^{(n)}.$$

Let $E\left(x_k^{(j)}\right) = \mu_k^{(j)}$ and therefore $\mu_k^{(1)} \leq \mu_k^{(2)} \leq \ldots \leq \mu_k^{(10)}$.

Then we build the regression model

$$x_{ik}^{(j)} = \alpha_k + \beta_k \mu_k^{(j)} + e_{ik}.$$

Theoretically when the rational subgroup values are in-control, then α_k is equal to zero and β_k is equal to one. However, we estimate the $\mu_k^{(j)}$'s values using the Phase I data and therefore these are not without error. Hence $\alpha_k = 0$ and $\beta_k = 1$ is not always true in practice. In addition, we need fairly large rational subgroups to estimate these regression coefficients accurately. If this regression model is fitted using only rational subgroup sample sizes of 10 or less, then these estimates can vary

substantially from the values expected theoretically. Therefore, we assume a rational subgroup of 20 for the remainder of the chapter, but note that traditional rational subgroups of size 20 are fairly rare in practice. Our focus on dispersion means we examine how β_k differs from the expected value of one. If it is significantly larger than one, then the rational subgroup has a larger standard deviation than in the in-control case. If it is significantly smaller than one, then the rational subgroup has a smaller standard deviation than in the in-control case.

2 Bivariate Control Charts: Monitoring Changes in Dispersion

Now we explore bivariate control charts with the aim of extending these to multivariate control charts for dispersion. The first idea was to look at data depth as a way of flagging increases in dispersion.

2.1 Bivariate Dispersion Monitoring Using Data Depth

A sample of 10,000 training data of rational subgroups of 20 observations were generated from one of the distributions. This training data were used to estimate the number of in-control points that are expected to have Liu (1990)'s data depth score of zero where a depth of zero indicates an outlier (see also Stoumbos et al. 2001; Liu et al. 1999). If we know that these outlying points don't cluster in a small region in the bivariate space then this is likely to be an outbreak in dispersion. In other words, too many extreme points that don't cluster in the two dimensional space indicates an increase in dispersion. The Liu (1990)'s depth score is not useful for assessing decreases in dispersion, but it can indicate increases in dispersion if we can demonstrate that these are not related to a shift in location. We decided to use the count of the number of Liu (1990)'s depth scores of zeros in the rational subgroup as a way of flagging increases in dispersion recognizing that this out-of-control criteria does not differentiate between changes in location and changes in dispersion. Despite this drawback, this statistic works comparatively well at flagging changes in dispersion as will be demonstrated later.

For normally distributed data in 10,000 simulation runs and a rational subgroup of 20: 12 of the 10,000 rational subgroups had one observation with a depth of zero and one with two zeros. Therefore the decision rule for declaring the process is out-of-control is taken when either:

1. Any rational subgroup of size 20 with two or more observations have depth equal to zero; or
2. Two or more consecutive rational subgroups of size 20 have one depth equal to zero.

Table 1 The number of flags for increased (decreased) dispersion for the various approaches; log-normal distribution

$d(a, b)$	Box-Cox Transformation				
	Log	Log			
	Volume	Perimeter			
	thresholds		Robust	Robust	Data depth
	Upper	Upper	regression	regression	(2 or more
	(Lower)	(Lower)	(P_i)	(Q_i)	depths=0 or
$LN(\mu, \sigma)$	(7.5393)	(5.6227)	$h = 6.259469$	$h_q = 5.33291$	two consecutive one depth=0)
$X \sim LN(0, 1)$, $Z \sim LN(0, \sqrt{0.75})$	9 (17)	7 (15)	9	11	8
$X \sim LN(0, 1.25)$, $Z \sim LN(0, \sqrt{0.75})$	278 (4)	**655** (2)	567	588	291
$X \sim LN(0, 1.5)$, $Z \sim LN(0, \sqrt{0.75})$	1728 (0)	**3680** (0)	2205	2391	1756
$X \sim LN(0, 1.75)$, $Z \sim LN(0, \sqrt{0.75})$	4063 (0)	**7042** (0)	4480	4631	4236
$X \sim LN(0, 2)$, $Z \sim LN(0, \sqrt{0.75})$	6504 (0)	**8940** (0)	6413	6473	6180
$X \sim LN(0, 1)$, $Z \sim LN(0, \sqrt{1.25})$	199 (2)	131 (0)	120	117	**310**
$X \sim LN(0, 1)$, $Z \sim LN(0, \sqrt{1.5})$	537 (0)	325 (3)	251	244	**1104**
$X \sim LN(0, 1)$, $Z \sim LN(0, \sqrt{1.75})$	872 (0)	533 (0)	478	467	**1757**
$X \sim LN(0, 1)$, $Z \sim LN(0, \sqrt{2})$	1400 (0)	886 (0)	909	811	**3032**
$X \sim LN(0, 1)$, $Z \sim LN(0, \sqrt{2.25})$	2028 (0)	1384 (0)	1558	1263	**4006**
$X \sim LN(0, 1)$, $Z \sim LN(0, \sqrt{2.5})$	2646 (0)	1865 (0)	2145	1746	**4918**
$X \sim LN(0, 1)$, $Z \sim LN(0, \sqrt{2.75})$	3165 (0)	2429 (0)	2820	2247	**5612**

The bold numbers in the table are the lowest ARL values for detecting the simulated outbreaks

For in-control normally distributed data this provides 1 to 2 false discoveries in 10,000 simulations. This same approach will be tried for all examples of bivariate data. We demonstrate in Table 1 that these rules for flagging an increase in dispersion work reasonably well.

The major issue with data depth measures is that it does not distinguish between changes in location and changes in dispersion, and the rules above fail to flag decreases in dispersion. The approach considered in the next section does differentiate between changes in location and dispersion as well as differentiating between increases and decreases in variation.

2.2 Bivariate Approach Using an Extension of the Robust Regression Approach: Outline for Univariate Distributions

Establish the median order statistic value across all 1000 rational subgroups of size 20 for both bivariate datasets ($k = 1, 2$), i.e., denote these $\mu_k^{(1)} \leq \mu_k^{(2)} \leq \ldots \leq \mu_k^{(n)}$ such that $\mu_k^{(j)} = median(x_{1,k}^{(j)}, x_{2,k}^{(j)}, \ldots, x_{1000,k}^{(j)})$ for all $j = 1, 2, \ldots, n$. The median is selected rather than the mean because it was more robust across the broad range of distributions considered. The values

$$\mu_k^{(1)} \leq \mu_k^{(2)} \leq \ldots \leq \mu_k^{(n)}$$

are the reference values as defined in the introduction section for the bivariate data $k = 1, 2$ which are used to gauge whether the dispersion of a rational subgroup of size 20 has increased.

Denote

$$\mathbb{X}_q = \begin{pmatrix} \mu_1^{(1)} & \mu_2^{(1)} \\ \mu_1^{(2)} & \mu_2^{(2)} \\ \vdots & \vdots \\ \mu_1^{(n)} & \mu_2^{(n)} \end{pmatrix} \quad \text{and} \quad \mathbb{X}_i = \begin{pmatrix} x_{i,1}^{(1)} & x_{i,2}^{(1)} \\ x_{i,1}^{(2)} & x_{i,2}^{(2)} \\ \vdots & \vdots \\ x_{i,1}^{(n)} & x_{i,2}^{(n)} \end{pmatrix}$$

with

$$\tilde{\mathbb{X}}_q = \mathbb{X}_q - \mathbf{1}_n^t \mathbb{X}_q / n \quad \text{and} \quad \tilde{\mathbb{X}}_i = \mathbb{X}_i - \mathbf{1}_n^t \mathbb{X}_i / n$$

where $\mathbf{1}_n$ is an n by 1 column vector of ones. Note that $\hat{\beta}_{i1}$ is the robust estimate of the regression slope using the rlm function in the MASS (Venables and Ripley 2002) package of R. Applying the usual quadratic form, we flag significant changes in dispersion when

$$P_i = \left(\hat{\beta}_{i1} - 1 \;\; \hat{\beta}_{i2} - 1\right) \tilde{\mathbb{X}}_i' \tilde{\mathbb{X}}_i \begin{pmatrix} \hat{\beta}_{i1} - 1 \\ \hat{\beta}_{i2} - 1 \end{pmatrix} / \operatorname{tr}\left(\tilde{\mathbb{X}}_i' \tilde{\mathbb{X}}_i\right) \; > \; h \,.$$

An alternative criterion that is more closely aligned with the traditional quadratic form that flags significant changes in dispersion when

$$Q_i = \left(\hat{\beta}_{i1} - 1 \;\; \hat{\beta}_{i2} - 1\right) \tilde{\mathbb{X}}_q' \tilde{\mathbb{X}}_q \begin{pmatrix} \hat{\beta}_{i1} - 1 \\ \hat{\beta}_{i2} - 1 \end{pmatrix} / \operatorname{tr}\left(\tilde{\mathbb{X}}_q' \tilde{\mathbb{X}}_q\right) \; > \; h_q \,.$$

Next, we outline the process of estimating the threshold necessary for delivering an acceptable false alarm rate for flagging significant bivariate changes in dispersion. We simulate ten thousand in-control rational sub-groups (i), estimate the parameters of the following simple linear model for the kth variable: $x_{ik}^{(j)} =$

$\alpha_k + \beta_k \mu_k^{(j)} + e_{ik}$. Denote these slope estimates $\hat{\beta}_{ik}$ for simulated rational subgroups $i = 1, \ldots, 10{,}000$ and $k = 1, 2$. These are used to establish the threshold for significant increases in dispersion by calculating Q_i/P_i for $i = 1, 2, \ldots, 10{,}000$. Estimate the quantile values for these statistics that correspond to an acceptable out-of-control false alarm rate. The statistics (P_i or Q_i) do not distinguish between increases and decreases in dispersion but if these are represented as a two dimensional plot of $(\hat{\beta}_{i1}\hat{\beta}_{i2})$ with the control limit as an ellipse as described in Sparks (1992), then diagnosing the nature of the significant changes is easy. Alternatively, observing the $\hat{\beta}_{ij}$ values independently for each j provides the necessary information for diagnosing the nature of significant changes.

2.3 Transformation to a Normal Distribution

Here we consider using a Box-Cox transformation to normality and then use charts derived for normally distributed bivariate data. The main advantage of this approach is that the plan simply involves finding the appropriate Box-Cox transformation, and then the existing traditional design of the chart for the normal distribution applies. However, this is not as simple as it sounds. For example, with log-normal data we know that the logarithm transform is the appropriate transform if the correlated variables X_1 and $X_2 = X_1^{\beta} Z$ where X_1 and Z are log-normally distributed. Then notice that X_2 is log-normally distributed. This means that the thresholds for $\log(X_1)$ and $\log(X_2) = \log(X_1^{\beta} Z) = \beta \log(X_1) + \log(Z)$ are easy to simulate and deliver appropriate thresholds. This is not that easy when the response variables select different transformations for the two variables X_1 and X_2 to individually approximate to normality. If f_1 and f_2 are the two transformations, then we need an approximation that simulates the appropriate thresholds for the bivariate normal approximation that will apply to bivariate variables $f_1(X_1)$ and $f_2(X_2)$. Assume that X and Z are independent random variables and that $X_1 = X$ and $X_2 = 0.5X + Z$ but this is hidden (unknown). We then find $E\big(f_1(X_1)\big) = \mu_1$, $E\big(f_2(X_2)\big) = \mu_2$, $Var\big(f_1(X_1)\big) = \sigma_1^2$, $Var\big(f_2(X_2)\big) = \sigma_2^2$ and $Cov\big(f_1(X_1), f_2(X_2)\big) = \sigma_{12}$ and this provides us with the appropriate normal distribution for setting up the thresholds for the control variables to follow. Although this approach is feasible, it at times fails to deliver a reasonable plan as we will see later in the section discussing the simulated examples. If we knew that $X_1 = X$ and $X_2 = 0.5X + Z$, then we would tranform $Z = X_2 - 0.5X$ to approximate normal, say using $f_3(Z)$ and then used $f_2(X_2) = 0.5 f_1(X) + f_3(Z)$. This would be deliver better results.

Mathematically if we knew the in-control covariance matrix Σ_0 and the transformed sample covariance matrix $\mathbb{S}$, then a control variable could be a function of the eigenvalues of $\Sigma_0^{-1}\mathbb{S}$. The difficulty is, this is not a meaningful measure for the control engineer. If we took instead the determinant of Σ_0 denoted $|\Sigma_0|$, then this is a measure of the volume of space (or area in the bivariate case) the in-control data usually "occupies" in the multivariate space, and the trace of Σ_0

denoted $\text{tr}(\Sigma_0)$ is proportional to the perimeter of space the data "occupies". These are both meaningful measures of variation and therefore $\text{tr}(\mathbb{S})$ and $|\mathbb{S}|$ are meaningful statistics worth monitoring. We flag increases in dispersion when either:

$$\text{tr}(\mathbb{S}) > \text{tr}(\Sigma_0) + h_{tr,upper} \quad \text{or} \quad |\mathbb{S}| > |\Sigma_0| + h_{d,upper}$$

and flagging a decrease in variation when

$$\text{tr}(\mathbb{S}) < \text{tr}(\Sigma_0) - h_{tr,lower} \quad \text{or} \quad |\mathbb{S}| < |\Sigma_0| - h_{d,lower}$$

where $h_{a,upper}$ and $h_{a,lower}$ are positive values with $a = tr$ or d depending on whether the trace or determinant is used, respectively. These thresholds are trained to deliver a specified false alarm rate. This does mean that these thresholds need to be trained as the in-control variance changes, but it is better to have a meaningful measure for the control engineer to use than one that is not. These thresholds are trained using normally distributed data. All of these relate to the eigenvalues of the sample covariance matrix or equivalently the singular values of the singular value decomposition (*svd*) of the departures the observations are from their sample means. These singular values are used because it limits the computational effort involved in the control plans. The product and sum of *svd* singular values of matrix $[\mathbb{X}_1 - \bar{x}_1 \quad \mathbb{X}_2 - \bar{x}_2]$ are proportional to the volume and perimeter, respectively. The thresholds for these are found by simulation based on the assumption that the transformed data are normally distributed. If the distribution is known, then we can do better than this by simulating data from this known distribution to find the thresholds.

2.4 Some Simulation Results

All bivariate distributions considered in this chapter are two parameter distributions denoted by $d(a, b)$ with parameters a and b. The simulated data used had $x \sim d(a, b1)$ and $y \sim 0.5x + z$ where $z \sim d(a, b2)$ independent of x, and a, $b1$ and $b2$ are defined in Table 1. The simulation results are included in Table 1. The bivariate in-control data were simulated using the distributions in red ink. Each simulation generated 10,000 independent rational subgroups and recorded the number of alarms in these 10,000 simulations.

The out-of-control simulated data are in black ink with either both control variables changing when the distribution of X_1 departs from the in-control distribution, or for the second variable (X_2) changing when only the Z variable changes from its in-control distribution. The thresholds were trained using a bootstrap sample from a Phase I dataset of 10,000 observation from the known distribution (but hidden from the user). An exception is the data depth method where the rules defined earlier were used. The in-control false alarms were then checked using in-control data and distribution-free methods; these are reported in Tables 1, 2, 3, 4, 5, 6 and 7, for

Table 2 The number of flags for increased (decreased) dispersion for the various approaches; inverse Gaussian distribution

$d(a, b)$	Box-Cox Transformation				
	−0.06	−0.03			
	Volume	Perimeter			
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)	Robust	Robust	
	1.0464	6.8207	regression (P_i)	regression (Q_i)	
$IG(\mu, \sigma)$	(0.3790)	(6.4754)	$h = 3.09924$	$h_q = 2.595764$	Data depth
$X \sim IG(1, 1)$,	16	9	14	10	9
$Z \sim IG(1, \sqrt{0.75})$	(3)	(23)			
$X \sim IG(1, 1)$,	98	60	55	62	**278**
$Z \sim IG(1, \sqrt{1.25})$	(0)	(16)			
$X \sim IG(1, 1)$,	218	81	84	97	**477**
$Z \sim IG(1, \sqrt{1.5})$	(0)	(0)			
$X \sim IG(1, 1)$,	413	124	138	104	**1619**
$Z \sim IG(1, \sqrt{1.75})$	(2)	(3)			
$X \sim IG(1, 1)$,	573	199	169	125	**2750**
$Z \sim IG(1, \sqrt{2})$	(0)	(3)			
$X \sim IG(1, 1)$,	890	312	231	18	**3520**
$Z \sim IG(1, \sqrt{2.25})$	(1)	(1)			
$X \sim IG(1, 1)$,	1086	384	325	178	**3269**
$Z \sim IG(1, \sqrt{2.5})$	(0)	(0)			
$X \sim IG(1, 1)$,	1086	384	325	178	**3269**
$Z \sim IG(1, \sqrt{2.5})$	(0)	(0)			
$X \sim IG(1, 1)$,	3667	1889	603	467	**8792**
$Z \sim IG(1, \sqrt{4.75})$	(0)	(0)			
$X \sim IG(1, 1)$,	6316	4993	784	628	**9931**
$Z \sim IG(1, \sqrt{8.75})$	(0)	(0)			
$X \sim IG(2, 1)$,	114	0	**4491**	4105	638
$Z \sim IG(1, \sqrt{0.75})$	(0)	(2270)			
$X \sim IG(3, 1)$,	308	0	**7931**	7726	2999
$Z \sim IG(1, \sqrt{0.75})$	(0)	(4928)			
$X \sim IG(4, 1)$,	0	0	**8886**	8697	3709
$Z \sim IG(1, \sqrt{0.75})$	(559)	(6415)			

The bold numbers in the table are the lowest ARL values for detecting the simulated outbreaks

a number of distributions. In Table 1, e.g., for log-normal data the false alarms are very similar as 9, 11 and 8 in the 10,000 simulations for robust regression using P, Q and data depth, respectively. For the log-normal X and Y, we assumed that we know that $X_1 = X$ and $X_2 = 0.5X + Z$ and this helps improve the design of the bivariate

Table 3 The number of flags for increased (decreased) dispersion for the various approaches; Weibull distribution

$d(a, b)$	Box-Cox Transformation				
	0.27	0.22			
	Volume	Perimeter			
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)	Robust	Robust	
	7.8774	8.0077	regression (P_i)	regression (Q_i)	
$WEI(\mu, \sigma)$	(2.8452)	(6.1948)	$h = 1.707337$	$h_q = 1.470695$	Data depth
$X \sim WEI(1, 1)$,	36	12	32	9	19
$Z \sim WEI(1, \sqrt{0.75})$	(44)	(20)			
$X \sim WEI(1.5, 1)$,	134	649	**701**	321	130
$Z \sim WEI(1, \sqrt{0.75})$	(56)	(1)			
$X \sim WEI(2, 1)$,	224	3433	**6933**	1287	950
$Z \sim WEI(1, \sqrt{0.75})$	(35)	(0)			
$X \sim WEI(2.5, 1)$,	331	6598	**8133**	3255	2521
$Z \sim WEI(1, \sqrt{0.75})$	(37)	(0)			
$X \sim WEI(3, 1)$,	514	8613	**9500**	6064	4402
$Z \sim WEI(1, \sqrt{0.75})$	(35)	(0)			
$X \sim WEI(1, 0.7)$,	922	469	**1697**	1434	1002
$Z \sim WEI(1, \sqrt{0.75})$	(3)	(23)			
$X \sim WEI(1, 0.6)$,	2509	1346	**3357**	3193	2878
$Z \sim WEI(1, \sqrt{0.75})$	(0)	(24)			
$X \sim WEI(1, 0.5)$,	5272	3328	**5707**	5581	5488
$Z \sim WEI(1, \sqrt{0.75})$	(35)	(13)			
$X \sim WEI(1, 0.4)$,	**8181**	6015	7706	7595	7942
$Z \sim WEI(1, \sqrt{0.75})$	(0)	(19)			
$X \sim WEI(1, 0.3)$,	**9696**	8447	9011	8906	9605
$Z \sim WEI(1, \sqrt{0.75})$	(0)	(14)			

The bold numbers in the table are the lowest ARL values for detecting the simulated outbreaks

control charts, otherwise it is difficult to get acceptable false alarm rates (e.g., if we don't assume this knowledge, estimate the correlation between X_1 and X_2, and then the in-control false alarms are 66 on the high-side and 86 on the low side for the measure of data volume and 47 on the high-side and 31 on the low side for the measure of data perimeter). This can be considered the best case scenario when the appropriate transformation to normality is known to be log. For the log-normal case in Table 1, notice that the data depth measure was more efficient at detecting the out-of-control situations than the measures of data volume and data perimeter when the change occurs in the second variable only.

Table 4 The number of flags for increased (decreased) dispersion for the various approaches; Gamma distribution

$d(a, b)$	Box-Cox Transformation				
	0.32	0.225			
	Volume	Perimeter			
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)	Robust	Robust	
	7.5292	8.8335	regression (P_i)	regression (Q_i)	
$Ga(shape, rate)$	(2.6899)	(7.5664)	$h = 0.6425606$	$h_q = 0.5919475$	Data depth
$X \sim Ga(3, 2)$,	30	23	31	34	13
$Z \sim Ga(3, \sqrt{1.75})$	(9)	(9)			
$X \sim Ga(3, 1.5)$,	99	**1548**	584	476	128
$Z \sim Ga(3, \sqrt{1.75})$	(3)	(0)			
$X \sim Ga(3, 1)$,	422	**9466**	6418	4960	2590
$Z \sim Ga(3, \sqrt{1.75})$	(0)	(0)			
$X \sim Ga(3, 0.8)$,	746	**9984**	9089	8158	5429
$Z \sim Ga(3, \sqrt{1.75})$	(0)	(0)			
$X \sim Ga(3, 0.6)$,	1211	**10,000**	9947	9785	9127
$Z \sim Ga(3, \sqrt{1.75})$	(0)	(0)			
$X \sim Ga(3, 2)$,	94	142	202	**219**	33
$Z \sim Ga(3, \sqrt{1.25})$	(8)	(2)			
$X \sim Ga(3, 2)$,	175	432	606	**657**	152
$Z \sim Ga(3, 1)$	(1)	(0)			
$X \sim Ga(3, 2)$,	441	1370	**1989**	1960	718
$Z \sim Ga(3, \sqrt{0.75})$	(1)	(0)			
$X \sim Ga(3, 2)$,	1026	4226	**5354**	5132	2300
$Z \sim Ga(3, \sqrt{0.5})$	(0)	(0)			
$X \sim Ga(3, 2)$,	2689	9202	**9451**	9322	7615
$Z \sim Ga(3, 0.5)$	(0)	(14)			

The bold numbers in the table are the lowest ARL values for detecting the simulated outbreaks

We have only simulated out-of-control data with increased dispersion for the rational subgroup because this is the usual out-of-control case. However, we recognize that this does not convey the full value for the robust regression method or the approaches using the Box-Cox transformations which are capable of flagging reduced dispersion as well. Firstly, note that the Box-Cox transformation to normality plans can't always deliver a reasonable plan that adequately controls the in-control false alarm rates. For example, note that the Inverse Gamma and Pareto2 distributions fail to deliver reasonable plans based on normal approximations, i.e., Inverse Gamma plan has false alarm rates out of 10,000 trials for high-side (low-side) of 0(9512) for the volume (area) measure and 154(37) for the perimeter

Table 5 The number of flags for increased (decreased) dispersion for the various approaches; inverse Gamma distribution

$d(a, b)$	Box-Cox Transformation				
	-0.15	-0.2			
	Volume	Perimeter			
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)	Robust	Robust	
	2.3849	6.1934	regression (P_i)	regression (Q_i)	
$IGa(\mu, \sigma)$	(0.8289)	(4.4893)	$h = 740.709$	$h_q = 659.1923$	Data depth
$X \sim IGa(2, 1)$, $Z \sim IGa(2, \sqrt{0.75})$	0 (9512)	154 (37)	20	22	9
$X \sim IGa(2, 1.25)$, $Z \sim IGa(2, \sqrt{0.75})$	Plan is not adequate		**1143**	1114	271
$X \sim IGa(2, 1.5)$, $Z \sim IGa(2, \sqrt{0.75})$			5331	**5403**	2365
$X \sim IGa(2, 1.75)$, $Z \sim IGa(2, \sqrt{0.75})$			**8813**	8767	5636
$X \sim IGa(2, 2)$, $Z \sim IGa(2, \sqrt{0.75})$			9815	**9819**	8023
$X \sim IGa(2, 2.25)$, $Z \sim IGa(2, \sqrt{0.75})$			9973	**9977**	9408
$X \sim IGa(2, 1)$, $Z \sim IGa(2, \sqrt{1.25})$			251	192	**600**
$X \sim IGa(2, 1)$, $Z \sim IGa(2, \sqrt{1.5})$			832	451	**1732**
$X \sim IGa(2, 1)$, $Z \sim IGa(2, \sqrt{1.75})$			1921	1027	**3271**
$X \sim IGa(2, 1)$, $Z \sim IGa(2, \sqrt{2})$			3504	2873	**4653**
$X \sim IGa(2, 1)$, $Z \sim IGa(2, \sqrt{2.25})$			5085	4363	**5856**
$X \sim IGa(2, 1)$, $Z \sim IGa(2, \sqrt{2.5})$			6434	5894	**6932**

The bold numbers in the table are the lowest ARL values for detecting the simulated outbreaks

measures, respectively. While the Pareto2 distribution example has 495(125) for the volume (area) measure and 320(32) for the perimeter measure when we are aiming for (14)14. Therefore, these plans do not always provide a solution and for this reason it is not recommended as a routinely acceptable approach. However, it may have merit if it is known that the transformation to normality is appropriate as is the case for log-normal data.

Table 6 The number of flags for increased (decreased) dispersion for the various approaches; Pareto distribution

$d(a, b)$	Box-Cox Transformation				
	log	log			
	Volume	Perimeter			
	thresholds				
	Upper	Upper			
$PARETO2 = Pa$	(Lower)	(Lower)	Robust	Robust	
	76.9411	19.9683	regression (P_i)	regression (Q_i)	
$Pa(\mu, \sigma)$	(20.0488)	(9.7517)	$h = 74,930.24$	$h_q = 39,975.8$	Data depth
$X \sim Pa(2, 1)$, $Z \sim Pa(2, \sqrt{0.5})$	495 (125)	320 (32)	25	33	6
$X \sim Pa(2, 0.7)$, $Z \sim Pa(2, \sqrt{0.5})$	Plan is not adequate		75	**110**	80
$X \sim Pa(2, 0.6)$, $Z \sim Pa(2, \sqrt{0.5})$			265	278	**341**
$X \sim Pa(2, 0.5)$, $Z \sim Pa(2, \sqrt{0.5})$			964	1004	**1165**
$X \sim Pa(2, 0.4)$, $Z \sim Pa(2, \sqrt{0.5})$			2572	3003	**3152**
$X \sim Pa(2, 0.3)$, $Z \sim Pa(2, \sqrt{0.5})$			6433	**6862**	6294
$X \sim Pa(2, 1)$, $Z \sim Pa(2, \sqrt{0.3})$			**277**	250	29
$X \sim Pa(2, 1)$, $Z \sim Pa(2, \sqrt{0.2})$			1128	**1140**	118
$X \sim Pa(2, 1)$, $Z \sim Pa(2, \sqrt{0.1})$			5097	**5106**	1510

The bold numbers in the table are the lowest ARL values for detecting the simulated outbreaks

The robust regression and depth measures seem to be good alternatives. Note that when the change is consistent with the correlation structure (i.e., the change is in the X variable), then generally the robust regression methods is more efficient. While if it is in-consistent with the correlation structure by the change being in the Z variable (and therefore only in variable X_2), then the data depth is generally more efficient. The Pareto and Reverse Gumbel cases are exceptions to this rule. It seems as if a robust plan should involve a combination of depth and robust regression. The advantages this has, besides delivering a robust plan, are: firstly it will differentiate between a location shift and an increase in dispersion, and secondly it will flag decreases in dispersion via the robust regression method. There is not much of a difference between the two robust regression plans but the P statistic appears to have the slight edge over the more traditional quadratic form. Although more work is needed on this topic, but the early signs are that depth measures and robust regression methods are worth further investigations in follow-up research.

Table 7 The number of flags for increased (decreased) dispersion for the various approaches; reverse Gumbel distribution

$d(a, b)$	Box-Cox Transformation				
	−2.68	−2.61			
	Volume	Perimeter			
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)	Robust	Robust	
	1.2774	0.4498	regression (P_i)	regression (Q_i)	
$RG(\mu, \sigma)$	(0.4452)	(0.3397)	$h = 0.7894239$	$h_q = 0.6935926$	Data depth
$X \sim RG(13, 1)$, $Z \sim RG(13, \sqrt{1.5})$	9 (14)	33 (23)	27	28	8
$X \sim RG(13, 1.5)$, $Z \sim RG(13, \sqrt{1.5})$	363 (1)	668 (133)	**1752**	1277	1201
$X \sim RG(13, 2)$, $Z \sim RG(13, \sqrt{1.5})$	2380 (2)	2467 (216)	**6631**	5804	5307
$X \sim RG(13, 3)$, $Z \sim RG(13, \sqrt{1.5})$	7423 (1)	6521 (189)	**9818**	9661	9072
$X \sim RG(13, 4)$, $Z \sim RG(13, \sqrt{1.5})$	9091 (1)	8442 (139)	**9995**	9986	9879
$X \sim RG(13, 5)$, $Z \sim RG(13, \sqrt{1.5})$	9577 (2)	9204 (130)	**10,000**	9999	9987
$X \sim RG(13, 1)$, $Z \sim RG(13, \sqrt{3})$	386 (0)	105 (5)	**832**	752	682
$X \sim RG(13, 1)$, $Z \sim RG(13, \sqrt{4})$	1309 (0)	188 (5)	**2301**	2042	2161
$X \sim RG(13, 1)$, $Z \sim RG(13, \sqrt{6})$	4216 (0)	489 (5)	**5734**	5065	5132
$X \sim RG(13, 1)$, $Z \sim RG(13, \sqrt{8})$	6573 (0)	1052 (1)	**8026**	7412	6854
$X \sim RG(13, 1)$, $Z \sim RG(13, \sqrt{10})$	8010 (0)	1693 (2)	**9074**	8641	7999

The bold numbers in the table are the lowest ARL values for detecting the simulated outbreaks

3 Example of Application

An application of bivariate control charts is in effluent monitoring of non-filterable residues (NFR) and total residual chlorine (TRC) at sewerage treatment plants. These are typically not normally distributed (e.g., Park 2007), and are routinely monitored over time at all treatment plants. The data we have involves daily measures of NFR and TRC from 1 July 1996 to 24 June 1999. A scatterplot of the data is provided in Fig. 1.

The number of observations in this dataset is not quite sufficient for both Phase I and Phase II SPC, therefore we took the first half of the data and fitted a best

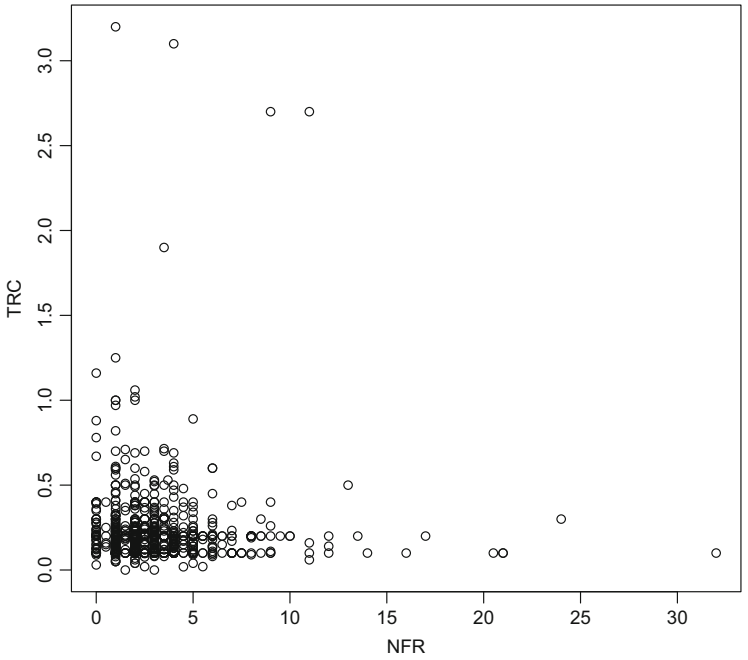


Fig. 1 Scatterplot of the sewerage treatment plant data

Box-Cox transform of the data to normality and then used a parametric bootstrap approach to set up the Phase II SPC process, and applied this to the second half of the data. The best transform for FCR was the value plus 0.11 inverted, and the best transform of TRC is the logarithm of the total of this measure plus one. The correlation between these two variables is not high at 0.05184, but nevertheless they are positively correlated. We construct bootstrap samples of the transformed TRC using $1/(x + 0.11) + 0.1855z \sim n(3.65, 1.397)$ and $z \sim n(0, 1)$ and the transformed NFR using $log(y + 1) + 0.1855z \sim n(1.233, 0.316)$ and the small positive correlation is induced by the normally distributed variable $z \sim n(0, 1)$. The autocorrelation for NFR is 0.279 and TRC is 0.563. The autocorrelation and partial autocorrelation functions suggested an AR(1) model for both transformed NFR and TRC. Therefore, we simulated the process $x_1 = log(y+1) \sim n(1.233 = 0.889/(1-0.279), 0.316-0.1855^2 = 0.2814/(1-0.279^2))$ and $x_2 = 1/(x+0.11) \sim n(3.65 = 1.5987/(1-0.562), 1.397-0.1855^2 = 0.9322/(1-0.562^2))$. Similarly, this means that we simulate $x_{2t} = 1.5987 + 0.562 \times x_{2t} + e_{1t}$ where $e_t \sim N(0, 0.9322)$ and $x_{1t} = 0.889 + 0.279 \times x_{1t} + e_{2t}$ where $e_t \sim N(0, 0.2814)$. Now $0.1855z$ is added to both of these and then simulated NFR and TRC values are taken as $exp(x_{1t} - 1$ and

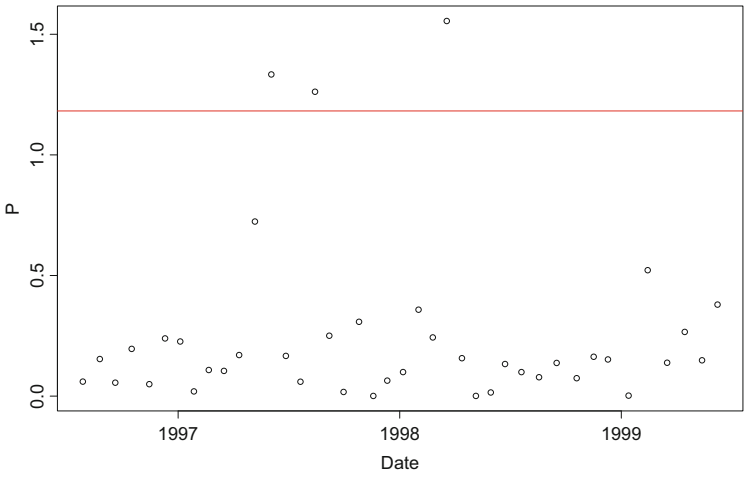


Fig. 2 Bivariate application flagging changes in dispersion of NFR and TRC using the robust regression method

$1/x_{2t} - 0.11$, respectively. These are used to find the thresholds for the dispersion monitoring plan using a parametric bootstrap approach.

The boostrap population $NFR = exp(x_{1t}) - 1$ and $TRC = 1/x_{2t} - 0.11$ is used to set up simulated data for training the bivariate process control charts for the second half of the data. This bootstrap sample indicated 279 false alarm signals in 10,000 in-control bootstrap samples for depth (this is a higher false alarm rate than we would like). The two robust regression procedures lead to identical conclusions and therefore only one is reported. In the robust regression cases we were able to train the methods to have a false alarm rate of 0.0027. The results are reported in Fig. 2. Notice that the robust regression approach only flags a change in dispersion for rational subgroup for data starting on 1998-06-18, whereas data depth flags whenever 2 or more depths are zero in the rational subgroup (equal to or above the depth red line in Fig. 3) and when two consecutive rational subgroups with exact one depth equal to zero (these are indicated by placing a cross at both the locations this occurs). In Fig. 4,

$$T = \text{tr}(\mathbb{S})/(\text{tr}(\Sigma_0) + h_{tr,upper}) \quad \text{and} \quad D = |\mathbb{S}|/(|\Sigma_0| + h_{d,upper})$$

So we are only signaling increases in variance.

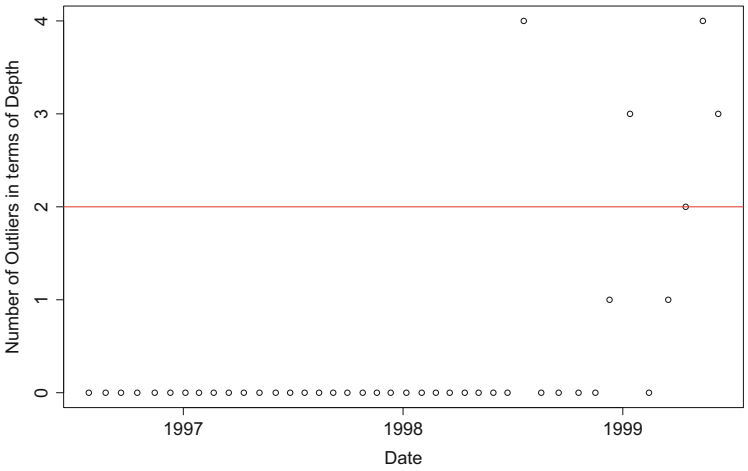


Fig. 3 Bivariate application flagging outliers for NFR and TRC using data depth method

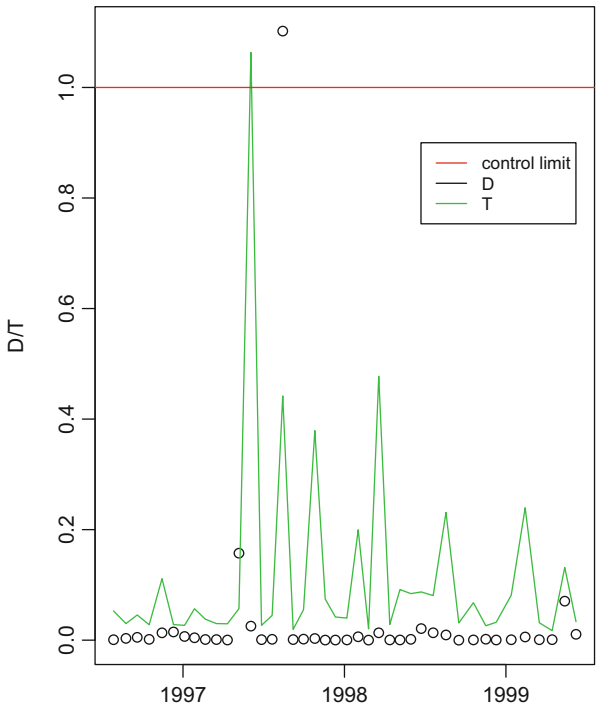


Fig. 4 Bivariate application flagging changes in dispersion of NFR and TRC using data BoxCox transformation to normality Volume and Perimeter measures

4 Concluding Remarks

The distribution-free method proposed in this chapter based on ranks does not work as well as the plan based on robust regression methods.

The biggest disadvantage with the robust regression approach is that many more numbers of rational subgroup samples are needed in Phase I to set-up this plan for Phase II monitoring. Although in many environmental settings data have been collected for decades and in several applications such data would be sufficient to establish the plan and in these cases data availability should not be a restriction. This is certainly the case in the Sydney Waterways. If we train the methods for a false alarm rate of 1 in 100, then we could get away with smaller samples in the Phase I stage, and so more work is needed in establishing the Phase I information needed to effectively design the robust regression plan. The other advantage that the robust regression has is that it distinguishes between increases in spread and decreases in spread, whereas data depth can't easily find reductions in dispersion. In addition, data depth will flag changes in location and therefore does not distinguish between changes in location and spread, but the robust regression approach does. In terms of their performance in detecting changes in dispersion quickly, there is little difference between the approaches, with the differences mostly being small except in the cases of the Inverse Gaussian distribution and the Inverse Gamma distribution. So the choice of which method to apply is going to depend on the individual application.

The relative performance for the robust regression is encouraging and therefore this plan is worth further investigation in settings that don't only involve positive measures. If users wish their monitoring plan to separate out the parameter influences on the process measures, then selecting the appropriate scale is important. This is demonstrated by the log-normal distribution where the log-scale applying the S-chart only flag changes in variance but not location.

References

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B, 26*, 211–252.

Cheng, S. W., & Xie, H. (2000). Control charts for lognormal data. *Tamkang Journal of Science and Engineering, 3*(3), 131–138.

Choi, D. J., & Park, H. (2001). A hybrid artificial neural network as a software sensor for optimal control of a wastewater treatment process. *Water Research, 35*(16), 3959–3967.

Dodds, W. K., Jones, J. R., & Welch, E. B. (1998). Suggested classification of stream trophic state: Distributions of temperate stream types by chlorophyll, total nitrogen, and phosphorus. *Water Research, 32*(5), 1455–1462.

Ferrell, E. B. (1958). Control charts for lognormal universe. *Industrial Quality Control*, **15**, 4–6.

Hamed, M. M., Khalafallah, M. G., & Hassanien, E. A. (2004). Prediction of wastewater treatment plant performance using artificial neural networks. *Environmental Modelling & Software, 19*(10), 919–928.

Joffe, A. D., & Sichel, H. S. (1968). A chart for sequentially testing observed arithmetic means from lognormal populations against a given standard. *Technometrics, 10*(3), 605–612.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Annals of Statistics, 18*(1), 405–414.

Liu, R. Y., Parelius, J. M., & Singh, K. (1999), Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). *Annals of Statistics, 27*, 783–858.

Morrison, J. (1958). The lognormal distribution in quality control. *Applied Statistics, 7*, 160–172.

Park, G.S. (2007). The role and distribution of total suspended solids in the macrotidal coastal waters of Korea. *Environmental Monitoring and Assessment, 135*, 153–162.

Saby, S., Djafer, M., & Chen, G. H. (2002). Feasibility of using a chlorination step to reduce excess sludge in activated sludge process. *Water Research, 36*(3), 656–666.

Sparks, R. S. (1992). Quality control with multivariate data. *Australian Journal of Statistics, 34*(3), 375–390.

Stoumbos, Z. G., Jones, L. A., Woodall, W. H., & Reynolds, M. R. (2001). On nonparametric multivariate control charts based on data depth. *Frontiers in Statistical Quality Control, 6*, 207–227.

Sukumar, R., Dattaraja, H. S., Suresh, H. S., Radhakrishnan, J., Vasudeva, R., Nirmala, S., & Joshi, N. V. (1992). Long-term monitoring of vegetation in a tropical deciduous forest in Mudumalai, southern India. *Current Science, 62*(9), 608–616.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Berlin: Springer.

Monitoring and Diagnosis of Causal Relationships Among Variables

Ken Nishina, Hironobu Kawamura, Kosuke Okamoto, and Tatsuya Takahashi

Abstract In statistical process control (SPC) there are two situations where monitoring multivariate is needed. One is that all of the variables monitored are product ones. The other is that the variables monitored are some product and process ones. In these cases, there are correlations among the variables. Therefore, application of multivariate control charts to such process control is useful.

In this chapter, the latter case of monitoring causality is addressed. It is known that T^2–Q control charts, which are modified from standard multivariate control charts utilizing Mahalanobis distance, are an effective SPC tool. However, in using multivariate control charts, diagnosis is not so easy. The objective in this chapter is to propose a diagnostic method for identifying an unusual causal relationship in a process causal model and then to examine its performance.

Our proposed method is to identify the nearest unusual model by utilizing the Mahalanobis distance between some supposed unusual models and the data indicating the out of control in Q charts.

Keywords Statistical process control · T^2–Q control charts · Unusual causal relationship · Mahalanobis distance

1 Introduction

In statistical process control (SPC) there are two situations where monitoring multivariate is needed.

One is that all of the variables monitored are product ones; for example, the remaining film thickness on the wafer surface after polishing in chemical mechanical polish process of semiconductor manufacturing process (see Nishina

K. Nishina (✉) · H. Kawamura · K. Okamoto · T. Takahashi
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Japan
e-mail: nishina@nitech.ac.jp

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_10

et al. 2011). In this case, correlations among the variables are strongly positive. Therefore, applying multivariate control charts to such process control is useful.

The other is that the variables monitored are some process and product ones; for example, some equipment parameters and product characteristics are monitored. In this case, it can be supposed that monitoring causality among the variables is needed. A case in which an environmental variable, which has an interaction with an equipment parameter, is suddenly varied can be illustrated as an unusual causal relationship. Another example is to lose control completely by a cyberattack. Applying multivariate control charts is also useful.

In this chapter, the latter case of monitoring causality is addressed. It is known that T^2–Q control charts, which are modified from standard multivariate control charts utilizing Mahalanobis distance, are an effective SPC tool (see Jackson and Mudholkar 1979).

The causal model consists of variables and causal relationships between the variables. In using multivariate control charts, diagnosis is not so easy because an unusual event may affect more than one variable. Moreover, if an unusual event may affect the causal relationship as mentioned above, it is more difficult to isolate the source of the causal unusualness. The objective of this chapter is to propose a method of diagnosis for isolating an unusual causal relationship in a process causal model and then to examine its performance.

Our proposed method is to identify the nearest unusual model by utilizing the Mahalanobis distance between some supposed unusual models and the data indicating the out of control in Q charts.

In our proposal, an unusual variable is isolated in the first step to narrow down the unusual causal relationship and then in the second step the unusual causal relationship is isolated. Kourti and MacGregor (1996) proposed a diagnostic method, called contribution plots, to isolate the unusual variable. Higashide et al. (2014) made slight improvement on the method. Another method is diagnosis by the MT (Mahalanobis–Taguchi) method (see Tatebayashi et al. 2008), which is a variable selection by using the two levels orthogonal array.

In our study, it is assumed that the causal relationship of the variables in the manufacturing process is known by causal analysis.

2 Outline of T^2–Q Control Charts and Their Application

T^2–Q control charts are modifications of the multivariate control charts using Mahalanobis distance. The statistic T^2, which is the Mahalanobis distance, is composed of major Principal Component Scores (PCSs). On the other hand, the statistic Q, which is the Euclidean distance, is composed of minor PCSs.

It is well known that the Mahalanobis distance D^2 in the p variables can be expressed as PCS z_k ($k = 1, 2, \ldots, p$) in Eq. (1).

$$D^2 = (1/\lambda_1)z_1^2 + (1/\lambda_2)z_2^2 + \cdots + (1/\lambda_m)z_m^2 + \cdots + (1/\lambda_p)z_p^2 \,, \tag{1}$$

where λ_k ($k = 1, 2, \ldots, p$) is the kth eigenvalue of the correlation coefficient matrix. The T^2 and Q statistic are modified slightly from the decomposition as shown in Eq. (1).

$$T_i^2 = \sum_{k=1}^{m} (1/\lambda_k) z_{ik}^2 \tag{2}$$

$$Q_i = \sum_{k=m+1}^{p} z_{ik}^2 \tag{3}$$

Decomposition of the Mahalanobis distance has a statistical meaning. Consider Eq. (1). In the Mahalanobis distance D^2, each minor PCS is divided by the much smaller eigenvalue, respectively. However, the much smaller eigenvalues are not so precise. This can lead to a much greater increment of type I error. On the other hand, the Q statistic is not affected by the much smaller eigenvalues because it is not Mahalanobis distance but Euclidean distance (see Nishina et al. 2011). Especially, when the number of variables becomes very large, the Mahalanobis distance D^2 faces singularity problems. T^2–Q control charts overcome this problem (see Kourti 2005).

Similarly, decomposition of Mahalanobis distance D^2 has a practical meaning. The T^2 and Q statistic have different roles in the process control, respectively. As mentioned earlier, the T^2 statistic consists of major PCS. This means that the T^2 statistic can monitor usual process variation. Out of control in T^2 charts indicates that the usual process variation becomes large; however, at that time the correlative structure does not change. On the other hand, the Q statistic can monitor unusual process variation. Out of control in Q charts indicates that the correlative structure changes. For example, process variation due to a parts deterioration is a usual variation. Such a variation is monitored by T^2 chart. Q charts have a role to control other miscellaneous factors, which may make the correlative structure change.

In this chapter, we focus on monitoring and diagnosis of causal relationships among variables. Therefore, in discussing hereafter, Q charts have an important role. In the simulation study of this chapter, m is determined as follows:

$$m = \arg\min_k \{\lambda_k - 1.0 \mid \lambda_k \geq 1.0\}.$$

The control lines (the control limits and the center line) of T^2–Q control charts are given as follows:

Control limits of T^2 charts:

$$UCL_\alpha = \frac{m(n+1)(n-1)}{n(n-m)} F_\alpha(m, n-m)$$

where $F_\alpha(\phi_1, \phi_2)$ is the upper $100\alpha\%$ percentile point of F distribution with (ϕ_1, ϕ_2) degrees of freedom and n is the sample size.

Center line of T^2 charts:

$$CL = \frac{m(n+1)(n-1)}{n(n-m-2)}.$$

The statistic Q can be approximated to the standard normal distribution by transforming to the statistic c as follows (see Jackson and Mudholkar 1979):

$$c = \frac{\theta_1\left[(Q/\theta_1)^{h_0} - 1 - \{\theta_2 h_0(h_0-1)/\theta_1^2\}\right]}{\sqrt{2\theta_2 h_0^2}},$$

$$\theta_i = \sum_{r=m+1}^{p} \lambda_r^i \quad (i = 1, 2, 3), \quad h_0 = 1 - \left(2\theta_1\theta_3/3\theta_2^2\right).$$

Therefore, the control limits of Q charts using c statistic are obtained the same as X control charts.

3 Proposals on Diagnosis

3.1 Isolation of the Unusual Variable

As mentioned earlier, the first step in the diagnosis of the source of causal unusualness is to narrow down the unusual variables. We evaluate two methods, that is, the contribution plots by Kourti and MacGregor (1996) and the MT method by Tatebayashi et al. (2008).

3.1.1 Modified Contribution Plots

The contribution plots can be extracted from the underlying PCA model. As shown in Eq. (3), the statistic Q consists of PCSs. The kth PCS of the ith sample (t_{ik}) can be decomposed as follows:

$$t_{ik} = w_{k1}x_{i1} + w_{k2}x_{i2} + \cdots + w_{kp}x_{ip},$$

where x_j ($j = 1, 2, \ldots, p$) is the jth centralized (or normalized) variable and w_{kj} ($k = 1, 2, \ldots, p$) is the element of eigenvector corresponding to the kth largest eigenvalue λ_k. Therefore, the original contribution of the variable x_j to the statistic Q can be measured as shown in Eq. (4) (see Kourti and MacGregor 1996).

$$\sum_{r=m+1}^{p} (w_{rj}x_j)^2 \quad (j = 1, 2, \ldots, p) \tag{4}$$

Higashide et al. (2014) gave slight modification for the original contribution plots as shown in Eq. (5).

$$C_j = \sum_{r=m+1}^{p} \left\{ I(r) w_{rj} x_j \right\}^2 \tag{5}$$

$$I(r) = \begin{cases} 1 & \text{if } \operatorname{sgn}(t_r) = \operatorname{sgn}(w_{rj} x_j) \\ 0 & \text{if } \operatorname{sgn}(t_r) \neq \operatorname{sgn}(w_{rj} x_j) \end{cases} \tag{6}$$

Equation (6) shows an essential point of the modification. This means that the degree to contribution of x_j, which is responsible for making the absolute value $|t_k|$ large, is inflated.

3.1.2 Diagnosis of Variables by MT System

The diagnosis of variables by the MT (Mahalanobis–Taguchi) system has been originally utilized as a method for selecting the variables so as to detect an unusual condition with more sensitivity. In the variable diagnosis the method is utilized to narrow down the unusual variable.

In this method, the orthogonal array with two factor levels is used. The candidate variables are assigned on each column; for example, in the case of using L_8 orthogonal array and lining up four candidate variables x_1, x_2, x_3 and x_4 an assignment is shown in Table 1. The level-0 means that the variable concerned is deleted and the level-1 means vice versa; for example, the causal model supposed in No. 7 experiment is that the variable x_1 and x_2 are retained but x_3 and x_4 are deleted. However, we regard the average of the result of other seven experiments as the result of No. 1 experiment.

The response is the Mahalanobis distance between the average of the usual dataset and the ith sample, which indicates the out of control, as follows:

$$D_i = (\boldsymbol{x}_{i(J)} - \bar{\boldsymbol{x}}_{(J)})' \, \boldsymbol{S}_{(J)}^{-1} \, (\boldsymbol{x}_{i(J)} - \bar{\boldsymbol{x}}_{(J)})$$

Table 1 Assignment to L_8 orthogonal array for diagnosis of variables

No.	x_1	x_2	x_3	x_4
1	0	0	0	0
2	0	0	1	1
3	0	1	0	1
4	0	1	1	0
5	1	0	0	1
6	1	0	1	0
7	1	1	0	0
8	1	1	1	1

where $\boldsymbol{x}_{i(J)}$ and $\bar{\boldsymbol{x}}_{(J)}$ is the ith observation vector and the average of the data set from the usual process, respectively; in addition the suffix (J) stands for "without the J variable set corresponding to the level-0 in the orthogonal array." $\boldsymbol{S}_{(J)}$ is the submatrix of $\boldsymbol{S}$ (covariance matrix of the usual dataset) without the J variable set.

As the result of the factorial effects, the variable, which has the largest factorial effect, is regarded as the unusual variable of the effect side in the unusual causal relationship.

3.2 Diagnosis of Unusual Causal Relationship

In our proposal for diagnosis of unusual causal relationship, the fundamental analysis is the Mahalanobis distance between the average of the dataset under a supposed unusual causal model, $\bar{\boldsymbol{x}}_{(u)}$, and the ith sample $\boldsymbol{x}_i$, which indicates the out of control in Q chart:

$$D_u = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(u)})' \, \boldsymbol{S}_{(u)}^{-1} \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(u)}) \, . \tag{7}$$

In the preceding step the unusual variable have been already isolated. In the next step the unusual causal relationship should be isolated among the causal relationships, which have the arrow line indicating the causality contained in the unusual variable isolated in the preceding step. Figure 1 shows a causal model. In the case of Fig. 1, if the isolated unusual variable is X_4, then the causal relationships to become an unusual candidate are α_{41}, α_{42} and α_{43}.

Now let the supposed unusual causal relationships be u and let the path coefficient of the causal relationship be α. Based on Eq. (7), the following u^* can be determined. As the result, the causal relationship u^* is isolated, that is, the unusual model with the shortest Mahalanobis distance among the supposed unusual models is regarded as the unusual causal relationship.

$$u^* = \arg\min_{u} \left[\min_{\alpha} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(\alpha;u)})' \, \boldsymbol{S}_{(u)}(\alpha)^{-1} \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(\alpha;u)}) \right]$$

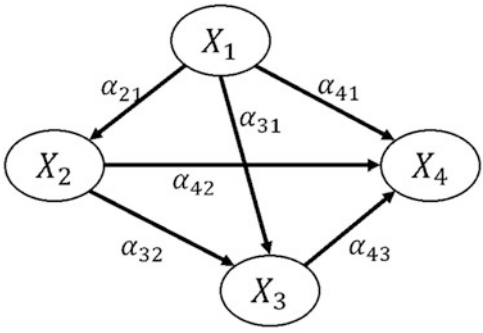


Fig. 1 An example of causal model

4 Examination of the Proposed Method by Simulation

4.1 Simulation Models and Simulation Experiments

We suppose the causal model shown in Fig. 1 again. The model is very simple, consisting of four variables; however, has the three kinds of causal relationships, which are the direct effect, indirect effect and the pseudo effect. The structural equations shown in Fig. 1 are as follows:

$$
\begin{aligned}
X_1 &= \varepsilon_1 \\
X_2 &= \alpha_{21} X_1 + \varepsilon_2 \\
X_3 &= \alpha_{31} X_1 + \alpha_{32} X_2 + \varepsilon_3 \\
X_4 &= \alpha_{41} X_1 + \alpha_{42} X_2 + \alpha_{43} X_3 + \varepsilon_4
\end{aligned}
$$

where αs and εs are path coefficients and random variables, respectively. Their variances, $Var(\varepsilon)$s, are determined so that $Var(X)$s form a unit. The random numbers are generated by NtRand of Mersenne twister. We suppose that one of the six paths in the model changes to an unusual situation.

We examine the proposed method in unusual cases of the two patterns shown in Table 2. As it is assumed that a unusual model has an unusual path coefficient, we suppose the 12 unusual models shown in Table 2; for example, one unusual model in the pattern 1 is that $\alpha_{21} = -2.1$, $\alpha_{31} = \alpha_{32} = \alpha_{41} = \alpha_{42} = \alpha_{43} = 0.4$. The pass coefficients of the unusual models in Table 2 are set to be ARL $\fallingdotseq$ 4.0 of the Q chart.

Our simulation study is carried out as follows: the sample size for determining the control limit, which is shown in Sect. 2, is 200. After constructing the control limit, 200 data under a unusual model are generated. Whenever Q chart indicates out of control, the unusual variable is isolated and then the unusual causal relationship is isolated. This is a simulation set. The set is carried out in 100 trials.

Table 2 Unusual models in our simulation study

Path coefficient	Pattern 1		Pattern 2	
	Usual model	Unusual model	Usual model	Unusual model
α_{21}	0.4	−2.1	0.6	−2.2
α_{31}	0.4	−2.1	0.7	−1
α_{32}	0.4	−2.1	0.4	−1.4
α_{41}	0.4	−2.2	0.2	−2.7
α_{42}	0.4	−2.2	−0.8	2
α_{43}	0.4	−2.2	0.3	−2.9

Let C_i and D_i be the successful count of the isolation of the unusuality and the count of searching the unusuality in the ith set of simulation, respectively. The performance index, which is called the success rate hereafter, is

$$\sum_{i=1}^{100} \frac{C_i}{D_i} \tag{8}$$

where D_i is about 50.

4.2 Comparison of Methods of Isolating Unusual Variable

As described in Sect. 3.1, we introduce two methods for isolation of an unusual variable. One is the modified contribution plots and the other is the diagnosis of variables by MT system. In this section we compare the performance of the methods. The performance is measured as the success rate shown in Eq. (8).

The results of the simulation study (the success rate of the isolation) are shown in Table 3. Table 3 indicates that the large difference of the performance appears in two cases of pattern 2, in which α_{31} and α_{32} are unusual. The reason is that the contribution plots are based on the correlation coefficient matrix. As known well, the correlation does not necessarily represent the causality. Table 3 shows that the performance of the MT system is not necessarily superior to the modified contribution plots with all cases, but the MT system can overcome the weak point of the modified contribution plots. We choose the diagnosis of variables by the MT system.

Table 3 Success rate of the isolation of unusual variable

Unusual path coefficient	Pattern 1		Pattern 2	
	Modified contribution plots	MT system	Modified contribution plots	MT system
α_{21}	0.985	0.814	0.931	0.893
α_{31}	0.950	0.985	0.249	0.939
α_{32}	0.948	0.985	0.491	0.954
α_{41}	0.874	0.993	0.957	0.960
α_{42}	0.884	0.996	0.868	0.964
α_{43}	0.848	0.993	0.945	0.977

4.3 Performance of the Proposed Method

Based on the results of Sect. 4.2, we choose MT method as the method for isolating an unusual variable. Table 4 shows the performance of the proposed method. The performance index in Table 4 is the success rate of the isolation of the unusual causal relationship in the cases of the twelve unusual models shown in Table 2.

Table 4 indicates that in the case of pattern 1 the success rates of the proposed method are relatively high but the results of some models in the case of pattern 2 are not so high. We examine the difference of the correlation coefficient matrix between the usual condition and the unusual condition for an example with the unusual path coefficient α_{43}. The success rate of this case is lowest in all the unusual models shown in Table 2. Table 5 shows the difference between the correlation coefficient matrices.

Table 5 indicates that r_{14} (correlation coefficient between X_1 and X_4) is quite different between the usual and the unusual as well as r_{34}, although α_{43} is unusual. It should be noted that this introduces the low success rate of the isolation of unusual relationships. The procedure for proposed method consists of the two steps, the isolation of the unusual variable and the isolation of an unusual relationship. As shown in this case, the proposed method may not isolate an unusual relationship and may simply show the priority order of the search. Even if the success rate of the isolation of unusual relationship is not so high, the unusual variable can be isolated. It is a remarkable property. In practice, after isolating an unusual variable, a method to search for unusuality in the order of the path with the small value of the Eq. (9),

Table 4 Success rate of the isolation of unusual causal relationship

Unusual path coefficient	Pattern 1	Pattern 2
α_{21}	0.814	0.893
α_{31}	0.824	0.631
α_{32}	0.828	0.639
α_{41}	0.742	0.628
α_{42}	0.756	0.842
α_{43}	0.703	0.589

Table 5 Difference of the correlation coefficient matrices between the usual and the unusual conditions (upper: usual causality; lower: unusual causality)

	X_1	X_2	X_3	X_4
X_1	1.000	0.600	0.940	0.002
X_2	0.600	1.000	0.820	−0.434
X_3	0.940	0.820	1.000	−0.168
X_4	0.002	−0.434	−0.168	1.000
	X_1	X_2	X_3	X_4
X_1	1.000	0.600	0.940	−0.857
X_2	0.600	1.000	0.820	−0.871
X_3	0.940	0.820	1.000	−0.960
X_4	−0.857	−0.871	−0.960	1.000

Table 6 Success rate including the second candidate

Unusual path coefficient	Pattern 1	Pattern 2
α_{41}	0.954	0.935
α_{42}	0.958	0.913
α_{43}	0.955	0.962

the Mahalanobis distance, is recommended.

$$\min_{\alpha}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(u)})' \, \boldsymbol{S}(\alpha)^{-1} \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(u)}) \, . \tag{9}$$

Table 6 shows the success rate including the value to the second candidate by Eq. (9) in the cases of α_{41}, α_{42} and α_{43}. The values indicate high rate. The result means that the alternative method proposed above proposed is useful.

5 Conclusive Remarks

In this chapter, we have proposed the diagnosis method in applying the T^2–Q charts. In general, it is not easy to make a diagnosis even if the multi-variate control chart indicates an out of control signal. Some methods have been proposed but the aim of these methods is to isolate an unusual variable. In this chapter, we can propose the diagnosis method with the aim of isolating an unusual relationship using the Mahalanobis distance.

In near future, we will try to apply the our proposed method to the process control of the facilities collection process such as the semiconductor process.

Acknowledgement This work was supported by KAKENHI (25750120).

References

Higashide, M., Nishina, K., & Kawamura, H. (2014). A practice of T^2–Q control charts in semiconductor manufacturing process. *Quality, 44*(3), 77–86 (in Japanese).

Jackson, J. E., & Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics, 21*(3), 341–349.

Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing, 19*(4), 213–246.

Kourti, T., & MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology, 28*(4), 409–428.

Nishina, K., Higashide, M., Hasegawa, Y., Kawamura, H., & Ishii, N. (2011). A paradigm shift from monitoring the amount of variation into monitoring the pattern of variation in SPC. In *Proceedings of ANQ Congress Ho Chi Minh City 2011*, Vietnam.

Tatebayashi, K., Teshima, M., & Hasegawa, Y. (2008). *Nyumon MT System*, Nikkagiren (in Japanese).

Statistical Monitoring of Multi-Stage Processes

Emmanuel Yashchin

Abstract In many complex processes, such as semiconductor manufacturing or production of mass storage systems, a large number of variables are monitored simultaneously. These variables can typically be impacted by several points of the manufacturing process, necessitating efforts that include not only monitoring but also diagnostics involving establishing change-points, regimes and potential stages of influence. We discuss statistical methods used to handle such multi-stage data and give examples of applying these methods in large-scale monitoring systems.

Keywords Average run length · Detection · False alarms · Statistical process control

1 Introduction

In today's applications of statistical process control one typically needs to handle large amounts of information produced by complex industrial and business processes. A number of publications address this aspect of process control, emphasizing the need for advanced statistical techniques. For example, Capizzi (2015) discusses the use of control charts in conjunction with variable selection methods. Woodall and Montgomery (2014) discuss current methods and directions in conjunction with large-scale monitoring applications, such as profile monitoring, health metrics monitoring and spatiotemporal analysis. Shmueli and Burkom (2010) discuss methods for detecting early outbreak of epidemics. Sparks (2015) discusses methods of detecting changes in communication rates between parties of interest in social networks. Duchesne et al. (2012) discuss the problem of image analysis and monitoring in process industries. Golosnoy et al. (2011) discuss the problem of detecting changes in weights of financial portfolio components. Yashchin (2012)

E. Yashchin (✉)
IBM, Thomas J. Watson Research Center, Yorktown Heights, NY, USA
e-mail: yashchi@us.ibm.com

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_11

describes a system for monitoring lifetime data and warranty data. Hryniewicz and Kaczmarek (2016) describe an approach to monitoring short series of dependent observations.

Every practical problem related to monitoring tends to have its own data setup, models, analytic/graphical component and decision framework. The common thread, however, is that methodologies that have proven statistical power tend to be of special value in large-scale applications, since achieving a high signal-to-noise ratio (or false alarm/sensitivity tradeoff) becomes imperative.

In this chapter, we discuss a large-scale monitoring system for multi-stage processes. Consider, for example, a semiconductor manufacturing line. Over the course of manufacturing (that can last several months), chips are processed as part of wafers—only in the last phase of the process are wafers diced into individual chips. Wafers are typically processed as parts of a lot (about 25 wafers per lot), and lots begin their journey as a collection of raw wafers (basically, silicon platters about 450 mm in diameter) and go through hundreds of operations, such as reactive ion etch (RIE), semiconductor device building stages, oxide deposition, chemical-mechanical planarization (CMP), rinsing, wiring, and multiple measuring and testing steps. The objective of the process is to make sure that the product characteristics (speed, reliability, thermal performance) are satisfactory and that the yields are acceptably high.

In the multi-stage process, we collect very large amounts of data related to various process stages, and one of the key issues is how to use this data to detect unfavorable trends. For example, consider the clock speed of chips and suppose that the target is 4 GHz. Many of the factors affecting speed are related to early stages of manufacturing and device formation. However, measurements of chip clock speeds are typically done at later stages; suppose that at some point we detected that the prevailing clock speed is 3.5 GHz. At this point, some defective product is likely to be present in the pipeline—so, it is very important to detect such a change as early as possible. Generally, we need more than just detection, as a number of early stages could be responsible for the drop in performance: we need methodology that will help us diagnose the stage that is the most likely culprit. The concept of timeslide analysis discussed in the chapter is the key instrument used in the problem of detection and diagnostics. The proposed methodology can be used in conjunction with other approaches and models for multi-stage processes, e.g. those described in Shi and Zhou (2009).

In the chapter, we will also describe a system for analysis of multi-stage process data, named QEWSV (Quality Early Warning System for Variables data). This system can be viewed as a kind of a *search engine* that sifts through the data on a periodic basis and selects stages and operations that merit engineering attention. In Sect. 2 we introduce the main components of the data setup. In Sect. 3 we discuss the basic system inputs and outputs. In Sect. 4 we discuss detection algorithms. In Sect. 5 we discuss some of the alarm attributes produced by the search engine. In Sect. 6 we discuss several operational and implementation issues.

2 Variables, Operations and Timeslides

The QEWSV system is organized around three basic notions that are referred to as *Variables, Operations* and *Timeslides*. The behavior of selected variables is monitored based on acceptable and unacceptable characteristics of the underlying process. These characteristics are converted to rules of the decision-making scheme used to decide whether a variable is "flagged" or not. Operations refer to points in the process that have a potential of influencing the stochastic behavior of the variables and are thus considered as prime suspects in cases where variables show unacceptable behavior. Timeslides are data structures that organize measurements pertaining to a particular variable with respect to a given operation of interest.

2.1 Variables

In many applications, variables correspond to measurements taken at particular points in the process (e.g., film thickness measurements in a process of semiconductor manufacturing). These variables can be continuous, discrete or mixed—the key assumption in QEWSV is that the variables are univariate. We will refer to these variables in terms of a character string that consists of a letter "v" and a set of four digits, e.g. v0001. The variables themselves can relate to process means, standard deviations, quantiles, multivariate characteristics (e.g., correlation coefficients or principal components), percentages of defective items and so forth. Associated with a variable is a name represented by an alphanumeric string; in what follows, this name is referred to as "`Meas_Name`". Furthermore, associated with a variable is a set of characteristics that establish what constitutes acceptable or unacceptable stochastic behavior, as well as additional quantities that are used in decisions on whether this variable is to be flagged. Variables typically correspond to Items that are measured. For example, a measurement of film thickness is usually taken on a semiconductor wafer—so, items corresponding to this particular variable are wafers.

In this chapter, we assume that the variables are constructed in a way that the Cusum-Shewhart procedure for detection of unfavorable changes can be relied upon to provide sufficient statistical power when detecting a change in the variable's mean from its acceptable region to the unacceptable region. This assumption holds, for example, for independent and identically distributed (iid) random variables belonging to the exponential family of distributions, by virtue of linearity of statistics corresponding to likelihood ratio tests, see Moustakides (1986). In most practical situations, one can transform variables in such a way that the above assumption is satisfied. Even when variables show serial correlation, it is often possible to transform them (or define them) in a way that this assumption holds. In what follows we assume, that all the necessary data pre-processing has been done and thus application of the straight Cusum-Shewhart detection scheme is justified.

2.2 Operations

In a typical data generating process, such as manufacturing process, variables are associated with points in the process that might influence their behavior. We refer to these points as "Operations". For example, consider the process of semiconductor manufacturing and a variable "`Avg_Speed`" that measures the average speed of a set of chips related to a given production lot. There are a number of points in the manufacturing process where "`Avg_Speed`" can be affected: e.g., at the point where a particular insulating film is deposited, at the point where metal wires connecting circuits are produced—and even at the point where the measurements of speed are taken (malfunctions of measurement systems are a real possibility in many processes). Our focus is on the set of Operations that can potentially affect "`Avg_Speed`"—this set must be pre-specified. In some cases, the set could cover hundreds of physical operations or measurement processes—but typically there will be just a few operations of real interest for any given variable under consideration.

A given Operation is typically associated with a set of Tools that also play a role in QEWSV processing. For example, the aforementioned process of depositing an insulating film can be performed in one of several tools or chambers of tools operating in parallel. For a given value of a Variable and a given Operation, it is assumed that one (and only one) of the Tools was associated with this value. The value of a Variable is related to an Item on which the measurement was taken—and this Item could be processed no more than by a single tool during a particular Operation.

When an Item is processed by a Tool in a course of an Operation, we record a timestamp that plays a key role in QEWSV processing. The timestamp corresponds to a moment in time, recorded up to the desired degree of accuracy. Notice that the timestamp has nothing to do with variables—it is a characteristic of an item with respect to a particular tool performing the operation.

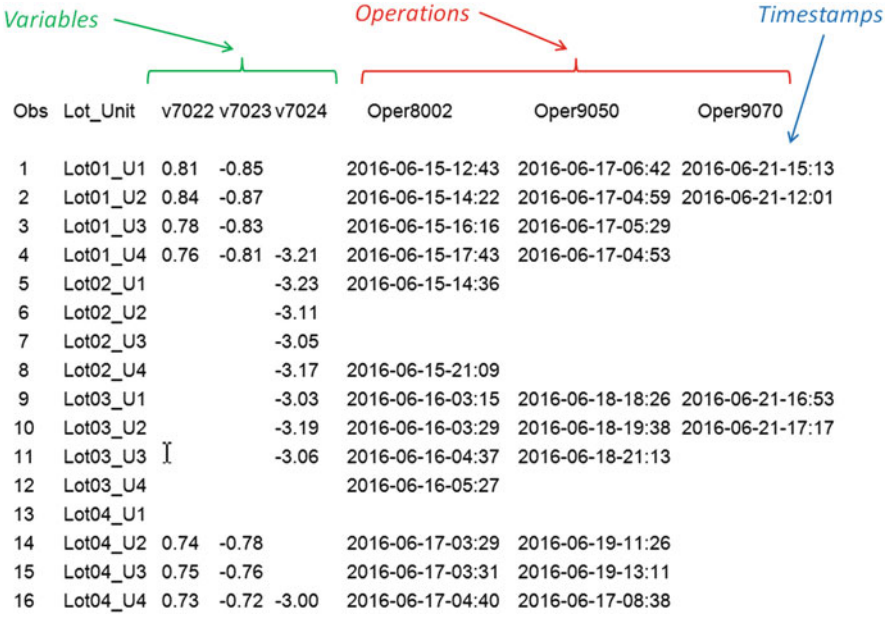
2.3 Timeslides

Consider a given Variable and a set of Operations that are deemed to be of potential importance with respect to it. For every Operation, we can order the values of this Variable in accordance with timestamps pertaining to Items recorded for Tools of this Operation. This type of an ordering is called a Timeslide. The main value of a Timeslide is related to its ability to assist in detection of changes, visualization of these changes and diagnostics. For example, if a particular tool has a negative impact on a given variable, and we order the values of the variable in accordance with processing time by the said tool, then we are likely to see some patterns. In particular, one could find that all variable values corresponding to items processed prior to a certain point in time are much higher than values corresponding to items processed later. This type of problem signature can be detected through QEWSV

analysis, leading to a flagged condition. Timeslides are typically tool-specific; it is quite possible that only one of the tools of a given Operation is causing problems—and we will rely on a Timeslide for this tool to detect that. If a particular operation is unrelated to changes in the variable of interest, then the corresponding Timeslides are not likely to show any signatures, and thus these Timeslides are not likely to be prioritized highly by the QEWSV analysis. Our focus on Timeslides reflects the fact that in a typical monitoring system, one is not only interested in detecting issues affecting the monitored variables, but also in diagnostics.

3 Multi-Stage Data Flow

In this section, we discuss an example related to monitoring characteristics of tape storage devices, where wrap loss measurements describe properties of the magnetic tape observed in performance tests. Consider the situation where three variables are of importance: v7022, v7023 and v7024 (called "`Wrap_loss_2`", "`Wrap_loss_3`" and "`Wrap_loss_4`", respectively) and these variables are expected to be influenced by operations named Oper8002, Oper9050 and Oper9070. We will typically be able to establish the data structure of type shown in Fig. 1. A



Obs	Lot_Unit	v7022	v7023	v7024	Oper8002	Oper9050	Oper9070
1	Lot01_U1	0.81	-0.85		2016-06-15-12:43	2016-06-17-06:42	2016-06-21-15:13
2	Lot01_U2	0.84	-0.87		2016-06-15-14:22	2016-06-17-04:59	2016-06-21-12:01
3	Lot01_U3	0.78	-0.83		2016-06-15-16:16	2016-06-17-05:29	
4	Lot01_U4	0.76	-0.81	-3.21	2016-06-15-17:43	2016-06-17-04:53	
5	Lot02_U1			-3.23	2016-06-15-14:36		
6	Lot02_U2			-3.11			
7	Lot02_U3			-3.05			
8	Lot02_U4			-3.17	2016-06-15-21:09		
9	Lot03_U1			-3.03	2016-06-16-03:15	2016-06-18-18:26	2016-06-21-16:53
10	Lot03_U2			-3.19	2016-06-16-03:29	2016-06-18-19:38	2016-06-21-17:17
11	Lot03_U3			-3.06	2016-06-16-04:37	2016-06-18-21:13	
12	Lot03_U4				2016-06-16-05:27		
13	Lot04_U1						
14	Lot04_U2	0.74	-0.78		2016-06-17-03:29	2016-06-19-11:26	
15	Lot04_U3	0.75	-0.76		2016-06-17-03:31	2016-06-19-13:11	
16	Lot04_U4	0.73	-0.72	-3.00	2016-06-17-04:40	2016-06-17-08:38	

Fig. 1 Data structure used in multi-stage process monitoring as compiled at the time of data processing, T_p. For the sake of brevity, we are not identifying tools that were used in the operations

row of this data structure represents a unit of a product (for example, a wafer in a semiconductor manufacturing context). The row gives the values of all relevant variables available for the unit at the time of processing. It also gives timestamps that are associated with the relevant operations. In what follows, we refer to the time of data processing as T_p.

For example, at the time of processing T_p, the unit "`Lot01_U1`" accumulated enough data to provide the values 0.81 and -0.85 for the variables v7022 and v7023, but not enough data for v7024. This unit "saw" all three operations, and the timestamp associated with the operation Oper9070 was 2016-06-21-15:13. In contrast, the unit "`Lot02_U2`" (row 6) had enough data to provide the value v7024 $= -3.11$ (but not v7022, v7023)—and this unit did not (maybe, yet) go through any of the three operations of interest. Note that the data structure is generally richer than that shown in Fig. 1: for example, it also identifies the tools that were associated with the three operations. In particular, Oper8002 is typically associated with tools of type "GHBx" (where x is a digit), Oper9050—with tools of type "GHMx" and Oper9070—with tools of type "GHUx".

Now let us prepare, based on the data in Fig. 1, timeslides for the variable v7022 (see Fig. 2). All the timeslides in Fig. 2 are sorted by the timestamp, as required. Note that the top two timeslides (against Oper8002 and Oper9050) have the same

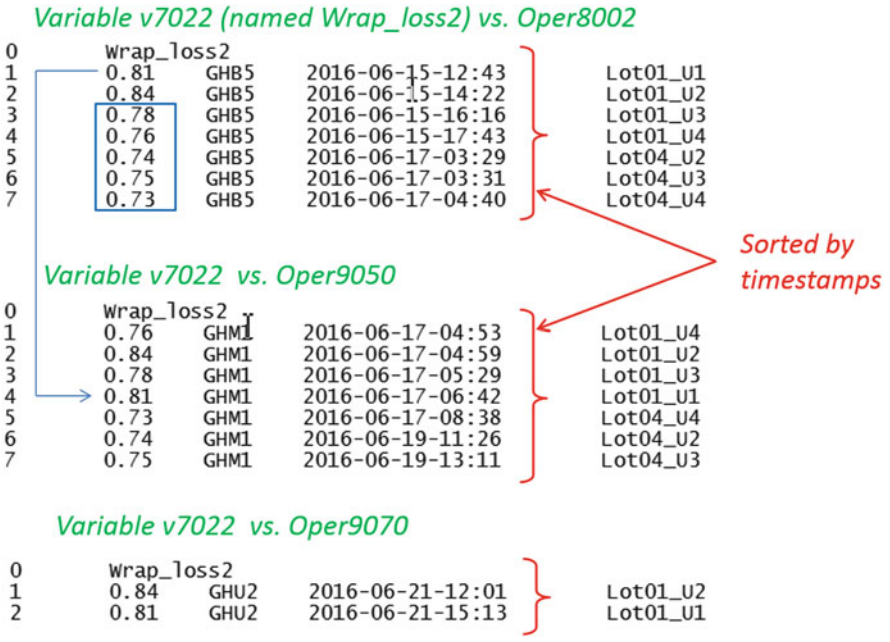


Fig. 2 Timeslides for the variable v7022 with respect to tools of Oper8002, Oper9050 and Oper9070 as compiled at the time of data processing, T_p based on data in Fig. 1. The tools involved in the operations are shown in the timeslides. Note that timeslides are always sorted by timestamp

seven values, but they are sorted in a different order. Two of these values (0.81 and 0.84) are also present in the bottom timeslide of v7022 against Oper9070.

As we will see from Sect. 4, the target mean for v7022 is 0.8 and standard deviation typically observed for this variable is $\sigma = 0.04$. Let us assume that (a) the distribution is Gaussian and (b) values of the mean that differ from the target by s or higher are considered unacceptable. Now let us produce control charts (say, of Cusum type) for all three timeslides. Then the top two timeslides are likely to produce alarms (i.e., get "flagged"), and the bottom timeslide will not show any alarms. Our immediate conclusion will be that there is a problem with the variable v7022—and this problem merits engineering attention. However, we can say more than that, based purely on the timeslide data analysis. Consider the top timeslide: it starts with values (0.81, 0.84) which are in line with the target behavior—but the next five measurements (0.78, 0.76, 0.74, 0.75, 0.73) are clearly more compatible with out-of-control regime. So, this timeslide indicates crisply that the tool GHB5 of Oper8002 might be considered the primary suspect: it looks as if something might have happened to (or around) 2016-06-15-14:22 and 2016-06-15-16:16. In contrast, the signal in the second timeslide is considerably weaker, as high and low values are intermixed in the range of timestamps (2016-06-17-04:53, 2016-06-17-06:42) of the tool GHM1. Finally, the third timeslide does not provide any evidence that the tool GHU2 is related to the detected unfavorable trend. The above argument illustrates usefulness of timeslide analysis for diagnostics. The main role of QEWSV is thus to provide a framework for detection and diagnostics based on analysis of timeslides.

3.1 The Process Inputs

The key part of the analysis is to make sure that data structures of type shown in Fig. 1 are, at least in principle, constructible based on the data in the multi-stage process. This will ensure that any plausible unfavorable trends can be detected by compiling a rich enough collection of data sets of this type and related timeslides. Let us focus on a given segment of data flow (for example, that limited to variables and operations discussed above. We can then specify, for example, that the timeslides "Oper8002 vs v7022", "Oper9050 vs v7022" and "Oper9070 vs v7022" are present in the list of requested timeslides—and this will ensure that the analysis of type shown above will be performed. In principle, one could include all the possible timeslides in the list—however, in practice this list will be limited to combinations that are considered relevant. For example, in the semiconductor manufacturing situation, it would make no sense to include timeslides of clock speed measurements against operations related to dicing of wafers into chips because there is no plausible way in which these operations could be related to slower clock speeds.

Next, one needs to ensure that the actual timeslides are present at the time of analysis (they can be either pre-compiled en masse or produced on demand). Finally, for every segment of the manufacturing process we also need to specify the

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id12345	v7022	v7023	v7024	v7025	v7026	v9570	v9571	v9572	v9574	v4008	v4009	v4015	v4016
2	Meas_Name	Wrap_ loss1	Wrap_ loss2	Wrap_ loss3	Wrap_ loss4	Wrap_ loss5	Wrap_ loss6	Wrap_ loss7	Wrap_ loss8	Wrap_ loss9	Wrap_ loss10	Wrap_ loss11	Wrap_ loss12	Wrap_ loss13
3	Sigma	0.04	0.04	3E-04	0.2	0.005	0.03	0.03	0.1	0.3	2	5	100	50
4	Target	0.8	-0.8	0	1.2	0	0	0	0.8	12.3	2.5	75	700	600
5	Accept_Level	0.8	-0.8	-3E-04	1.21	0	0.05	0.05	0.8	12.3	2.5	72	600	500
6	Unaccept_Level	0.85	-0.75	-6E-04	1.26	0.01	0.1	0.1	1	13	4.5	60	500	400
7	Type_of_Control	2	2	1	2	2	1	2	2	2	2	2	2	2
8	False_Alarm_Rate	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
9	Unaccept_Factor_Sigma	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5

Fig. 3 A typical parameter file used in QEWSV analysis. Column B contains parameters for the variable v7022. We use these parameters to analyze timeslide of the variable v7022 with respect to Operation 8002

Row No.	Name	Type	Description
2	Meas_Name	Char []	Name of the measurement
3	Sigma	Double	Assumed standard deviation of the measurements. This value should be always greater than zero.
4	Target	Double	The most desirable value for the center of the measurement population (typically interpreted as the best level for the *mean of the measurements*.)
5	Accept_ Level	Double	The level of *mean of the measurements* that is still considered *acceptable.* Typically, this level is close to the Target and it reflects the amount of "wiggling room" that is left for the population mean around the target. In many cases involving low process capability this level will coincide with the Target, indicating that there is no wiggling room for the population mean.
6	Unaccept_ Level	Double	The level of *mean of the measurements* that is considered *unacceptable.* This is the level for which we want good detection capability. Generally, *unacceptable* level should be further away from the Target than the *acceptable* level. It is also advisable to keep a certain degree of separation between the acceptable and unacceptable levels (say, no lower than 0.2*Sigma, where possible).
7	Type_of_ Control	Integer	1 means that the control is one-sided (we are only interested in detecting changes up or down). 2 means that the control is two-sided: both types of deviation from the Target are considered unacceptable. If Type_of_Control = 1 and Accept_Level < Unaccept_Level, then of interest are only changes of the process mean *up.* If Type_of_Control = 1 and Accept_Level > Unaccept_Level, then of interest are only changes of the process mean *down.* If Type_of_Control = 2 then we could specify either Accept_Level < Unaccept_Level or Accept_Level > Unaccept_Level, with the understanding that the acceptable and unacceptable levels of the two-sided procedure will be positioned *symmetrically* around the target.
8	False_Alarm_ Rate	Double	Default = 1000. This means that the detection procedure will produce a rate of false alarms of 1 per 1000 points (i.e., values of in timeslide) when the population mean is located at the Accept_Level.
9	Unaccept_ Factor_Sigma	Double	Will not be discussed in this paper.

Fig. 4 Rows of the parameter file

parameters that govern detection and analysis. An example of this file is shown in Fig. 3. The columns of this file are related to monitored variables for which timeslide analysis is required. The upper-left cell of the file contains the ID of the analysis (i.e., ID of the parameter file) and the first row contains the names of the variables. The other entries in the parameter file columns are described in Fig. 4.

Consider, for example, the column G of the parameter file shown in Fig. 3. This column specifies control characteristics for the variable v9570 (named `Wrap_Loss6`). The value of Sigma that the designer wishes to use for this variable

is 0.03 and the process should ideally be centered at zero. However, if the process mean deviates from zero and settles at the *acceptable* level of 0.05—this deviation is not considered a reason for concern. A signal triggered under these conditions would be considered a *false alarm*. If, however, the process level (i.e., population mean) settles as far as 0.1 or beyond—this is considered *unacceptable*, and we expect to detect such a condition as quickly as possible. Between 0.05 and 0.1 lies the "grey zone": alarms will tend to be relatively rare when the mean is near 0.05 and they will become increasingly likely when the mean gets close to 0.1. The detection scheme is one-sided: we are only interested in detection of the process changes *up*. The rate of false alarms of 1000 (default value) indicates that for the process level 0.05 (i.e. the edge of the acceptable zone) the rate of alarms should be 1 in 1000 points.

Now consider the column D (variable v7024 named `Wrap_loss3`). This variable has assigned Sigma of 0.0003 and Target = 0. The process level of -0.0003 is considered acceptable, and the level of -0.0006 is considered unacceptable. Since we require one-sided control for this variable, the configuration of acceptable and unacceptable levels indicates that we would like to detect the process level changes *down*.

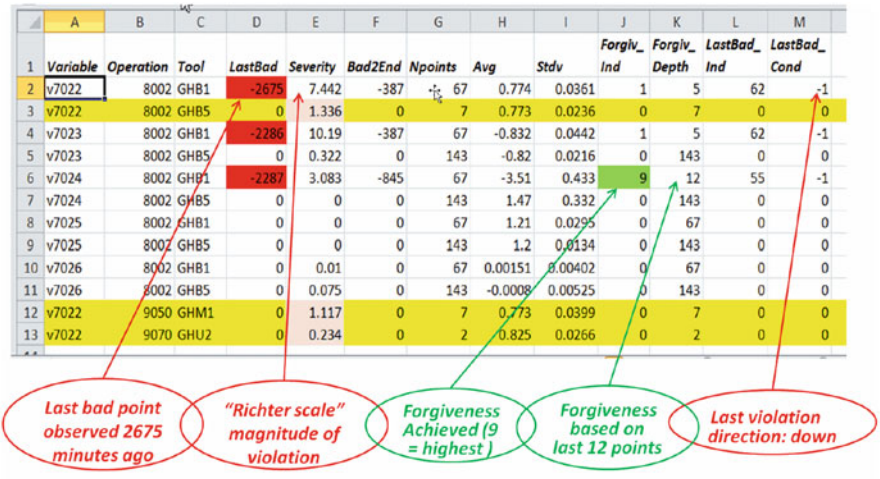
Finally, consider the column E (variable v7025 named `Wrap_loss4`). Its assigned Sigma is 0.2 and Target = 1.2. The process level of 1.21 is considered acceptable (i.e., there is a very small amount of "wiggling room") and the level of 1.26 is considered unacceptable. Since we are requesting two-sided control, the levels of mean between 1.2 and $1.2 - 0.01 = 1.19$ are also considered acceptable, and the levels of mean below $1.2 - 0.06 = 1.14$ are considered unacceptable, by symmetry.

The outlined design only handles the case of symmetric two-sided monitoring; if asymmetric monitoring is required, we will define, in the parameter file, two identically valued variables (with different names) and then apply one-sided upper and lower detection schemes to these variables.

3.2 Outputs

As a result of QEWSV processing, we obtain information on variables (timeslides) that are flagged as well as alarm attributes that are useful in diagnostics and alarm prioritization. In addition, the output contains supporting information, including graphics that help one to further analyze data involved in monitoring or prepare reports.

One of the key output files is the analysis logbook file; it contains lines corresponding to analyses performed, one line per analysis. As noted earlier, QEWSV operates as a *search engine* that sifts through all timeslides specified for a run. The number of timeslides can run into millions, considering the fact that a given variable can be timeslided against hundreds of operations that can potentially (and mostly negatively) affect the distribution of its values. The logbook records



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Variable	Operation	Tool	LastBad	Severity	Bad2End	Npoints	Avg	Stdv	Forgiv_ Ind	Forgiv_ Depth	LastBad_ Ind	LastBad_ Cond
2	v7022	8002	GHB1	-2675	7.442	-387	67	0.774	0.0361	1	5	62	-1
3	v7022	8002	GHB5	0	1.336	0	7	0.773	0.0236	0	7	0	0
4	v7023	8002	GHB1	-2286	10.19	-387	67	-0.832	0.0442	1	5	62	-1
5	v7023	8002	GHB5	0	0.322	0	143	-0.82	0.0216	0	143	0	0
6	v7024	8002	GHB1	-2287	3.083	-845	67	-3.51	0.433	9	12	55	-1
7	v7024	8002	GHB5	0	0	0	143	1.47	0.332	0	143	0	0
8	v7025	8002	GHB1	0	0	0	67	1.21	0.0295	0	67	0	0
9	v7025	8002	GHB5	0	0	0	143	1.2	0.0134	0	143	0	0
10	v7026	8002	GHB1	0	0.01	0	67	0.00151	0.00402	0	67	0	0
11	v7026	8002	GHB5	0	0.075	0	143	-0.0008	0.00525	0	143	0	0
12	v7022	9050	GHM1	0	1.117	0	7	0.773	0.0399	0	7	0	0
13	v7022	9070	GHU2	0	0.234	0	2	0.825	0.0266	0	2	0	0

Fig. 5 The analysis logbook. A total of 12 analyses were performed. Three timeslides, corresponding to red boxes in column D, are flagged (Color figure online)

give summaries of timeslide analysis (one row per timeslide) and the logbook is the primary driver of the user interface, such as dashboard or list of timeslides flagged during the course of the search engine run. The logbook corresponding to example discussed earlier is shown in Fig. 5. In this case, QEWSV performed 12 analyses. Consider the first analysis, based on timesliding the variable v7022 with respect to the operation 8002. One can see that there are two tools involved in the operation 8002: GHB1 and GHB5. The first line of the logbook file is related to analysis of GHB1. The column D contains the return code of the analysis: a negative return code indicates that the analysis was flagged. The value −2675 of the return code, also marked as *LastBad*, gives the number of minutes that elapsed since the last data point that can be attributed to unacceptable process level; this amounts to 1.86 days. The value 7.442 in column E is the severity of detected deviation from the acceptable process behavior. This is a logarithmic measure, and one can think of it as a kind of a "Richter scale". The value −387 (2nd return code, marked as *Bad2End*) gives the number of minutes elapsed between the timestamp corresponding to the last bad point and the timestamp of the last data point. The value −1 would have indicated that the detected bad condition persisted till the end of the data in the timeslide—however, in our case it appears that the data points observed within the last 387 min of the data showed behavior consistent with the acceptable level. The columns G-I tell us that the data consisted of 67 points, and that the mean and standard deviation of these points were 0.774 and 0.0361, respectively. The columns J and K are related to forgiveness conditions (see Sect. 5.3). The column K indicates that the last five data points supported the evidence that the process might have returned to acceptable behavior. However, 1 in column J tells us that the degree of forgiveness (0–9) is very low. The value 62 in column L gives us the index of the

Col No.	Name	Type	Description
1	Variable	Char []	Variable ID (e.g., "v7022")
2	Operation	Char []	Operation ID
3	Tool	Char []	Tool ID
4	Return Code (RC)	Double	Zero for a normal run. Positive values correspond to error conditions. Negative values give the time elapsed since the last point consistent with the unacceptable process level. A negative return code means that the analysis was "flagged".
5	Severity	Double	Degree of violation of acceptable conditions detected during the analysis. This is a base-10 logarithm of the "conformance to acceptable process level" test p-value, see Sec. 5.1.
6	Second Return Code (2nd RC)	Double	Zero for a normal run. If the analysis was flagged then the 2nd RC is the time elapsed between the timestamp corresponding to the last bad point and the timestamp of the last data point, taken with a minus sign. The value -1 is a maximum that can be reported under the flagged conditions – and this value indicates that the detected bad condition persisted till the end of the data.
7	Number of Points	Integer	Total number of points in the analysis (that covers the given Variable-Operation-Tool combination)
8	Average of Points	Double	Average of all points in the analysis
9	Standard Dev. of Points	Double	Sample standard deviation of all points in the analysis
10	Forgiveness Index	Double	Degree of forgiveness observed within the last acceptable period of the data (as measured by Forgiveness Depth (next field). The value 0 for a flagged analysis indicates that there was no information suggesting that the process might be returning to acceptable behavior. The highest level of 9 is achieved when the forgiveness criteria (as governed by the built-in option *forgiveness*) are satisfied, see Sec. 5.3.
11	Forgiveness Depth	Integer	The number of last data points that are consistent with the acceptable process level. This is the number of points since the last bad point that plays the role in establishing the values of RC and 2nd RC. This is also the length of the Last Good Period (LGP), see Sec. 5.2.
12	Index of Last Bad Point	Integer	The sequential index of the point that concludes that last regime in the data that is consistent with an unacceptable process behavior. Points observed since the last bad point are consistent with acceptable process levels.
13	Last Bad Condition	Double	0 if the analysis was not flagged, 1 if the last bad point was related to the process change upwards, -1 if the last bad point was related to the process change downwards.

Fig. 6 Columns of the logbook file shown in Fig. 5

last data point consistent with unacceptable behavior. The value −1 in column M tells us that the last bad condition was related to the process mean being too low. Note that the analysis could detect several bad conditions present in the data, and the severity index of the analysis would reflect that. However, the focus in columns J-M is on the nature of the last bad condition, as it is likely to be most relevant as far as diagnostics, prioritization or corrective actions are concerned.

In the next section, we will show supplemental information (chart, table) pertaining to this analysis. The summary of fields of the logbook file is given in Fig. 6.

Information contained in the logbook file is sufficient for many objectives of monitoring: for example, it can be used as a basis for multi-layer dashboards or displays with sortable columns. In essence, the subset of *flagged rows* of this file can be viewed as an output of a search engine that can be further manipulated in order to establish priorities. These priorities will depend on the role of the analyst. For example, people responsible for assessing a financial impact of flagged conditions might be more interested in cases where most severe violations of acceptable

behavior have been detected. On the other hand, people responsible for process control might be most interested in conditions where the severity is low but the RC is close to −1, reflecting "freshness" of the detected bad condition. In the initial phases of the developing problem, we are not likely to see a very severe violation of the acceptable behavior. The severity index of such condition is likely to be low—yet sufficient for flagging the analysis, placing the detected weak signal on the "radar screen". Focusing on "fresh" signals would enable one to spot such brewing bad conditions early. A flagged analysis with high severity (even with RC close to −1, indicating ongoing bad behavior) is likely not to be new—such results could persist for a while until the effects of corrective actions become sufficiently transparent.

It is also important to keep in mind the timesliding dimension of the analysis. As noted earlier, a given variable can participate in a number of analyses (timeslides) corresponding to different operations of potential impact. If the data for this variable contains a sufficiently high number of points consistent with the unacceptable process level, then this variable will end up flagged for several (and sometimes all) timeslides. Under such conditions, we might need to do some detective work to establish the relative degree of importance of the flagged timeslides for this variable, and this is where other alarm attributes can be handy. Of special importance to the problem of timeslide comparison are also visualization techniques—and this is why supplemental information is provided for every flagged condition. This information consists of chart/table pairs, and we describe it next. Let us explore a particular chart/table pair corresponding to the variable 7022, see Figs. 7 and 12. The alarm attributes corresponding to this flagged analysis are given in line 1 of the logbook file and discussed above. The parameters used in the analysis are given in the Column B of the parameter file, see Fig. 3.

The plot contains of a pair of horizontal strips: the upper strip gives the data plot and the bottom strip gives the corresponding evidence chart. On the top plot, the horizontal lines are shown for the Target (black) and for the unacceptable levels (dashed blue). The "Rng" in the plot header gives the range of timestamps for the 67 points included in the analysis. In the upper-right corner is the date of the analysis. The x-axis gives indices of points, and further information about the points is available in the accompanying table, see Fig. 12.

The bottom plot gives the two-sided Page's (Cusum-Shewhart) chart. We call it the Evidence Chart. The term "Evidence" helps users who are not familiar with the Cusum Technique to understand the meaning of the chart: when the data is explained better by the acceptable process level than by the unacceptable level, the Evidence trajectory will take on values near zero. When the data is better explained by the unacceptable level, the evidence trajectory will take off towards the threshold (signal level). A horizontal slope of the trajectory indicates that the data is equally consistent with both acceptable and unacceptable levels. The upper part of the chart is responsible for detecting changes of the process level up, and the lower part is responsible for detecting changes down. We found that users who absorb just this basic information are able, after some experience, to interpret the charts correctly.

We can see that an alarm was produced by the lower evidence chart. The severity of the detected condition (7.442) is shown in the footer line of the plot and the last

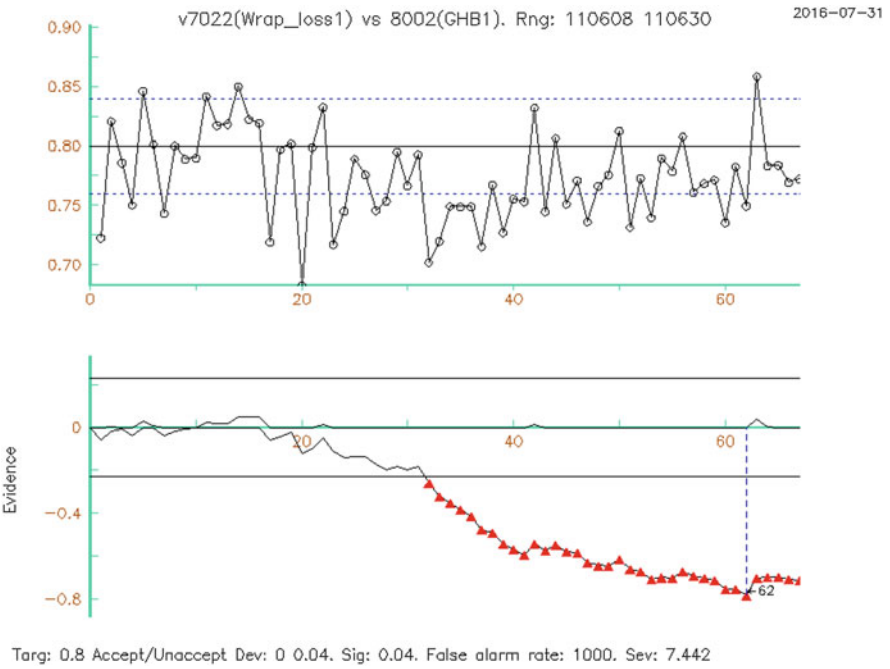


Fig. 7 Timeslide of v7022 with respect to the Operation 8002. Only the subset of data corresponding to the tool GHB1 is used in the analysis. The data set contains 67 points. The last bad point (no. 62) is marked by the dashed vertical line. The horizontal lines on the top plot correspond to the target and unacceptable levels

bad point (No. 62) is marked. Therefore, only $67 - 62 = 5$ last points provide information on the degree of forgiveness, as reported in the first line of the logbook file, see Fig. 5. The timestamps of points 62 and 67 are 2011-06-30-14:11:30 and 2011-06-30-20:38:28; they are separated by 387 min—this is also reported in the logbook file (Col. F). Note that the acceptable/unacceptable levels (see bottom line of the plot) are represented in terms of deviation from the Target: according to the v7022 column of the parameter file, the acceptable/unacceptable levels of the process mean are 0.8 and 0.85, respectively—so the acceptable/unacceptable deviations from the Target are 0 and 0.05.

4 The Detection Algorithms

Consider a data set $X_1, X_2, \ldots, X_N$ consisting of N points and the problem of detecting unacceptable deviations in the underlying process mean. Based on the "`Type_of_Control`" setting in the parameter file (Fig. 3), we apply either a one-sided or two-sided detection algorithm.

4.1 One-Sided Detection Schemes

Denote the acceptable and unacceptable levels of the underlying process mean by μ_0 and μ_1, respectively. Let us assume that our task is to detect process changes up (i.e., $\mu_0 < \mu_1$). In such case, we should specify Target < μ_0, i.e., the acceptable level must be closer to the Target than the unacceptable level. Let us transform the data process to the process of control scheme values $S_0, S_1, \ldots, S_N$ in the following way:

$$S_0 = 0,\ S_i = max\,[0, S_{i-1} + (X_i - k)], \qquad i = 1, 2, \ldots, N \tag{1}$$

where

$$k = (\mu_0 + \mu_1)/2. \tag{2}$$

As mentioned in Sect. 3.2, in QEWS modules, we refer to the set $S_i, i = 0, 1, \ldots, N$ as the *Evidence Chart*. In the literature, the process S_i is known as the one-sided (upper) Page's scheme, see Hawkins and Olwell (1998). Since it can be viewed as representation of the degree of evidence that the process is better explained by the unacceptable level than by acceptable level, here and in what follows we will refer to S_i as "Evidence", e.g. see the y-axis label "Evidence" in Fig. 7.

A timeslide represented by a data set X_i, $i = 1, \ldots, N$ is flagged if for some i the value of the scheme (1) exceeds a threshold h. The value of h is selected so as to achieve a pre-specified rate of false alarms defined as "`False_Alarms_Rate`" in the parameter file, see Fig. 3. This value is the on-target Average Run Length (ARL0). It is computed under the assumption that the process mean is μ_0 and that the process standard deviation is equal to Sigma specified in the parameter file.

Note that as the length N of the data series increases, the probability that the corresponding scheme flags this series increases as well. For example, if ARL0 = 1000 and our data series (timeslide) is of length $N = 2000$, then the probability of the analysis getting flagged is approximately $1 - \exp\,[-(1/1000)*2000] = 1 - \exp[-2] = 0.86$, i.e., quite high. Therefore, if it is known a priori that length of timeslides typically encountered in the analysis is around 2000, then the choice of ARL0 = 1000 is unsuitable. Generally, it is a good policy to select ARL0 at least 50× the expected window size N; in this case, the false flagging probability per analysis is about $1 - \exp[-1/50] \approx 1/50 = 0.02$.

Once the value ARL0 has been established, the corresponding value of the threshold h can be obtained by solving the equation

$$ARL(h|k, \mu_0, s) = ARL0, \tag{3}$$

where $ARL(h|k, \mu_0, s)$ is the theoretical value of the Average Run Length as a function of h, computed under the assumption that the scheme (1) is applied to

the data, the mean of the data is at the edge of the acceptable region (i.e., μ_0) and its standard deviation is s (i.e., the nominal Sigma, see Fig. 3). Equation (3) can be solved using the Markov Chain modeling of Cusum-Shewhart schemes, e.g., see Yashchin (1985). However, in many cases one can make additional assumptions that (a) the data sequence X_i is Gaussian and (b) its terms are independent and identically distributed. In such cases, we can also use the approximation

$$ARL(h|k, \mu_0, s) \approx 2\tilde{h}^2 * \frac{1}{(2a)^2}[exp(-2a) - 2a + 1] \tag{4}$$

where

$$\tilde{h} = h/\sigma + 1.16, \quad a = -\tilde{h} * (k - \mu_0)/\sigma, \tag{5}$$

motivated by the Brownian Motion approximation for the Cusum process, e.g., see Bagshaw and Johnson (1975) and Siegmund (1985).

Finding h enables one to establish the primary detection rule. However, in practice we need supplemental rules that enhance our ability to detect very large changes quickly. As noted in Hawkins and Olwell (1998), this can be achieved by introducing a supplemental Shewhart's limit c and flagging the analysis if $X_i > c$ for some i . The value c used in QEWSV analysis is chosen based on the equation

$$c = k + c_a \tag{6}$$

where the value c_a is chosen so as to have a minimal impact on the nominal false alarm rate. In particular, the decrease in ARL0 caused by the additional signal criterion is about 5%.

One can see that use of Cusum-Shewhart methodology here is specific to the problem of timeslide analysis. In many applications of the Cusum technique, one can preserve the state of analysis in terms of the scheme values: at the next time point of analysis, it is only necessary to update these values using the new data. In contrast, new data in timeslides can appear anywhere in the time series, as sorting can re-shuffle the data points in accordance with the new timestamp information. Therefore, the whole data set typically needs to be re-analyzed from scratch at every time point of analysis.

An example of the upper scheme for the variable v9570 (timeslided against the operation 9050, tool GHM3) is shown in Fig. 8; the parameters of the monitoring process are shown in Fig. 3, col. G.

4.2 Lower and Two-Sided Detection Schemes

In case of one-sided control with $\mu_0 > \mu_1$, we are interested in detecting changes *down*. Since the problem of detection of change in the sequence $\{X_i\}$ *down* is

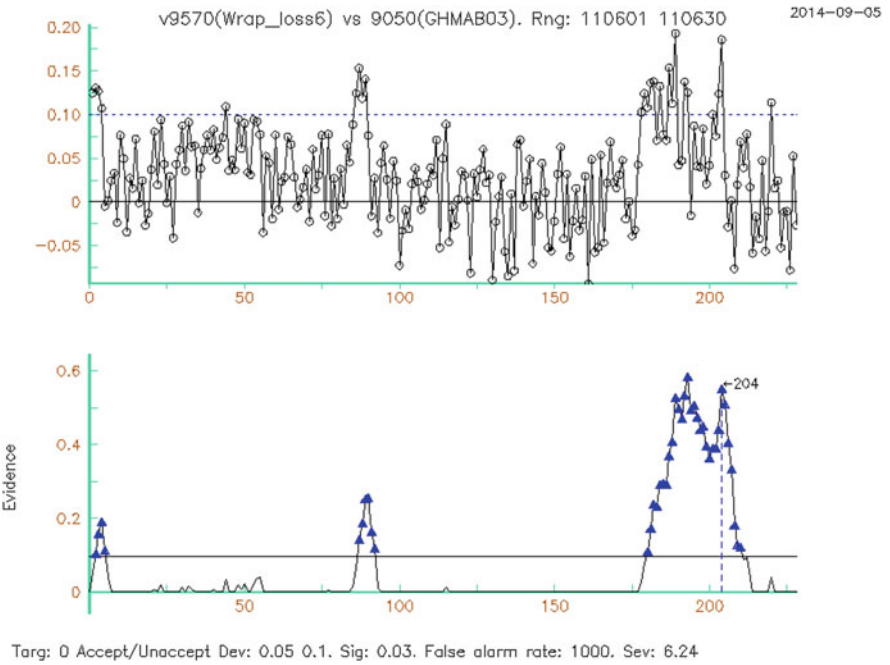


Fig. 8 Timeslide of the variable v9570 with respect to the Operation 9050 (upper scheme). Only the subset of data corresponding to the tool GHM3 is used in the analysis. Parameters governing the analysis are shown in Fig. 3, col. G

equivalent to the problem of detection of changes in the reflected sequence $\{-X_i\}$ *up*, the corresponding one-sided lower detection scheme can be defined as follows:

$$S_0^- = 0,\ S_i^- = max\,[0, S_{i-1}^- + (-X_i - k^-)], \qquad i = 1, 2, \ldots, N \tag{7}$$

where

$$k^- = -(\mu_0 + \mu_1)/2, \tag{8}$$

i.e., the reference value k^- is also defined in terms of the reflected sequence. One can see that the lower scheme is also non-negative, so that the threshold $h^- \geq 0$ is applied in order to decide whether the analysis is to be flagged. The supplemental Shewhart's limit c^- is defined automatically, in a way similar to (6). In practice, however, it is often convenient to represent the lower scheme with a negative y-axis, effectively plotting the reflected values of the scheme, see Fig. 9. This figure shows the timeslide for the variable v7024 with respect to the operation 8002, tool GHB1; the corresponding parameters are shown in Fig. 3, col. D.

Two-sided schemes. Two-sided monitoring is implemented as a combination of upper and lower Page's schemes.

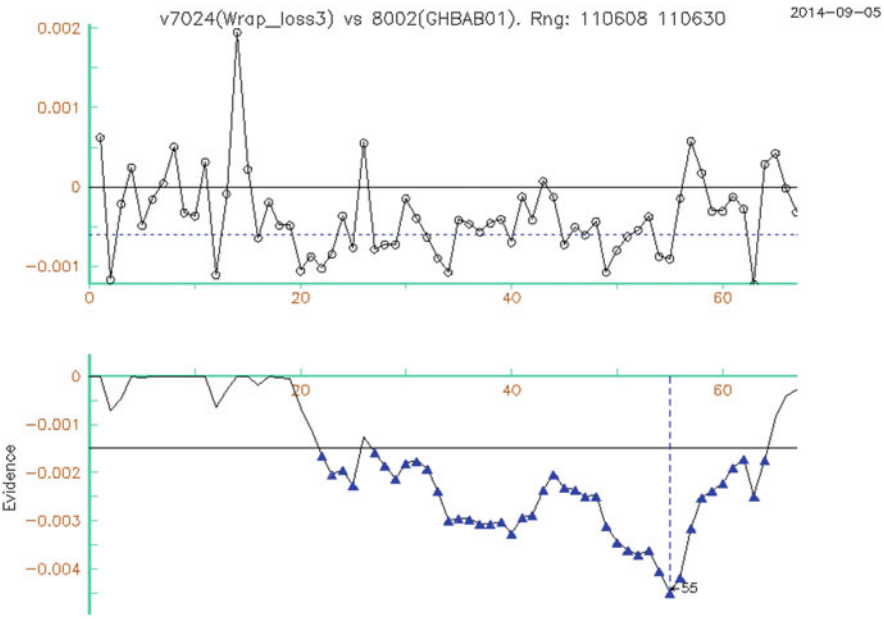


Fig. 9 Data and lower Page's scheme plots corresponding to variable v7024 with respect to the Operation 8002, tool GHB1

5 Alarm Attributes

In the wake of an alarm (flagging) event, it is important to provide alarm attributes to facilitate diagnostics and alarm prioritization. Accordingly, the logbook file (Fig. 5) gives a number of fields that serve this purpose. In this section, we describe algorithms related to the three main attributes: *severity, last good period* and *forgiveness*.

5.1 Severity

Severity reflects the degree of deviation from acceptable process conditions detected over the course of analysis. This attribute consists of two components: (a) severity associated with the maximal value of the Page's scheme (i.e., Evidence) achieved in the run and (b) severity based on the end value of the Evidence trajectory. First, consider a one-sided scheme S_i computed using (1) and values $X_1, X_2, \ldots, X_N$ of the monitored variable. Denote $S = max(S_1, S_2, \ldots, S_N)$.

The statistic S can be used as a basis of a test that the process mean remained in the acceptable region throughout the process. Let s be the value of S observed in the run. Then Sev_1, the first component of severity, is defined as a base-10 logarithm of

the test p-value:

$$Sev_1 = -ln_{10}[Prob\{S > s|N, s, \mu = \mu_0\}] \tag{9}$$

Similarly, let s_N be the end value of the evidence trajectory observed for the data set. Then Sev_2, the second component of severity is related to the p-value of the test based on S_N,

$$Sev_2 = -ln_{10}[Prob\{S_N > s_N|N, s, \mu = \mu_0\}]. \tag{10}$$

This test puts emphasis on the last portion of the data set. The combined severity Sev is a function of the individual severities; for example, one can use the average

$$Sev = (Sev_1 + Sev_2)/2. \tag{11}$$

It is this measure of severity that is shown in tables and plots of this chapter.

The exact computations of severity components are quite complex; for the first component, we can use the approximation

$$Prob\{S > s|N, \sigma, \mu = \mu_0\} \approx 1 - exp\left[-\frac{tN}{ARL(s|k, \mu_0, \sigma)}\right], \tag{12}$$

computed based on formulas (4)–(5) with the coefficient $t = 1$ in the above formula. For the second component, we can use the approximation

$$\begin{aligned} &Prob\{S_N > s_N|N, \sigma, \mu = \mu_0\} \approx \\ &\quad 1 - max\left\{0, 1 - t * exp\left[-\left(\tfrac{2(k-\mu_0)}{\sigma}\right)\left(\tfrac{s_N}{\sigma} + 0.65\right)\right]\right\} \end{aligned} \tag{13}$$

computed based on Brownian Motion approximation to the distribution of the endpoint of Evidence trajectory (e.g., see Cox and Miller (1977)) with the continuity correction. The value $t = 1$ is used in (13), as in the earlier formula.

For lower schemes one can use the same formulas: as noted earlier, these detection schemes can be treated as instances of upper control schemes applied to reflected values of the variable. For two-sided schemes, we define the values S and S_N as maxima of the respective values for one-sided schemes, e.g., $S = max\{S^{(upper)}, S^{(lower)}\}$. The severity for two-sided schemes is computed based on (9)–(11), with the value $t = 2$ used in formulas (12)–(13).

5.2 Last Good Period

In general, severity by itself is not sufficient to decide on how to prioritize a flagged analysis. The fact that the deviation from the acceptable process level is severe does

not tell us, for example, how long ago were bad trends last seen in the data, i.e., how recent was the last detected alarming condition. In order to find the "last bad point" corresponding to the last unfavorable data regime, we use the procedure that inspects segments of data starting from the last point and going progressively deeper into history. For example, just an examination of the last point could lead one to a conclusion that the process level is unacceptable at this point, effectively terminating the search. On termination, the search will yield the "Index of the last bad point" (see Fig. 5, Col. E and Fig. 6).

For a one-sided (upper) scheme, the search returns a window of depth M corresponding to indices ranging from $i = N - M + 1$ to N if a window of depth $M_0 > M$ can be identified for which each of the following four conditions is met:

1. Starting from zero at time $i_0 = N - M_0$, the scheme of type (1) does not exceed the threshold h that is used for scheme flagging;
2. Starting from zero at time $i_0 = N - (M_0 + 1)$, however, the scheme (1) does exceed the threshold h.
3. The maximum value of the scheme (1), when started from zero at the index $i_0 = N - (M_0 + 1)$ is achieved at time $i_{max} = N - M$.
4. For none of the last M points was the Shewhart's supplemental criterion triggered, i.e., $X_i < c$ for every $i = N, N - 1, \ldots, N - M + 1$.

Computations for a one-sided lower case are analogous. For a two-sided case, we first establish the type of condition associated with the last bad regime. This is achieved by examining the values of the scheme and information about the indices for which either primary or supplemental criteria exceeded respective thresholds. Once we know, for example, that the last bad condition was associated with abnormally high process levels, the last bad point can be established via analysis of the one-sided (upper) scheme, as described above.

The window M obtained via the above algorithm is referred to as the length of the *last good period* (LGP). This value can be inferred from the column in the logbook file that gives the index of the last bad point i_{max}. and the column "Npoints" that gives the number of points in the series. For example, for the first row of the logbook file (v7022 vs Oper8002, Fig. 5) the number of points is 67 and the index of the last bad point is 62; therefore, the last good period consists of $M = 67 - 62 = 5$ points.

The LGP plays a critical role in alarm prioritization. For example, in Fig. 10 we show a display of flagged timeslides for the segment of a semiconductor manufacturing line. The cases (rows) of the table can be rank-ordered by alarm attributes, and this table is ordered by severity. Such an ordering might be of importance to people dealing with financial impact of out-of-control conditions: indeed, high severity signals are usually associated with more severe financial consequences. Moreover, severity is one of the factors helping one to identify the operations/tools that are more likely to be the culprits or are, in some sense, "close" to the likely culprits. On the other hand, people dealing with early warning are more likely to rank order the flagged timeslides by the return code (named "days ago" in Fig. 10) or by the LGP. Indeed, a freshly emerging unfavorable condition is not

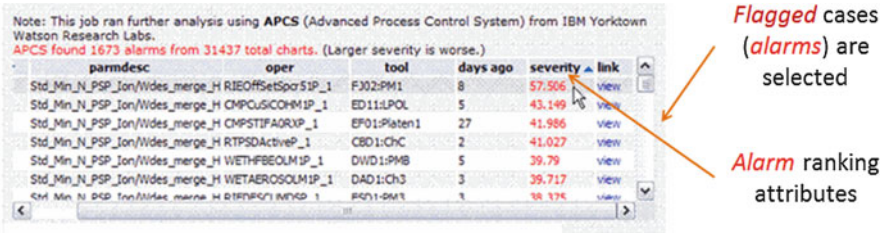


Fig. 10 Display of selected timeslides (i.e., those that got flagged) in the segment of a semiconductor manufacturing line

likely to be of high severity—however, alarm attributes like return code or LGP (possibly, jointly with forgiveness factors defined in the next section) will typically identify such a condition as "fresh" and ongoing.

5.3 Forgiveness Criteria

Even in cases where there is some history suggesting that the process level might be returning to the acceptable region, there is still a question on the statistical strength of the evidence. The forgiveness index enables one to judge the degree of statistical significance in the observed "return to normal" process. Note that dashboard-level decisions on turning off alarm lights will typically take the forgiveness index into account—however, information that is external to QEWSV might also play an important role.

Forgiveness attribute of a flagged analysis is governed by two parameters described below. In general, degree of forgiveness depends on the amount of evidence that the process mean has returned to a level that is considered satisfactory. This level is governed by $\delta \in [0, 1]$. The default level is $\delta = 0.5$. The forgiveness level $\mu_{0\delta}$ for a one-sided (upper) control scheme is defined as follows:

$$\mu_{0\delta} = \mu_0 + \delta(k - \mu_0), \tag{14}$$

where k is the reference value (2). Selection of $\delta = 0$ would require the data following the last bad point (as defined in Sect. 5.2) to support strongly the hypothesis that the process mean μ has settled at μ_0 or lower. Selection of $\delta = 1$ leads to a weaker requirement: the data needs to support the hypothesis $\mu \leq k$ at a high level of statistical significance.

The strength of evidence that $\mu \leq \mu_{0\delta}$ is measured in terms of the "return" level of confidence $(1 - \epsilon_r)$, which is the second parameter governing forgiveness. Its default value is 0.95, i.e., $\epsilon_r = 0.05$.

Based on the LGP value M obtained in Sect. 5.2, we compute the scores

$$Z_m = \frac{\bar{X}_{[m]} - \mu_{0\delta}}{\sigma/\sqrt{m}}, \quad m = 1, 2, \ldots, M \tag{15}$$

where

$$\bar{X}_{[m]} = \frac{X_N + X_{N-1} + \ldots + X_{N-m+1}}{m}, \quad m = 1, 2, \ldots, M \tag{16}$$

Let us denote the corresponding p-values $p[m] = \Phi(Z_m), m = 1, 2, \ldots, M$ (Φ is the standard normal cdf) and by m^* the value of m that maximizes $p[m]$ (i.e., the window of the worst observed significance). The value 9 of the forgiveness index is returned if both $p[M] < \epsilon_r$ and $p[m^*] < 0.5$, i.e., not only is the estimate of the process mean based on the LGP value M in the acceptable domain $\mu \leq \mu_{0\delta}$ with high degree of confidence, but also every sub-segment of m last points supports the hypothesis that the process level has "returned to normal", at least to some degree. The value 8 is returned if $p[M] < \epsilon_r$, but $p[m^*] \geq 0.5$, i.e., the overall evidence for the "return to normal" hypothesis is strong—however, there exist sub-segments which do not support it strongly. The summary of the forgiveness indices is given in Fig. 11.

Forgiveness indices for the one-sided lower case are computed in a similar way. In the case of two-sided detection scheme, one needs first to establish the type of condition associated with the last bad regime, see the end of Sect. 5.2. Once we know the nature of the last bad regime (i.e., whether it is associated with unacceptable changes in the process level up or down), we can compute the forgiveness index based on the corresponding one-sided detection scheme.

Fig. 11 Values of the forgiveness index returned by QEWSV analysis. Note that the value 0 is also returned in cases when the corresponding analysis was not flagged (i.e., forgiveness is not needed)

Forgiveness index	Condition
9	$p[M] < \varepsilon_r$ and $p[m^*] < 0.5$
8	$p[M] < \varepsilon_r$ and $p[m^*] \geq 0.5$
7	$p[M] < 2\varepsilon_r$ and $p[m^*] < 0.5$
6	$p[M] < 2\varepsilon_r$ and $p[m^*] \geq 0.5$
5	$p[M] < 3\varepsilon_r$ and $p[m^*] < 0.5$
4	$p[M] < 3\varepsilon_r$ and $p[m^*] \geq 0.5$
3	$p[M] < 4\ \varepsilon_r$ and $p[m^*] < 0.5$
2	$p[M] < 4\ \varepsilon_r$ and $p[m^*] \geq 0.5$
1	*All other cases where* $M \geq 1$
0	*M=0 (i.e., no sign of improvement)*

v7022(Wrap_loss1) vs 8002(GHBAB01). Rng: 110608 110630

1	515AF1203	110608	05:49:11	0.723
2	505RA0455	110620	16:40:46	0.8211
3	515AF2549	110624	04:47:27	0.7861
4	515AF2612	110624	04:56:55	0.7505
5	515AF2604	110624	05:05:58	0.8465
6	515AF2605	110624	05:25:34	0.8017
7	515AF2755	110629	10:10:08	0.7437
8	515AF2731	110629	11:24:17	0.8002
9	515AF2733	110629	11:36:39	0.7891
10	515AF2732	110629	11:48:00	0.7902
11	515AF2740	110629	12:14:54	0.8419
12	515AF2746	110629	12:27:06	0.8177
13	515AF2741	110629	12:43:01	0.8187
14	515AF2712	110629	13:06:33	0.8503
15	515AF2721	110629	14:28:41	0.8227
16	515AF2647	110629	16:38:50	0.8195
17	515AF2643	110629	16:49:42	0.7195
18	515AF2266	110629	16:58:29	0.7973
19	515AF2645	110629	17:11:38	0.8021
20	515AF2339	110629	17:21:15	0.6828
21	515AF2736	110629	17:31:55	0.7989
22	515AF2343	110629	17:42:34	0.8329
23	515AF2646	110629	17:52:51	0.7174
24	515AF2727	110629	18:03:18	0.7454
25	515AF2644	110629	18:13:06	0.7894
26	515AF2344	110629	18:34:32	0.7762
27	515AF2725	110629	18:53:41	0.7461
28	515AF2307	110629	19:08:46	0.7541
29	515AF2641	110629	19:17:29	0.7952
30	515AF2735	110629	19:29:11	0.7668
31	515AF2334	110629	19:38:57	0.7931
32	515AF2393	110629	19:50:51	0.7024
33	515AF2378	110629	20:01:20	0.72
34	505RA1046	110629	20:11:39	0.7495
35	505RA1069	110629	20:21:23	0.7493
36	505RA1073	110629	20:32:29	0.7492
37	505RA1059	110629	20:43:00	0.7156
38	505RA1071	110629	20:52:19	0.7674
39	505RA1068	110629	21:12:16	0.7273
40	515AF2276	110629	21:22:07	0.7557
41	505RA1067	110629	21:40:31	0.7536
42	505RA1054	110629	21:54:15	0.8324
43	515AF2792	110630	03:37:13	0.7451
44	515AF2796	110630	03:49:07	0.8069
45	515AF2774	110630	04:01:24	0.7514
46	515AF2771	110630	04:14:29	0.7711
47	515AF2776	110630	04:27:15	0.7365
48	515AF2778	110630	04:45:32	0.7665
49	515AF2775	110630	05:04:32	0.7759
50	515AF2787	110630	05:16:18	0.8129
51	515AF2807	110630	05:28:28	0.7318
52	515AF2765	110630	05:44:43	0.773
53	515AF2806	110630	05:59:24	0.7401
54	515AF2800	110630	06:13:11	0.7899
55	515AF2785	110630	06:33:41	0.7789
56	515AF2891	110630	12:53:28	0.8081
57	515AF2867	110630	13:09:06	0.7614
58	515AF2673	110630	13:24:03	0.7686
59	515AF2892	110630	13:35:24	0.7719
60	515AF2902	110630	13:49:08	0.7358
61	515AF2876	110630	14:01:14	0.7828
62	515AF2880	110630	14:11:30	0.7499
63	515AF3002	110630	14:45:30	0.859
64	505RA1075	110630	15:58:01	0.7838
65	505RA1076	110630	19:58:13	0.7842
66	505RA1083	110630	20:08:21	0.7697
67	515AF2977	110630	20:38:28	0.7725

Fig. 12 Output table accompanying the plot in Fig. 7

6 Discussion

Change detection techniques based on Likelihood Ratios are most promising in the modern high-volume and high-intensity data environments. Of special importance is the Cusum-Shewhart methodology, which provides high statistical power while maintaining good designability and interpretability. In our experience, these properties are highly valued by the users. Over the years, several large-scale systems based on use of this methodology were deployed in IBM, the earliest one described in Yashchin (1985). These systems were used not only for manufacturing operations (semiconductors, storage, personal systems, servers) but also for business processes such as finance, pension fund management and investment portfolio monitoring, e.g., see Philips et al. (2003).

In this chapter, our primary focus is on statistical methodology used in conjunction with multi-stage process data. In practice, management of EWSs is much more complex; it requires substantial analytics and data-handling capabilities that we could not describe in detail. We refer the readers to Baseman et al. (2010), Civil et al. (2013) and Negandhi et al. (2015) for information related to implementation and system properties.

In many implementations, the issue of *outlier management* is of critical importance. We found it useful to address the outlier issue prior to submitting the data to the search engine. The *robustness* issue is also very important: one needs to be well prepared to handle situations when the assumptions outlined in Sect. 2 cannot be justified. Some forms of assumption violations can be accommodated through manipulation of process parameters. For example, when a moderate (and known) amount of serial correlation is present, one could get away with simple adjustment of the acceptable false alarm rate in the parameter file (see Fig. 3). Effects of other types of violations, e.g., some skewness or presence of discretized data instead of the assumed continuous data, can be ameliorated by manipulating Sigma in the parameter file. In general, however, one needs to be ready for applying transformations and running timeslide analysis in the transformed space.

Composition of the QEWSV engine output files, including logbook, is another important implementation issue. For example, in some applications one could prefer reporting individual measures of severity (9)–(10) instead of the combined measure (11). The engine must provide enough attributes in the logbook to enable operation of the user interface, such as a *dashboard*. In general, dashboards deploy their own logic, and statistical information provided by the engine needs to be supplemented by operational or business information to decide on prioritizing alarms, disregarding them, or taking other actions. Indeed, from the perspective of the dashboard administrator, achievement of the highest forgiveness index may not be sufficient for declaring that the process has returned to acceptable conditions, and de-prioritizing the alarm. If the forgiveness computation is based, say, on only 2 days of fresh data, the dashboard logic could demand at least three additional days of conforming data before changing the alarm status.

Finally, management of computing resources related to engine deployment needs to be carefully planned. If the engine is activated at pre-specified time points, one needs to make sure that the mode of deployment can support the pace. For example, if the data intensity is high and the engine is activated every second, one could choose the deployment mode that limits production of graphics, or eliminates it altogether. The use of multi-processor hardware typically enables one to enhance efficiency substantially due to parallel nature of timeslide processing.

Acknowledgements I would like to thank Aaron Civil, Reynaldo Corral, Jeff Komatsu, Tony Spielberg, John Wargo and Paul Zulpa from the IBM Supply Chain organization for their kind help, feedback and effort in developing a solution based on this methodology. I am also thankful to David L. Jensen and Brian F. White (IBM Research) for help with software design and development and to Robert J. Baseman for his valuable advice and feedback. My deepest appreciation goes to Steven Ruegsegger and William K. Hoffman from the IBM Microelectronics organization for help in developing software and for integrating the methodology into the ecosystem of tools. I am most thankful to Ishan Sehgal and Jayashree Ravichandran (IBM Internet of Things) for their help and support in productizing this methodology. I am also very indebted to the Editor and the Referee for their insightful comments and suggestions.

References

Bagshaw, M., & Johnson, R. A. (1975). The effect of serial correlation on the performance of CUSUM tests II. *Technometrics, 17*, 73–80.

Baseman, R. J., Hoffman, W. K., Ruegsegger, S., & Yashchin, E. (2010). System for monitoring multi-orderable measurement data. US Patent Publication 20100017009A1.

Capizzi, G. (2015). Recent advances in process monitoring: Nonparametric and variable-selection methods for phase I and phase II (with discussion). *Quality Engineering, 27*, 44–80.

Civil, A. D., Komatsu, J. G., Ng, A. S., Liang, Y., Wargo, J., Yashchin, E., et al. (2013). Hybrid analysis of emerging trends for process control. US Patent Publication 20130041626A1.

Cox, D. R., & Miller, H. D. (1977). *The theory of stochastic processes.* Boca Raton, FL: Chapman & Hall/CRC.

Duchesne, C., Liu, J. J., & MacGregor, J. F. (2012). Multivariate image analysis in the process industries: A review. *Chemometrics and Intelligent Laboratory Systems, 117*, 116–128.

Golosnoy, V., Ragulin, S., & Schmid, W. (2011). CUSUM control charts for monitoring optimal portfolio weights. *Computational Statistics and Data Analysis, 55*(11), 2991–3009.

Hawkins, D. M., & Olwell, D. H. (1998). *Cumulative sum charts and charting for quality improvement.* New York: Springer.

Hryniewicz, O., & Kaczmarek, K. (2016). Monitoring of short series of dependent observations using a control chart approach and data mining techniques. In S. Knoth (Ed.), *Proceedings of the XII International Workshop on Intelligent Statistical Quality Control* (pp. 143–161).

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Annals of Statistics, 14*(4), 1379–1387.

Negandhi, V., Sreenivasan, L., Giffen, R., Sewak, M., & Rajasekharan, A. (2015). *IBM predictive maintenance and quality 2.0 technical overview*. Armonk, NY: IBM Redbooks.

Philips, T., Yashchin, E., & Stein, D. (2003). Using statistical process control to monitor active managers. *Journal of Portfolio Management, 30*(1), 86–94.

Shi, J., & Zhou, S. (2009). Quality control and improvement for multistage systems: A survey. *IIE Transactions, 41*(9), 744–753.

Shmueli, G., & Burkom, H. (2010). Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics, 52*(1), 39–51.

Siegmund, D. (1985). *Sequential analysis*. New York: Springer.

Sparks, R. (2015). Social network monitoring: Aiming to identify periods of unusually increased communications between parties of interest. In S. Knoth & W. Schmid (Eds.), *Frontiers in statistical quality control* (Vol. 11, pp. 3–13). Berlin: Springer.

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

Yashchin, E. (1985). On analysis and design of Cusum-Shewhart control schemes. *IBM Journal of Research and Development, 29*, 377–391.

Yashchin, E. (2012). Design and implementation of systems for monitoring lifetime data. In H. J. Lenz, W. Schmid & P. Wilrich (Eds.), *Frontiers in statistical quality control* (Vol. 10, pp. 171–195). Berlin: Springer.

Control Charts for Time-Dependent Categorical Processes

Christian H. Weiß

Abstract The monitoring of categorical processes received increasing research interest during the last years, but usually on the premise of the underlying process being serially independent. We start with a brief survey of approaches for modeling and analyzing serially dependent categorical processes. Then we consider two general strategies for monitoring a categorical process: if the process evolves too fast to be monitored continuously, then segments are taken in larger intervals and a corresponding statistic is plotted on a control chart; here, one has to carefully consider the serial dependence within the sample. If a continuous process monitoring is possible, then the serial dependence between the plotted statistics has to be taken into account. For both scenarios, we propose appropriate control charts and investigate their performance through simulations.

Keywords Attributes data · Categorical time series · Pearson chart · Gini chart · CUSUM chart · Literature survey

1 Introduction

Methods of *statistical process control* (SPC) help to monitor and improve processes in manufacturing and service industries. For such a process, relevant quality characteristics are measured at times $t \in \mathbb{N} = \{1, 2, \ldots\}$ thus leading to a stochastic process $(X_t)_{\mathbb{N}}$ of continuous-valued or discrete-valued random variables (*variables data* or *attributes data*, respectively). The most important SPC tool is the *control chart*, which requires the relevant quality characteristics to be measured online. Control charts are applied to a process operating in a stable state (*in control*), i.e., $(X_t)_{\mathbb{N}}$ is assumed to be stationary according to a specified in-control model. As a new measurement arrives, this is used to compute a statistic (possibly also incorporating

C. H. Weiß (✉)
Helmut Schmidt University, Department of Mathematics and Statistics, Hamburg, Germany
e-mail: weissc@hsu-hh.de

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_12

past values of the quality characteristic) which is then plotted on the chart with its control limits. If the statistic violates the limits, an alarm signals that the process may not be stable anymore (*out of control*) and requires corrective actions. More details about these terms and concepts can be found in the textbook by Montgomery (2009) or in the survey chapters by Woodall (2000), Woodall and Montgomery (2014).

In this article, we shall be concerned with a particular type of attributes data processes $(X_t)_{\mathbb{N}}$: the range of X_t is assumed to be of *categorical* nature. So X_t has a discrete and non-metric range consisting of a finite number $m + 1$ of categories with $m \in \mathbb{N}$ (*state space*). In some applications, the range exhibits at least a natural ordering; it is then referred to as an *ordinal* range. In other cases, not even such an inherent order exists (*nominal* range). Here, we shall consider this latter, most general case, i.e., even if there would be some ordering, we would not make use of it but assume that each random variable X_t takes one of a finite number of *unordered* categories. To simplify notations, it is assumed that the range of $(X_t)_{\mathbb{N}}$ is coded as $\mathcal{S} = \{0, \ldots, m\}$. But as emphasized before, this does *not* imply that there is any natural order between the states in $\mathcal{S}$, except a lexicographic order. In view of quality-related applications, X_t often describes the result of an inspection of an item, which either leads to classification $X_t = i$ for an $i = 1, \ldots, m$ iff the t^{th} item was non-conforming of type 'i', or $X_t = 0$ for a conforming item. In the example described by Mukhopadhyay (2008), a non-conforming ceiling fan cover is classified according to the most predominant type of paint defect, e.g., 'poor covering' or 'bubbles', while Ye et al. (2002) reports the monitoring of network traffic data with different types of audit events.

Since a few years, there seems to be increasing research interest in the monitoring of categorical processes, which manifests itself in some recent articles like Chen et al. (2011) (traditional χ^2-chart, see Sect. 3 below, but with additional inspection error), or Ryan et al. (2011) and Weiß (2012) (charts for continuous process monitoring, see Sect. 4 below); further references can be found in Woodall (1997), Topalidou and Psarakis (2009). But when looking for existing literature, it is important to precisely define the kind of data one is concerned with. In this article, we do not only concentrate on *unordered* categories, but also on *mutually exclusive* ones (i.e., different categories cannot appear together). This is in contrast to the recent articles by Li et al. (2012) and Yashchin (2012), which are "multivariate" in a sense by considering "multi-attribute processes". Finally, we restrict to statistical methods, while part of the literature is about methods based on fuzzy theory instead (Woodall 1997; Topalidou and Psarakis 2009).

Although more and more articles deal with categorical attributes data processes, there is one important restriction with all these works: the underlying process is assumed to be *serially independent* in its in-control state, i.e., $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.). Probably the main reason why researchers and practitioners are often ill at ease when being concerned with time-dependent categorical data is that concepts for expressing categorical forms of serial dependence are not well communicated yet, and also simple stochastic models for such processes, i.e., which are of a simplicity being comparable to that of the well-known autoregressive moving average (ARMA) models for autocorrelated variables

data processes, are not known to a broader audience. Therefore, we start in Sect. 2 with a brief survey of approaches for modeling and analyzing categorical processes. Then we consider two general strategies for monitoring a categorical process: if the process evolves too fast to be monitored continuously, one may take segments from the process at selected times. Then a statistic is computed from the resulting sample and plotted on a control chart, see Sect. 3. Here, it is important to carefully consider the serial dependence *within* the sample. In other cases, it is possible to continuously monitor the process, but then the serial dependence has to be taken into account *between* the plotted statistics, see Sect. 4. For any of these two scenarios, we propose appropriate control charts and investigate their performance through simulations. Finally, we outline possible directions for future research in Sect. 5.

2 Modeling and Analyzing Categorical Processes

If being concerned with stationary *real-valued* time series, then a huge toolbox for analyzing and modeling such time series is readily available and well-known to a broad audience. To highlight a few basic approaches, the time series is visualized by simply plotting the observed values against time, marginal properties such as location and dispersion may be measured in terms of mean and variance, and serial dependence is commonly quantified in terms of autocorrelation. Depending on the observed dependence structure, a model of the ARMA family itself might turn out to be appropriate, or one of its innumerable extensions, see the recent survey by Holan et al. (2010) or any textbook about time series analysis.

Things change if the available time series is *categorical*. In the ordinal case, a time series plot is still feasible by arranging the possible outcomes in their natural ordering along the Y axis, and location could be measured at least by the median. In the purely nominal case as considered in this article, not even these basic analytic tools are applicable. Therefore, tailor-made solutions are required when analyzing a (stationary) categorical process $(X_t)_{\mathbb{Z}}$ with range $\mathcal{S} = \{0, \ldots, m\}$, $m > 1$. In the sequel, we denote the time-invariant marginal probabilities by $\boldsymbol{\pi} := (\pi_0, \ldots, \pi_m)^\top$ with $\pi_i := P(X_t = i) \in (0; 1)$ and $\pi_0 = 1 - \pi_1 - \ldots - \pi_m$. As their sample counterpart, we consider the vector $\hat{\boldsymbol{\pi}}$ of relative frequencies computed from the process' segment $X_1, \ldots, X_T$ being of length T.

Although there are a few proposals for a *visual analysis* of a categorical time series (Weiß 2008), a reasonable substitute of the simple time series plot is still missing. But a number of non-visual tools are available. Concerning *location*, the (empirical) mode seems to be the only established solution. Categorical *dispersion* measures compare the actual marginal distribution with the two possible extremes of a one-point distribution (no dispersion; maximal concentration) and a uniform distribution (maximal dispersion; no concentration). Several measures have been proposed for this purpose, see the survey in Appendix A of Weiß and Göb (2008).

In the author's opinion, the (empirical) *Gini index*,

$$\nu_{\mathrm{G}} \;=\; \tfrac{m+1}{m}\,(1-\textstyle\sum_{j=0}^{m}\pi_j^2) \quad\text{and}\quad \hat{\nu}_{\mathrm{G}} \;=\; \tfrac{m+1}{m}\,\tfrac{T}{T-1}\,(1-\textstyle\sum_{j=0}^{m}\hat{\pi}_j^2), \tag{1}$$

is the most preferable dispersion measure, not only because of its simplicity, but also because of attractive stochastic properties of the empirical Gini index $\hat{\nu}_{\mathrm{G}}$ (like unbiasedness in the i.i.d. case; see Section 3 in Weiß (2013a) for a detailed discussion). The theoretical Gini index ν_{G} has range $[0;1]$, where increasing values indicate increasing dispersion, with the extremes $\nu_{\mathrm{G}} = 0$ iff X_t has a one-point distribution, and $\nu_{\mathrm{G}} = 1$ iff X_t has a uniform distribution.

Since autocorrelation is not defined in the categorical case, several alternative measures of *serial dependence* have been proposed, see the references in Weiß and Göb (2008), Weiß (2013a). These measures usually rely on lagged bivariate probabilites, $p_{ij}(k) := P(X_t = i, X_{t-k} = j)$, with the empirical counterpart $\hat{p}_{ij}(k)$ being the relative frequency of (i,j) within the pairs $(X_{k+1},X_1),\ldots,(X_T,X_{T-k})$. Again, there seems to be a preferable solution, namely (empirical) *Cohen's* κ

$$\kappa(k) \;=\; \frac{\sum_{j=0}^{m}\left(p_{jj}(k)-\pi_j^2\right)}{1-\sum_{j=0}^{m}\pi_j^2} \quad\text{and}\quad \hat{\kappa}(k) \;:=\; \frac{1}{T} \;+\; \frac{\sum_{j=0}^{m}\left(\hat{p}_{jj}(k)-\hat{\pi}_j^2\right)}{1-\sum_{j=0}^{m}\hat{\pi}_j^2}. \tag{2}$$

The range of $\kappa(k)$ is given by $[-\frac{\sum_{j=0}^{m}\pi_j^2}{1-\sum_{j=0}^{m}\pi_j^2}\;;\;1]$, where 0 corresponds to serial independence. So it includes both positive and negative values in analogy to the range of the autocorrelation function. In fact, Weiß and Göb (2008) argued that $\kappa(k)$ is a measure of *signed* serial dependence: While we have perfect (unsigned) serial dependence at lag $k \in \mathbb{N}$ iff for any j, the conditional distribution $p_{\cdot|j}(k)$ is a one-point distribution, we have perfect *positive* (*negative*) dependence iff all $p_{i|i}(k) = 1$ (all $p_{i|i}(k) = 0$). So like positive autocorrelation implies that large values tend to be followed by large values, for instance, positive dependence implies that the process tends to stay in the state it has reached (and vice versa). Besides this analogy to the autocorrelation function, again the empirical version, $\hat{\kappa}(k)$, has attractive properties (also see below). Among others, it is nearly unbiased in the i.i.d. case, and its distribution is well approximated by the normal distribution $\mathrm{N}(0,\sigma^2)$ with $T\sigma^2 = 1-\left(1+2\sum_{j=0}^{m}\pi_j^3-3\sum_{j=0}^{m}\pi_j^2\right)/\left(1-\sum_{j=0}^{m}\pi_j^2\right)^2$, which, in turn, allows to test for significant serial dependence in a categorical time series (Weiß 2011).

Next, we turn to the question of how to model a categorical process. Perhaps the most obvious approach is to use a Markov model. $(X_t)_{\mathbb{Z}}$ is said to be a p^{th} order *Markov process* with $\mathrm{p} \in \mathbb{N}$ if for all t and for each $x_t \in \mathcal{S}$, we have

$$P(X_t = x_t \mid X_{t-1} = x_{t-1}, \ldots) \;=\; P(X_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_{t-\mathrm{p}} = x_{t-\mathrm{p}}). \tag{3}$$

The special case $\mathrm{p} = 1$ ("memory of length 1") is usually referred to as a *Markov chain*, with its stochastic properties being solely determined by the (one-step)

transition probabilities $p_{i|j} = P(X_t = i \mid X_{t-1} = j)$ or the corresponding transition matrix $\mathbf{P} = (p_{i|j})_{i,j}$, respectively (Feller 1968, Chapter XV). General p^{th} order Markov processes (i.e., where the conditional probabilities are not further restricted by parametric assumptions), however, have the practical disadvantage of a huge number of model parameters, $(m+1)^{\mathrm{p}} \cdot m$. For this reason, more parsimonious models for categorical processes have been proposed in the literature, e.g., the variable length Markov model by Bühlmann and Wyner (1999) or the mixture transition distribution model by Raftery (1985).

An even more parsimonious model class, which also allows for non-Markovian forms of serial dependence, are the new discrete ARMA (NDARMA) models by Jacobs and Lewis (1983),[1] which are motivated by the standard ARMA models for real-valued processes. As shown in Weiß and Göb (2008), the NDARMA process $(X_t)_{\mathbb{Z}}$ can be defined as follows:

Let $(\epsilon_t)_{\mathbb{Z}}$ be i.i.d. with marginal distribution $\boldsymbol{\pi}$ and, independently, let

$$\boldsymbol{D}_t \;=\; (\alpha_{t,1}, \ldots, \alpha_{t,\mathrm{p}}, \beta_{t,0}, \ldots, \beta_{t,\mathrm{q}})$$

be a $(\mathrm{p}+\mathrm{q}+1)$-dimensional vector, where exactly one of the components takes the value 1 (either an $\alpha_{t,i}$ with probability ϕ_i or a $\beta_{t,j}$ with probability φ_j; $\phi_1 + \ldots + \varphi_{\mathrm{q}} = 1$) and all others are equal to 0. Both ϵ_t and $\boldsymbol{D}_t$ are assumed to be independent of $(X_s)_{s<t}$. Then $(X_t)_{\mathbb{Z}}$ defined by the random mixture

$$X_t \;=\; \alpha_{t,1} \cdot X_{t-1} + \ldots + \alpha_{t,\mathrm{p}} \cdot X_{t-\mathrm{p}} \;+\; \beta_{t,0} \cdot \epsilon_t + \ldots + \beta_{t,\mathrm{q}} \cdot \epsilon_{t-\mathrm{q}} \tag{4}$$

is said to be an *NDARMA process* of order (p, q).

Although written down in an ARMA-like manner, recursion (4) states that X_t chooses either one of $X_{t-1}, \ldots, X_{t-\mathrm{p}}$ or $\epsilon_t, \ldots, \epsilon_{t-\mathrm{q}}$. Therefore, this approach is applicable to categorical processes. In fact, it can be applied to any kind of processes, but already for ordinal data, the selection mechanism would not be very plausible anymore because it is not able to deal with an order between the possible outcomes. If $\mathrm{q} > 0$, then $(X_t)_{\mathbb{Z}}$ is not Markovian, while the model order (p, 0) leads to a special type of p^{th} order Markov process, the *DAR process* of order p. In the latter case, the transition probabilities are given by

$$P(X_t = x_0 \mid X_{t-1} = x_1, \ldots, X_{t-\mathrm{p}} = x_{\mathrm{p}}) \;=\; \varphi_0 \pi_{x_0} + \textstyle\sum_{r=1}^{\mathrm{p}} \delta_{x_0,x_r} \phi_r, \tag{5}$$

where $\delta_{a,b}$ denotes the Kronecker delta. Generally, the NDARMA process is stationary with marginal distribution $\boldsymbol{\pi}$, and if serial dependence is measured in terms of Cohen's κ, then $\kappa(k)$ satisfies a set of Yule-Walker-type equations in analogy to the standard ARMA case (Weiß and Göb 2008):

$$\kappa(k) \;=\; \textstyle\sum_{j=1}^{\mathrm{p}} \phi_j \cdot \kappa(|k-j|) \;+\; \sum_{i=0}^{\mathrm{q}-k} \varphi_{i+k} \cdot r(i) \qquad \text{for } k \geq 1, \tag{6}$$

[1] The ARMA model discussed by Biswas and Song (2009) is equivalent to the NDARMA model.

where the $r(i)$ are determined by $r(i) = 0$ for $i < 0$, $r(0) = \varphi_0$, and

$$r(i) \;=\; \sum_{j=\max\{0,i-\mathrm{p}\}}^{i-1} \phi_{i-j} \cdot r(j) \;+\; \sum_{j=0}^{\mathrm{q}} \delta_{i,j} \cdot \varphi_j \qquad \text{for } i > 0.$$

This implies to use the empirical version, $\hat{\kappa}(k)$, not only for uncovering significant serial dependence, but also for identifying the model order of an NDARMA process and for estimating the model parameters in analogy to the method of moments. The empirical analyses in Weiß (2013a), Maiti and Biswas (2018) showed that $\hat{\kappa}(k)$ is often better suited for this purpose than alternative measures of serial dependence.

3 Sample-Based Monitoring of Categorical Processes

From now on, we turn to the question of monitoring a categorical process. If the process $(X_t)_{\mathbb{N}}$ cannot be monitored continuously, then (non-overlapping) segments $X_{t_k}, \ldots, X_{t_k+n-1}$ from the process (of a certain length $n > 1$, taken at times $t_1, t_2, \ldots$ with $t_k - t_{k-1} > n$ sufficiently large) are analyzed and evaluated. Here, it is important to carefully consider the serial dependence *within* the segments. But since the time distance $t_k - t_{k-1}$ between successive segments is assumed to be quite large, at least the serial dependence *between* the segments can be ignored.

Remark 1 (Bulk Sampling) As stated before, we shall assume the sample size $n > 1$ in the sequel. The reviewer pointed out that in the field of bulk sampling, also the inspection of single items (i.e., $n = 1$) is common, still with negligible between-sample dependence due to a large distance between successive items (exceeding the "correlation length"). The methods described in this section rely on frequencies, so sample size $n > 1$ is essential. But the methods described in Sect. 4 could be used instead for bulk sampling, because we may understand the case $n = 1$ as a continuous monitoring of the virtually i.i.d. process $(X_{t_k})_{k\in\mathbb{N}}$.

3.1 Sample-Based Monitoring: Binary Case

In the special case of a binary process with range $\{0, 1\}$, one commonly determines either the sample sum $N_k^{(n)} = X_{t_k} + \ldots + X_{t_k+n-1}$ (e.g., count of non-conforming items) or the corresponding sample fraction of '1's. Then this count or fraction is either plotted directly on a Shewhart-type chart (*np chart* or *p chart*, respectively, see Montgomery (2009)), or this quantity is used for an advanced control scheme like an exponentially weighted moving average (EWMA) chart or cumulative sum (CUSUM) chart, see Gan (1990, 1993) for instance.

Concerning the distribution of the sample count (the sample fraction differs from the count only by the factor $1/n$), the serial dependence structure of the underlying binary process $(X_t)_{\mathbb{N}}$ is important. If $(X_t)_{\mathbb{N}}$ is i.i.d. with $P(X_t = 1) = \pi \in (0; 1)$

(e.g., probability for a non-conforming item), then each sample sum $N_k^{(n)} = X_{t_k} + \ldots + X_{t_k+n-1}$ is binomially distributed according to $\mathrm{Bin}(n, \pi)$, and the statistics $(N_k^{(n)})_{\mathbb{N}}$ constitute themselves an i.i.d. process of binomial counts. But if $(X_t)_{\mathbb{N}}$ exhibits serial dependence, in contrast, the distribution of $N_k^{(n)}$ will deviate from a binomial one.

In Deligonul and Mergen (1987), Bhat and Lal (1990), Weiß (2009), the case of $(X_t)_{\mathbb{N}}$ being a binary Markov chain with success probability $\pi \in (0; 1)$ and autocorrelation parameter $\rho \in \left(\max\{\frac{-\pi}{1-\pi}, -\frac{1-\pi}{\pi}\}; 1\right)$ is considered, i.e., with transition matrix

$$\mathbf{P} = \begin{pmatrix} p_{0|0} & p_{0|1} \\ p_{1|0} & p_{1|1} \end{pmatrix} = \begin{pmatrix} (1-\pi)(1-\rho)+\rho & (1-\pi)(1-\rho) \\ \pi(1-\rho) & \pi(1-\rho)+\rho \end{pmatrix}. \tag{7}$$

In this case, $N_k^{(n)} = X_{t_k} + \ldots + X_{t_k+n-1}$ follows the so-called *Markov binomial distribution* $\mathrm{MB}(n, \pi, \rho)$ (which coincides with $\mathrm{Bin}(n, \pi)$ iff $\rho = 0$). While the mean of $N_k^{(n)}$ is not affected by the serial dependence, especially the variance changes (*extra-binomial variation* if $\rho > 0$):

$$E\left[N_k^{(n)}\right] = n\pi, \qquad V\left[N_k^{(n)}\right] = n\pi(1-\pi)\,\frac{1+\rho}{1-\rho}\,\underbrace{\left(1 - \frac{2\rho(1-\rho^n)}{n(1-\rho^2)}\right)}_{\approx 1 \text{ for large } n};$$

these and further well-known properties of the MB-distribution are summarized in Table II in Weiß (2009). If the time distance $t_k - t_{k-1}$ between successive segments from $(X_t)_{\mathbb{N}}$ is sufficiently large, the resulting process of counts $(N_k^{(n)})_{\mathbb{N}}$ can still be assumend to be approximately i.i.d. (note that the correlation $\rho^{|t-s|}$ between X_t and X_s decays exponentially), but with a marginal distribution being different from a binomial one. This difference in the distribution of $N_k^{(n)}$ certainly has to be considered very carefully when designing a corresponding control chart (see Weiß (2009) for the case of an *np* or EWMA chart). An alternative approach was recently proposed by Adnaik et al. (2015), who do not use the sample sums $N_k^{(n)}$ as the chart's statistics, but compute a likelihood ratio statistic for each segment.

3.2 Sample-Based Monitoring: i.i.d. Case

Let us return to the truly categorical case, i.e., where the range of $(X_t)_{\mathbb{N}}$ consists of more than two states, $\mathcal{S} = \{0, \ldots, m\}$ with $m > 1$, and has time-invariant marginal probabilities $\boldsymbol{\pi} := (\pi_0, \ldots, \pi_m)^\top$, see Sect. 2. If the number of different states, $m + 1$, is small, it would be feasible to monitor the process by m simultaneous binary charts, e.g., by using the *p-tree method* described in Duran and Albin (2009). But here, we shall concentrate on such charting procedures, where the information about the process is comprised in a univariate statistic: after having taken a sample

or segment from the process, we first compute the resulting frequency distribution as a summary, which then serves as the base for deriving the statistic to be plotted on the control chart. To keep it consistent with the binary case from before, we concentrate on absolute frequencies: $\boldsymbol{N}_k^{(n)} = (N_{k;\,0}^{(n)}, \ldots, N_{k;\,m}^{(n)})^\top$ with $N_{k;\,i}^{(n)}$ being the absolute frequency of the state 'i' in the sample $X_{t_k}, \ldots, X_{t_k+n-1}$ such that $N_{k;\,0}^{(n)} + \ldots + N_{k;\,m}^{(n)} = n$. If the underlying categorical process $(X_t)_\mathbb{N}$ is serially independent (hence i.i.d.), then the distribution of each $\boldsymbol{N}_k^{(n)}$ is a multinomial one.

Remark 2 (Multinomial Distribution) The *multinomial distribution* is defined by summing up n independent copies of a binary random vector $\boldsymbol{Y}$, where exactly one of the components takes the value 1, all others are equal to 0. So the possible range of $\boldsymbol{Y}$ consists of the unit vectors $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_m \in \{0, 1\}^{m+1}$, where $\boldsymbol{e}_j = (e_{j,0}, \ldots, e_{j,m})^\top$ is defined by $e_{j,i} = \delta_{j,i}$ ($\boldsymbol{e}_j$ has a one in its j^{th} component) for $j = 0, \ldots, m$, and $P(\boldsymbol{Y} = \boldsymbol{e}_j) = \pi_j$ is assumed. Then $\boldsymbol{N} := \sum_{i=1}^n \boldsymbol{Y}_i \sim \text{MULT}(n;\ \pi_0, \ldots, \pi_m)$ has the range $\left\{\boldsymbol{n} \in \{0, \ldots, n\}^{m+1} \;\middle|\; n_0 + \ldots + n_m = n\right\}$, and its probability mass function (PMF) is given by

$$P(\boldsymbol{N} = \boldsymbol{n}) \;=\; \binom{n}{n_0, \ldots, n_m} \cdot \pi_0^{n_0} \cdots \pi_m^{n_m}.$$

The covariance matrix equals

$$n \cdot \boldsymbol{\Sigma}, \quad \text{where } \boldsymbol{\Sigma} = (\sigma_{ij}) \quad \text{is given by } \sigma_{ij} \;=\; \begin{cases} \pi_i(1 - \pi_i) & \text{if } i = j, \\ -\pi_i \pi_j & \text{if } i \neq j. \end{cases}$$

Each component N_j of $\boldsymbol{N}$ is binomially distributed according to $\text{Bin}(n, \pi_j)$.

The importance of the multinomial distribution for i.i.d. categorical samples arises from the fact that the binary random vector $\boldsymbol{Y}$ can be understood as a *binarization* of a categorical random variable X, by defining $\boldsymbol{Y} = \boldsymbol{e}_j$ if $X = j$. Then $\boldsymbol{N}$ represents the realized absolute frequencies of n independent replications of X.

So according to Remark 2, the categorical process $(X_t)_\mathbb{N}$ might be represented equivalently by the process $(\boldsymbol{Y}_t)_\mathbb{N}$ of its binarizations, and hence $\boldsymbol{N}_k^{(n)} = \boldsymbol{Y}_{t_k}^{(n)} + \ldots + \boldsymbol{Y}_{t_k+n-1}^{(n)}$ in analogy to the above binary situation.

Using that $\boldsymbol{N}_k^{(n)}$ is multinomially distributed if $(X_t)_\mathbb{N}$ is i.i.d., Duncan (1950), Marcucci (1985), Nelson (1987), and Mukhopadhyay (2008) proposed to plot *Pearson's χ^2-statistic* on a control chart,

$$C_k^{(n)} \;=\; \sum_{j=0}^{m} \frac{(N_{k;\,j} - n\,\pi_{0;\,j})^2}{n\,\pi_{0;\,j}}, \tag{8}$$

where $\boldsymbol{\pi}_0 := (\pi_{0;\,0}, \ldots, \pi_{0;\,m})^\top$ refers to the in-control values of the categorical probabilities. So in the in-control case, the process $(C_k^{(n)})_\mathbb{N}$ is i.i.d. with a marginal distribution that might be approximated by a χ^2_m-distribution (Horn 1977).

This approximate distribution may be used for chart design, i.e., for finding an appropriate upper control limit u_C.

As an alternative, Weiß (2012) proposed to use a control statistic based on a categorical dispersion measure such as the *Gini index* (1). This suggestion is motivated by the fact that for most production processes, the probability of a unit being conforming, say $\pi_{0;\,0}$, is much larger than any defect probability, i.e., $\pi_{0;\,0} \gg \pi_{0;\,1}, \ldots, \pi_{0;\,m}$ and thus we have low categorical dispersion. A relevant out-of-control scenario, in turn, will be one where π_0 gets reduced, while $\pi_1, \ldots, \pi_m$ are increased (leading to increased categorical dispersion). Therefore, an upper-sided Gini chart is reasonable for quality-related applications. If $(X_t)_{\mathbb{N}}$ is i.i.d., following the in-control model, then

$$G_k^{(n)} = \frac{1 - n^{-2} \sum_{j=0}^{m} N_{k;\,j}^2}{1 - \sum_{j=0}^{m} \pi_{0;\,j}^2} \tag{9}$$

is approximately normally distributed with mean $1 - 1/n$ and variance $\frac{4}{n}\left(\sum_{j=0}^{m} \pi_{0;\,j}^3 - (\sum_{j=0}^{m} \pi_{0;\,j}^2)^2\right)/\left(1 - \sum_{j=0}^{m} \pi_{0;\,j}^2\right)^2$, see Weiß (2011), which can be used to determine an appropriate upper limit u_G.

Remark 3 (np Chart) In the situation described before, where π_0 expresses the probability of a unit being conforming and where $\pi_1, \ldots, \pi_m$ are the defect probabilities, a further alternative for process monitoring could be to use the *np* chart from Sect. 3.1 by only distinguishing between conforming and non-conforming. Certainly, we loose the information about the defects' distribution with such a monitoring strategy, but we shall include it as a benchmark in our performance analyses in Sect. 3.4.

Remark 4 (Ordinal Data) As already briefly pointed out in Sect. 1, in some applications, the possible categories might exhibit an inherent order, i.e., the categorical data are indeed *ordinal* data. All control charts discussed in this article could be applied to such ordinal data, too. In fact, such an example is given by Marcucci (1985), where the above χ^2-chart (designed for nominal data) is applied to ordinal data from a brick manufacturing process. However, the ordinal nature of the data is completely ignored by such a monitoring approach.

There are a few proposals for sample-based control charts, which make use of the inherent order in the range of an i.i.d. ordinal process. Tucker et al. (2002) assume a latent variable Z_t with a continuous distribution behind each ordinal observation X_t, e.g., following a normal distribution. The real axis is partitioned into $m+1$ intervals, and if (the unobservable) Z_t falls into the j^{th} interval, then X_t takes the category j. To obtain a control statistic from the k^{th} sample, the maximum likelihood estimate (MLE) of the location parameter of Z_t's distribution is computed, and the standardized MLE is then plotted on a control chart.

Another approach is used by Cozzucoli (2009), who picks up the idea of a demerits control chart (Jones et al. 1999). Each category is assigned a weight, which reflects the severeness of the respective type of quality defect (and which accounts for the ordinality of the range in this way). Using these weights, the control statistic for the k^{th} sample is defined as a weighted sum of the observed defect proportions.

We conclude this section by pointing out the relationship between the sample frequencies and so-called *compositional data*.

Remark 5 (Compositional Data) If the number n of replications becomes very large, say $n \to \infty$, then the vector of random proportions becomes a continuous random vector with the $(m+1)$-*part unit simplex* as its range,

$$\mathbb{S}^{m+1} := \left\{ \boldsymbol{x} \in (0;1)^{m+1} \,\middle|\, x_0 + \ldots + x_m = 1 \right\}.$$

The corresponding data, which express the "proportions of some whole" (Aitchison 1986, p. 1), are referred to as *compositional data (CoDa)*. Excellent books about this topic are the ones by Aitchison (1986); Pawlowsky-Glahn and Buccianti (2011). Approaches for monitoring i.i.d. compositional data have been investigated by Boyles (1997) and Vives-Mestres et al. (2014a,b).

3.3 Sample-Based Monitoring of Serially Dependent Categorical Processes

From now on, we allow $(X_t)_{\mathbb{N}}$ to be serially dependent. Then, in general, the distribution of $\boldsymbol{N}_k^{(n)}$ will not be multinomial anymore, and consequently, also the distributions of $C_k^{(n)}$ and $G_k^{(n)}$ will deviate from the ones given above for the i.i.d. case. As argued in Weiß (2012), especially $C_k^{(n)}$ is extremely sensitive with respect to serial dependence. This is also illustrated by the asymptotic results in Weiß (2013a), which refer to the case of an underlying NDARMA process (see (4) before). If we define the model-dependent constant (remember the Yule-Walker equations (6) for Cohen's κ (2))

$$c := 1 + 2 \cdot \sum_{i=1}^{\infty} \kappa(i) < \infty \qquad (c = 1 \text{ in the i.i.d. case}),$$

then $C_k^{(n)}/c$ is approximately χ_m^2-distributed, and the distribution of $G_k^{(n)}$ is still approximately normal, but with the mean being deflated by the factor $(n-c)/(n-1)$ and the variance being inflated by the factor c (Weiß 2013a).

For illustration, we discuss the example of an underlying DAR(1) process (as an instance of a Markov chain) in more detail. To keep the notation consistent with the above binary Markov chain, we denote $\rho := \phi_1$. Using formula (5), the transition

matrix of $(X_t)_{\mathbb{N}}$ follows as

$$\mathbf{P} \;=\; (p_{i|j})_{i,j} \;=\; \begin{pmatrix} \pi_0(1-\rho)+\rho & \pi_0(1-\rho) & \cdots & \pi_0(1-\rho) \\ \pi_1(1-\rho) & \pi_1(1-\rho)+\rho & & \vdots \\ \vdots & & \ddots & \\ \pi_m(1-\rho) & \pi_m(1-\rho) & \cdots & \pi_m(1-\rho)+\rho \end{pmatrix}, \qquad (10)$$

and we have $c = (1+\rho)/(1-\rho)$ since $\kappa(i) = \rho^i$ according to (6). The distribution of $\boldsymbol{N}_k{}^{(n)}$ is called the *Markov multinomial distribution* by Wang and Yang (1995), say $\mathrm{MM}(n;\ \pi_0, \ldots, \pi_m;\ \rho)$. A closed-form formula for the joint probability generating function of $\boldsymbol{N}_k{}^{(n)}$ is provided by Wang and Yang (1995). An asymptotic approximation of the distribution is derived in Weiß (2013a), a normal distribution with mean vector $n\boldsymbol{\pi}$ and covariance matrix $c \cdot n\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is given as in Remark 2. So compared to the multinomial distribution (case $\rho = 0$), the (asymptotic) covariance matrix of $\mathrm{MM}(n;\ \pi_0, \ldots, \pi_m;\ \rho)$ is inflated by the factor c. Note that the j^{th} component $N_{k;\,j}{}^{(n)}$ follows the $\mathrm{MB}(n, \pi_j, \rho)$ distribution, since for this particular type of Markov chain, also each component of the binarization $(\boldsymbol{Y}_t)_{\mathbb{N}}$ is itself a binary Markov chain.

Remark 6 (Multinomial CUSUM Chart) Besides plotting the statistics $C_k{}^{(n)}$ or $G_k{}^{(n)}$ on a Shewhart-type control chart, one may also consider a type of CUSUM control chart (Page 1954) as an alternative. Generally, such CUSUM charts are known to be more sensitive to small changes in the process, since they accumulate information about the process' past in contrast to the memoryless Shewhart charts. Picking up a proposal by Steiner et al. (1996) and Ryan et al. (2011) defined a multinomial CUSUM chart based on the log-likelihood ratio of the process $(\boldsymbol{N}_k{}^{(n)})_{k\in\mathbb{N}}$ (such an approach was also considered by Höhle (2010) in the context of a categorical logit model). Due to $(\boldsymbol{N}_k{}^{(n)})_{\mathbb{N}}$ being i.i.d., the contribution to the log-likelihood ratio by the k^{th} sample simply equals $L_k = \ln\left(P_{\boldsymbol{\pi}_1}(\boldsymbol{N}_k{}^{(n)})/P_{\boldsymbol{\pi}_0}(\boldsymbol{N}_k{}^{(n)})\right)$, where $\boldsymbol{\pi}_1$ expresses a likely out-of-control scenario that is to be detected. Furthermore, since $(X_t)_{\mathbb{N}}$ is i.i.d., $\boldsymbol{N}_k{}^{(n)}$ is multinomially distributed (Remark 2), so the expression for L_k simplifies to

$$L_k \;=\; \sum_{j=0}^{m} N_{k;\,j}^{(n)} \,\ln\frac{\pi_{1;\,j}}{\pi_{0;\,j}}.$$

Now the CUSUM statistics are defined in the usual way as $S_k = \max\{0,\ S_{k-1} + L_k\}$.

The CUSUM statistics are easily computed in the above i.i.d. situation, and as shown by Ryan et al. (2011), the CUSUM chart quickly detects an out-of-control situation provided that this situation is in the direction anticipated by $\boldsymbol{\pi}_1$. Things change, however, if the underlying process $(X_t)_{\mathbb{N}}$ becomes serially dependent. As we have seen before, a closed-form formula for the PMF of $\boldsymbol{N}_k{}^{(n)}$ is not yet known even in the case of the rather simple Markov dependence. As a consequence,

the computation of the CUSUM statistics becomes difficult. An exception is the boundary case $n = 1$ (continuous process monitoring, see Sect. 4 below); a feasible CUSUM chart for the case $n > 1$ (truly sample-based monitoring) appears to be a relevant issue for future research.

3.4 Sample-Based Monitoring: ARL Performance

Design and performance of the Pearson chart (8) with upper limit u_C as well as of the Gini chart (9) with upper limit u_G are investigated through simulations. As some relevant in-control scenarios, we choose marginal distributions that have already been analyzed in the literature, namely $\boldsymbol{\pi}_0 = (0.54, 0.25, 0.12, 0.09)^\top$ (Duncan 1950), $\boldsymbol{\pi}_0 = (0.65, 0.24, 0.07, 0.04)^\top$, $(0.83, 0.104, 0.04, 0.026)^\top$, $(0.99, 0.005, 0.004, 0.001)^\top$ (Cozzucoli 2009), and $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.021, 0.023, 0.022, 0.025)^\top$ (Mukhopadhyay 2008), with dispersion $\nu_G \approx 0.831$, 0.685, 0.397, 0.026 and 0.463, respectively. For these marginals, we consider both the i.i.d. case ($\rho = 0$) as well as DAR(1) dependence with parameter value $\rho > 0$. While the serial dependence within the samples $X_{t_k}, \ldots, X_{t_k+n-1}$ being used for computing ${C_k}^{(n)}$ and ${G_k}^{(n)}$, respectively, is explicitly considered, we assume that the resulting processes $({C_k}^{(n)})_{\mathbb{N}}$ and $({G_k}^{(n)})_{\mathbb{N}}$ are i.i.d. (since the time distance $t_k - t_{k-1}$ between successive samples is sufficiently large). So as for any Shewhart chart, we can define u_C and u_G as appropriate quantiles from the in-control distributions of ${C_k}^{(n)}$ and ${G_k}^{(n)}$, respectively. Since the ARL is computed as

$$\mathrm{ARL}_C(\boldsymbol{\pi}) = \frac{1}{P_{\boldsymbol{\pi}}({C_k}^{(n)} > u_C)} \quad \text{and} \quad \mathrm{ARL}_G(\boldsymbol{\pi}) = \frac{1}{P_{\boldsymbol{\pi}}({G_k}^{(n)} > u_G)},$$

respectively, we always determine the $(1 - 1/\mathrm{ARL}_0)$-quantile for a specified in-control level ARL_0. Here, we choose $\mathrm{ARL}_0 \in \{100, 200, 370, 500\}$, and the sample size as $n \in \{50, 100, 150, 200, 250\}$.

Remark 7 (ARL vs. ATS) An ARL-based chart design has to be treated with some caution. If we have fixed sampling intervals $t_k - t_{k-1} = K > n$, say $t_k := k \cdot K - n + 1$, for instance, and if the chart triggers its first alarm after plotting the r^{th} sample statistic (corresponds to run length r), then the number of manufactured items until this alarm is much larger, given by $r \cdot K$. Therefore, it would be preferable to look at the *average time to signal (ATS)* instead, where "time" refers to the original process $(X_t)_{\mathbb{N}}$, not to the number of plotted statistics. In the given example, we have $\mathrm{ATS} = K \cdot \mathrm{ARL}$. But for the sake of simplicity, we shall continue the simulation study by considering the ARL performance of the control charts.

The main focus of our investigations is on finding an appropriate in-control design. For this purpose, 1 million i.i.d. samples ${\boldsymbol{N}_k}^{(n)}$ are simulated for each situation,

Table 1 Asymptotic compared to exact chart design of Pearson and Gini chart for $n = 150$, $\boldsymbol{\pi}_0 = (0.83, 0.104, 0.04, 0.026)^\top$, $\mathrm{ARL}_0 = 370$

ρ	$\mathrm{ARL}_{C;\,\mathrm{as}}$	$u_{C;\,\mathrm{as}}$	u_C	$\mathrm{ARL}_{G;\,\mathrm{as}}$	$u_{G;\,\mathrm{as}}$	u_G
0	221.1	14.154	15.554	476.4	1.4250	1.4147
0.25	128.4	23.590	30.512	467.5	1.5462	1.5317
0.5	84.0	42.462	63.989	605.3	1.7277	1.6933
0.75	50.9	99.079	189.423	1508.3	2.0955	1.9677

and $C_k^{(n)}$ and $G_k^{(n)}$ are always computed. Then we determine

- the true ARL if deriving u_C, u_G from the asymptotic approximations, and
- the true limits u_C, u_G as the (empirical) $(1 - 1/\mathrm{ARL}_0)$-quantiles.

The complete tables of control limits and ARLs are available from the author upon request; here, we just summarize and illustrate the main findings. First of all, in nearly any case, the asymptotic approximation of u_C or u_G is rather bad, so these approximations can only be recommended as a starting value when searching for the true value. For the Pearson chart (8), the asymptotic limits are always too small (hence, also the true in-control ARL becomes too small), and the difference becomes worse with decreasing n, with decreasing dispersion in $\boldsymbol{\pi}_0$, and with increasing ρ. For the Gini chart (9), in contrast, except for situation $\boldsymbol{\pi}_0 = (0.99, 0.005, 0.004, 0.001)^\top$, the asymptotic limits are always too large, and now worse with increasing dispersion in $\boldsymbol{\pi}_0$, see Table 1 as an example. In the case of distribution $\boldsymbol{\pi}_0 = (0.99, 0.005, 0.004, 0.001)^\top$ with its extremely low degree of dispersion, we have $u_{G;\,\mathrm{as}} < u_G$.

Next, we analyze the effect of serial dependence in more detail. Table 1 already indicated that the actual dependence level ρ has to be considered when designing the control chart (widened limits for increasing ρ). In fact, if we just take the i.i.d. design ($\rho = 0$) but apply it to a DAR(1) process with $\rho > 0$, the resulting ARL is severely affected. Already values of ρ being only slightly above 0 lead to an enormous decrease in the ARL, independent of the marginal distribution $\boldsymbol{\pi}_0$ and of the sample size n, but even more severely for the Pearson chart (8) than for the Gini chart (9). This is illustrated by Fig. 1, which shows the ARL against ρ in the situation $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.021, 0.023, 0.022, 0.025)^\top$ (Mukhopadhyay 2008) with $n = 150$ and $\mathrm{ARL}_0 = 370$. On the other hand, this implies that especially the Pearson chart might be used for uncovering increases in ρ. Figure 1 also includes the ARLs of the upper-sided *np* chart (dotted line) as a benchmark, see Remark 3, which are quite close to those of the Gini chart. Note that for $\rho > 0$, the ARLs of the *np* chart are determined by the Markov binomial distribution $\mathrm{MB}(n, 1 - \pi_0, \rho)$, see the discussion before Remark 6 as well as Table II in Weiß (2009), and can thus be computed numerically.

Even if the chart design is chosen appropriately with respect to the serial dependence level ρ, we usually will observe an effect on the out-of-control performance. As an example, assume that the probability π_0 of having no

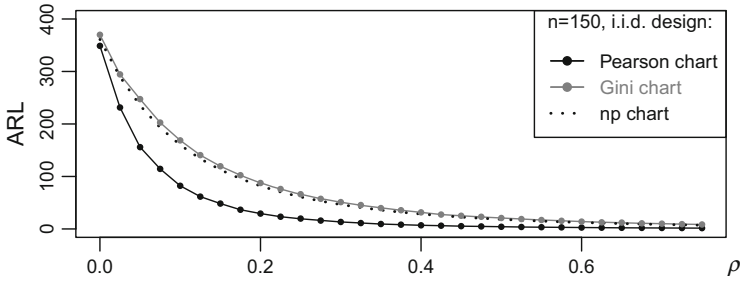


Fig. 1 ARL performance of Pearson ($u_C = 22.3043$), Gini ($u_G = 1.324642$) and *np* chart ($u_{np} = 49$) for $n = 150$, $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.021, 0.023, 0.022, 0.025)^\top$

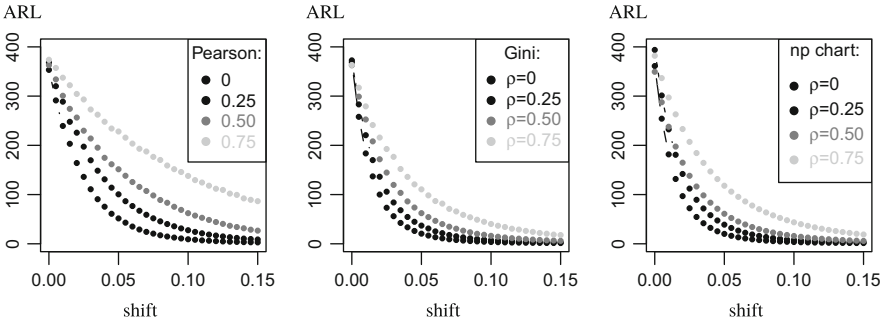


Fig. 2 ARL performance of Pearson, Gini and *np* chart ($\text{ARL}_0 \approx 370$) concerning $\pi_{1;\,0} = (1 - \text{shift})\,\pi_{0;\,0}$, $n = 150$, $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.021, 0.023, 0.022, 0.025)^\top$

defect is shifted downwards by a certain relative amount, i.e., $\pi_{1;\,0} = (1 - \text{shift})\,\pi_{0;\,0}$, and all other probabilities are increased in equal measure, $\pi_{1;\,k} = \frac{1-\text{shift}\cdot\pi_{0;\,0}}{1-\pi_{0;\,0}}\,\pi_{0;\,k}$. Independent of the marginal distribution $\boldsymbol{\pi}_0$, it can be observed that the out-of-control performance becomes worse for increasing ρ. As an illustration, Fig. 2 shows some ARL graphs for the marginal distribution $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.021, 0.023, 0.022, 0.025)^\top$, where all charts are designed to give roughly the same in-control ARL. Again the *np* chart is included as a benchmark, although its in-control ARLs show more variation than those of the other charts. For this particular out-of-control scenario, the Gini chart is preferable, which is reasonable since the dispersion strongly increases with increasing shift size (the ARL performance is again similar to that of the *np* chart). In some other scenarios, e.g., if $\pi_{1;\,k} = \pi_{0;\,k}$ for $k = 1, \ldots, m-1$ and $\pi_{1;\,m} = \pi_{0;\,m} + \pi_{0;\,0} - \pi_{1;\,0}$ as suggested by Cozzucoli (2009), the Pearson chart is superior (at least for larger shift amounts), but again with a worse performance for increasing ρ.

4 Continuous Monitoring of Categorical Processes

In this section, we consider the case, where a continuous monitoring of the categorical process $(X_t)_{\mathbb{N}}$ is possible. In this case, it would still be possible to form adjacent segments and to calculate the corresponding sample statistics, like for the strategy discussed in Sect. 3. Such a strategy, however, does not appear to be particularly useful: one would have to deal with two kinds of dependence, the within-sample dependence and the between-sample dependence, and the minimal delay of detecting an out-of-control situation could not become smaller than the sample size, also see the discussion in Remark 7. For this reason, only the following type of continuous monitoring is considered here: as a new categorical observation X_t arrives, the next statistic is computed and plotted on the control chart.

4.1 Continuous Monitoring: Binary Case

Again, we start by looking at the binary case first. Perhaps the most well-known approach for (quasi) continuously monitoring a binary process is by plotting run lengths on the chart, i.e., the number of '0's between two successive '1's (Bourke 1991; Xie et al. 2000). This is a reasonable approach especially for high-quality processes, where $\pi = P(X_t = 1)$ is very small. If '1's are observed more frequently, and hence the usual runs become quite short, one may modify the definition of a run, e.g., by waiting until the r^{th} occurrence of a '1' (Bourke 1991) or until the occurrence of a segment of '1's (Weiß 2013b). Bourke (1991) also proposed a CUSUM procedure to monitor the run lengths in $(X_t)_{\mathbb{N}}$. This *geometric CUSUM* control chart is essentially equivalent to the *Bernoulli CUSUM* control chart of Reynolds and Stoumbos (1999) and shall be discussed in some more detail below. Generally, while it is quite natural to check for runs in a binary process, it is more difficult to define a run for the truly categorical case in a reasonable way. One possible solution was discussed in Weiß (2012), but as pointed out there, also waiting times for different types of patterns might be relevant. Because of this ambiguity, we shall not further consider the monitoring of runs in a categorical process here.

Another approach for continuously monitoring a binary process would be the EWMA chart (Roberts 1959), which was applied to binary processes by, among others, Yeh et al. (2008) and Weiß and Atzmüller (2010). In view of generalizing to the truly categorical case and of incorporating serial dependence, however, it appears that again the CUSUM approach is more feasible (an EWMA-based categorical approach is discussed by Ye et al. (2002)). A CUSUM chart for an i.i.d. binary process $(X_t)_{\mathbb{N}}$ was first proposed by Reynolds and Stoumbos (1999), and it was extendend to the case of a binary Markov chain as in (7) by Mousavi and Reynolds (2009). Here, the idea is as sketched in Remark 6: the contribution to the log-likelihood ratio by the t^{th} observation equals $L_t = \ln\left(P_{\pi_1}(X_t)/P_{\pi_0}(X_t)\right)$ (i.i.d.

case) or $L_t = \ln\left(P_{\pi_1}(X_t|X_{t-1})/P_{\pi_0}(X_t|X_{t-1})\right)$ (Markov case), respectively, which is then used to compute the t^{th} CUSUM statistic. Again, π_1 refers to the relevant out-of-control parameter value of π, while π_0 represents the in-control value.

4.2 Continuous Monitoring: Categorical Case

At this point, let us return to the truly categorical case, where $(X_t)_{\mathbb{N}}$ has range $\mathcal{S} = \{0, \ldots, m\}$ with an $m > 1$. The true marginal probabilities are denoted again by $\boldsymbol{\pi} := (\pi_0, \ldots, \pi_m)^\top$, with $\boldsymbol{\pi}_0$ representing the corresponding in-control value. For defining a CUSUM monitoring scheme, we also have to consider a relevant out-of-control value, say $\boldsymbol{\pi}_1$. Such a CUSUM scheme, assuming that the underlying process is i.i.d., was proposed by Ryan et al. (2011) (also see the discussion in Remark 6 before). If $L_t = \ln\left(P_{\boldsymbol{\pi}_1}(X_t)/P_{\boldsymbol{\pi}_0}(X_t)\right)$, then the CUSUM statistic at time t is

$$S_t = \max\{0,\ S_{t-1} + L_t\}, \qquad \text{where} \quad S_0 := 0. \tag{11}$$

Note that $P_{\boldsymbol{\pi}}(X_t = i)$ just equals π_i, so we can denote $P_{\boldsymbol{\pi}}(X_t) = \pi_{X_t}$, and hence $L_t = \ln\left(\pi_{1;\,X_t}/\pi_{0;\,X_t}\right)$. An alarm is triggered once S_t violates the upper control limit $h > 0$ for the first time.

In analogy to Mousavi and Reynolds (2009), we can extend this categorical CUSUM approach to any kind of Markov-dependent categorical process by defining

$$L_t = \ln\left(\frac{P_{\boldsymbol{\pi}_1}(X_t|X_{t-1}, \ldots, X_{t-\mathrm{p}})}{P_{\boldsymbol{\pi}_0}(X_t|X_{t-1}, \ldots, X_{t-\mathrm{p}})}\right).$$

For illustration, to keep it simple, we shall focus again on the special case of an underlying DAR(1) process (10), where we denote the dependence parameter by $\rho := \phi_1$ as before. It then follows that

$$L_t = \ln\left(\frac{(1-\rho)\,\pi_{1;\,X_t} + \delta_{X_t,X_{t-1}}\,\rho}{(1-\rho)\,\pi_{0;\,X_t} + \delta_{X_t,X_{t-1}}\,\rho}\right) \quad \text{for } t \geq 2, \qquad L_1 = \ln\left(\frac{\pi_{1;\,X_1}}{\pi_{0;\,X_1}}\right). \tag{12}$$

4.3 Continuous Monitoring: ARL Performance

To investigate the effect of serial dependence on the categorical CUSUM chart, we pick up the four situations discussed by Ryan et al. (2011). The assumed

Table 2 CUSUM chart (11) with i.i.d. design and adjusted design, CUSUM chart (12)

		CUSUM (11)			CUSUM (11)			CUSUM (12)		
Case	ρ	h	ARL_0	ARL_1	h	ARL_0	ARL_1	h	ARL_0	ARL_1
1	0	2.95	280.4	21.9						
	0.25	2.95	116.3	20.9	4.3	278.4	31.1	2.85	304.5	30.6
	0.5	2.95	72.0	20.3	6.1	274.0	43.4	2.5	306.0	41.0
	0.75	2.95	59.6	21.8	9.5	280.6	65.0	1.9	289.4	59.6
2	0	2.8	501.8	36.3						
	0.25	2.8	245.7	37.2	3.85	509.8	52.4	2.55	503.4	45.6
	0.5	2.8	170.8	39.3	5.2	500.2	72.6	2.25	508.4	58.8
	0.75	2.8	155.2	48.3	7.6	500.7	107.8	1.7	514.7	86.0
4	0	3.25	284.6	20.6						
	0.25	3.25	103.9	18.9	4.7	285.7	28.9	3	293.0	27.6
	0.5	3.25	52.9	17.0	6.9	280.8	40.8	2.6	289.1	37.1
	0.75	3.25	35.2	15.6	11.5	284.6	63.1	2.05	298.1	56.9

in-control marginal distributions and the corresponding anticipated out-of-control scenarios are

Case 1: $\boldsymbol{\pi}_0 = (0.65, 0.25, 0.10)^\top$, $\boldsymbol{\pi}_1 = (0.4517, 0.2999, 0.2484)^\top$;
Case 2: $\boldsymbol{\pi}_0 = (0.94, 0.05, 0.01)^\top$, $\boldsymbol{\pi}_1 = (0.8495, 0.0992, 0.0513)^\top$;
Case 3: $\boldsymbol{\pi}_0 = (0.994, 0.005, 0.001)^\top$, $\boldsymbol{\pi}_1 = (0.9848, 0.0099, 0.0053)^\top$;
Case 4: $\boldsymbol{\pi}_0 = (0.65, 0.20, 0.10, 0.05)^\top$, $\boldsymbol{\pi}_1 = (0.3960, 0.3283, 0.1734, 0.1023)^\top$.

The first three cases have three states and show decreasing dispersion ($\nu_\mathrm{G} \approx 0.758, 0.171, 0.018$), while the fourth case has four states ($\nu_\mathrm{G} = 0.7$).

Ryan et al. (2011) assumed the categorical process to be i.i.d. and, hence, applied the CUSUM chart (11) for process monitoring. The corresponding chart designs h for Cases 1, 2 and 4 (Case 3 is discussed separately for reasons explained below) are shown in the first block of Table 2, together with simulated (zero-state) ARL values (100,000 replications). Here, ARL_0 always refers to the in-control marginal distribution $\boldsymbol{\pi}_0$, while ARL_1 refers to the special out-of-control situation $\boldsymbol{\pi}_1$.

If the chart design is done assuming i.i.d. observations, but if serial dependence according to a DAR(1) model with parameter value $\rho > 0$ is present (see the first block of Table 2), then the true in-control performance deviates heavily from the expected one. The values for ARL_0 decrease severely with increasing ρ such that false alarms will be observed much too often. One solution is to retain chart type (11) but with adjusted control limit h, as it is shown in the second block of Table 2. It can be observed that the control limit has to be widened to make the chart sufficiently robust (which, inevitably, goes along with a worse out-of-control performance).

The recommended solution, however, is to use the CUSUM chart (12), which is designed to deal with DAR(1) dependence. Appropriate chart designs are shown in the third block of Table 2. Although the out-of-control performance is still worse than in the i.i.d. case (the price one has to pay for serially dependent data), it is visibly better than for the adjusted i.i.d.-CUSUM (11).

Finally, let us have a look at Case 3. Here, $\boldsymbol{\pi}_0$ shows very little dispersion, most of the probability mass concentrates on the state '0'. Certainly, if serial dependence is present but ignored, the chart's performance is affected, see

ARL_0 if $\rho =$				ARL_1 if $\rho =$			
0	0.25	0.5	0.75	0	0.25	0.5	0.75
500.8	488.5	543.6	839.8	124.2	143.7	184.9	309.4

However, for such an extreme marginal distribution, a monitoring of the process is rather problematic if additional serial dependence is present, since then the process nearly always leads constant sample paths. For instance, if $\rho = 0.75$, then $p_{0|0} \approx 0.9994$ according to (10), so we will hardly ever leave the state '0'. This increasing tendency to constantly observing '0' also explains the non-monotonic behaviour observed for ARL_0 before.

5 Conclusions and Future Research

Two scenarios of monitoring a serially dependent categorical process were discussed: a sample-based approach, where the dependence *within* the samples has to be considered, and a continuous monitoring approach, where the dependence *between* successive observations has to be taken into account for chart design. Concerning the first scenario, a Shewhart chart based on a dispersion measure is plausible in view of quality-related applications, while a likelihood-ratio-based CUSUM approach is feasible in the second scenario. In both cases, simulations are required for chart design and performance evaluation. As already pointed out in Remark 6, the development of a sample-based CUSUM chart for serially dependent categorical processes would be an interesting direction for future research.

Besides this, much more work is required concerning both models and control charts for serially dependent *ordinal* data (Remark 4). In view of Remark 5, the development of control charts being able to deal with both time-dependent categorical and *compositional* data would be a promising topic for future research. It also seems that the *Phase I application* of categorical control charts, in particular, the effect of parameter estimation on the charts' performance (Jensen et al. 2006; Jones-Farmer et al. 2014), has not been investigated yet.

Finally, another traditional SPC topic has been ignored completely until now regarding categorical data: process capability analysis. A popular tool for evaluating the actual process capability are *process capability indices*. If it is possible to define a specification region for the categorical distribution $\boldsymbol{\pi}$ in a reasonable way, then one may pick up the idea of Perakis and Xekalaki (2005) and define an index based on the actual "proportion of conformance". The estimation of such an index from time-dependent categorical in-control data has to be investigated.

Acknowledgement The author thanks the reviewer for useful comments on an earlier draft of this article.

References

Adnaik, S. B., Gadre, M. P., & Rattihalli, R.N. (2015). Single attribute control charts for a Markovian-dependent process. *Communications in Statistics—Theory and Methods, 44*(17), 3723–3737.

Aitchison, J. (1986). *The statistical analysis of compositional data*. New York: Chapman and Hall Ltd.

Bhat, U. N., & Lal, R. (1990). Attribute control charts for Markov dependent production processes. *IIE Transactions, 22*(2), 181–188.

Biswas, A., & Song, P. X.-K. (2009). Discrete-valued ARMA processes. *Statististics & Probability Letters, 79*(17), 1884–1889.

Bourke, P. D. (1991). Detecting a shift in fraction nonconforming using run-length control charts with 100% inspection. *Journal of Quality Technology, 23*(3), 225–238.

Boyles, R. A. (1997). Using the chi-square statistic to monitor compositional process data. *Journal of Applied Statistics, 24*(5), 589–602.

Bühlmann, P., & Wyner, A. J. (1999). Variable length Markov chains. *Annals of Statistics, 27*(2), 480–513.

Chen, L., Chang, F. M., & Chen, Y. (2011). The application of multinomial control charts for inspection error. *International Journal of Industrial Engineering, 18*(5), 244–253.

Cozzucoli, P. (2009). Process monitoring with multivariate *p*-control charts. *International Journal Quality, Statistics and Reliability, 2009*, 11

Deligonul, Z. S., & Mergen, A. E. (1987). Dependence bias in conventional *p*-charts and its correction with an approximate lot quality distribution. *Journal of Applied Statistics, 14*(1), 75–81.

Duncan, A. J. (1950). A chi-square chart for controlling a set of percentages. *Industrial Quality Control, 7*, 11–15.

Duran, R. I., & Albin, S. L. (2009). Monitoring and accurately interpreting service processes with transactions that are classified in multiple categories. *IIE Transactions, 42*(2), 136–145.

Feller, W. (1968). *An introduction to probability theory and its applications – volume I* (3rd ed.). New York: John Wiley & Sons, Inc.

Gan, F. F. (1990). Monitoring observations generated from a binomial distribution using modified exponentially weighted moving average control chart. *Journal of Statistical Computation and Simulation, 37*(1–2), 45–60.

Gan, F. F. (1993). An optimal design of CUSUM control charts for binomial counts. *Journal of Applied Statistics, 20*(4), 445–460.

Höhle, M. (2010). Online change-point detection in categorical time series. In T. Kneib & G. Tutz (Eds.), *Statistical modelling and regression structures* (pp. 377–397). Heidelberg: Physica Verlag.

Holan, S. H., Lund, R., & Davis, G. (2010). The ARMA alphabet soup: A tour of ARMA model variants. *Statistics Surveys, 4*, 232–274.

Horn, S. D. (1977). Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics, 33*(1), 237–248.

Jacobs, P. A., & Lewis, P. A. W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis, 4*(1), 19–36.

Jensen, W. A., Jones-Farmer, L. A., Champ, C. W., & Woodall, W.H. (2006). Effects of parameter estimation on control chart properties: A literature review. *Journal of Quality Technology, 32*(4), 395–409.

Jones, L. A., Woodall, W. H., & Conerly, M. D. (1999). Exact properties of demerit control charts. *Journal of Quality Technology, 31*(2), 207–216.

Jones-Farmer, L. A., Woodall, W. H., Steiner, S. H., & Champ, C. W. (2014). An overview of phase I analysis for process improvement and monitoring. *Journal of Quality Technology, 46*(3), 265–280.

Li, J., Tsung, F., & Zou, C. (2012). Directional control schemes for multivariate categorical processes. *Journal of Quality Technology, 44*(2), 136–154.

Maiti, R., & Biswas, A. (2018). Time series analysis of categorical data using auto-odds ratio function. *Statistics, 52*(2), 426–444.

Marcucci, M. (1985). Monitoring multinomial processes. *Journal of Quality Technology, 17*(2), 86–91.

Montgomery, D. C. (2009). *Introduction to statistical quality control* (6th ed.). New York: John Wiley & Sons, Inc.

Mousavi, S. & Reynolds, M. R. Jr. (2009). A CUSUM chart for monitoring a proportion with autocorrelated binary observations. *Journal of Quality Technology, 41*(4), 401–414.

Mukhopadhyay, A. R. (2008). Multivariate attribute control chart using Mahalanobis D^2 statistic. *Journal of Applied Statistics, 35*(4), 421–429.

Nelson, L. S. (1987). A chi-square control chart for several proportions. *Journal of Quality Technology, 19*(4), 229–231.

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

Pawlowsky-Glahn, V., & Buccianti, A. (Eds.). (2011). *Compositional data analysis — theory and practice*. Chichester: John Wiley & Sons, Ltd.

Perakis, M., & Xekalaki, E. (2005). A process capability index for discrete processes. *Journal of Statistical Computation and Simulation, 75*(3), 175–187.

Raftery, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society B, 47*(3), 528–539.

Reynolds, M. R. Jr., & Stoumbos, Z. G. (1999). A CUSUM chart for monitoring a proportion when inspecting continuously. *Journal of Quality Technology, 31*(1), 87–108.

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

Ryan, A. G., Wells, L. J., & Woodall, W. H. (2011). Methods for monitoring multiple proportions when inspecting continuously. *Journal of Quality Technology, 43*(3), 237–248.

Steiner, S. H., Geyer, P. L., & Wesolowsky, G. O. (1996). Grouped data-sequential probability ratio tests and cumulative sum control charts. *Technometrics, 38*(3), 230–237.

Topalidou, E., & Psarakis, S. (2009). Review of multinomial and multiattribute quality control charts. *Quality and Reliability Engineering International, 25*(7), 773–804.

Tucker, G. R., Woodall, W. H., & Tsui, K.-L. (2002). A control chart method for ordinal data. *American Journal of Mathematical and Management Sciences, 22*(1–2), 31–48.

Vives-Mestres, M., Daunis-i-Estadella, J., Martín-Fernández, J.A. (2014a). Out-of-Control signals in three-part compositional T^2 control chart. *Quality and Reliability Engineering International, 30*(3), 337–346.

Vives-Mestres, M., Daunis-i-Estadella, J., & Martín-Fernández, J. A. (2014b). Individual T^2 control chart for compositional data. *Journal of Quality Technology, 46*(2), 127–139.

Wang, Y. H., & Yang, Z. (1995). On a Markov multinomial distribution. *Mathematical Scientist, 20*, 40–49.

Weiß, C. H. (2008). Visual analysis of categorical time series. *Statistical Methodology, 5*(1), 56–71.

Weiß, C. H. (2009). Group inspection of dependent binary processes. *Quality and Reliability Engineering International, 25*(2), 151–165.

Weiß, C. H. (2011). Empirical measures of signed serial dependence in categorical time series. *Journal of Statistical Computation and Simulation, 81*(4), 411–429.

Weiß, C. H. (2012). Continuously monitoring categorical processes. *Quality Technology and Quantitative Management, 9*(2), 171–188.

Weiß, C. H. (2013a). Serial dependence of NDARMA processes. *Computational Statistics & Data Analysis, 68*, 213–238.

Weiß, C. H. (2013b). Monitoring *k*-th order runs in binary processes. *Computational Statistics, 28*(2), 541–563.

Weiß, C. H., & Atzmüller, M. (2010). EWMA control charts for monitoring binary processes with applications to medical diagnosis data. *Quality and Reliability Engineering International, 26*(8), 795–805.

Weiß, C. H., & Göb, R. (2008). Measuring serial dependence in categorical time series. *Advances in Statistical Analysis, 92*(1), 71–89.

Woodall, W. H. (1997). Control charts based on attribute data: Bibliography and review. *Journal of Quality Technology, 29*(2), 172–183.

Woodall, W. H. (2000). Controversies and contradictions in statistical process control. *Journal of Quality Technology, 32*(4), 341–350.

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

Xie, M., Goh, N., & Kuralmani, V. (2000). On optimal setting of control limits for geometric chart. *International Journal of Reliability, Quality and Safety Engineering, 7*(1), 17–25.

Yashchin, E. (2012). On detection of changes in categorical data. *Quality Technology & Quantitative Management, 9*(1), 79–96.

Ye, N., Masum, S., Chen, Q., & Vilbert, S. (2002). Multivariate statistical analysis of audit trails for host-based intrusion detection. *IEEE Transactions on Computers, 51*(7), 810–820.

Yeh, A. B., McGrath, R. N., Sembower, M. A., & Shen, Q. (2008). EWMA control charts for monitoring high-yield processes based on non-transformed observations. *International Journal of Production Research, 46*(20), 5679–5699.

Monitoring of Short Series of Dependent Observations Using a XWAM Control Chart

Olgierd Hryniewicz and Katarzyna Kaczmarek-Majer

Abstract Many different control charts have been proposed during the last 30 years for monitoring processes with autocorrelated observations (measurements). The majority of them are developed for monitoring residuals, i.e., differences between the observed and predicted values of the monitored process. Unfortunately, statistical properties of these chart are very sensitive to the accuracy of the estimated model of the underlying process. In this chapter we consider the case when the information from the available data is not sufficient for good estimation of the model. Therefore, we use the concept of model weighted averaging in order to improve model prediction. The novelty of the proposed XWAM control chart consists in the usage of computational intelligence methodology for the construction of alternative models, and the calculation of their weights.

Keywords Control chart · Residuals · Autocorrelated data · Short time series · XWAM control chart

1 Introduction

Control charts were originally devised for monitoring production processes when long series of quality-related measurements are observed. Later on, they have also been successfully applied in cases of short production runs. Problem arise, however, when consecutive observations are statistically dependent. First pioneering works in the area of process control in presence of dependent (autocorrelated) data, such as, e.g., Box et al. (1974), were published in the 1970s. Since that time many chapters devoted to this problem have been published, and they can be, in general, divided into two groups. Authors of the first group of chapters, such as, e.g., Vasilopoulos and Stamboulis (1978), Montgomery and Mastrangelo (1991),

O. Hryniewicz (✉) · K. Kaczmarek-Majer
Systems Research Institute, Polish Academy of Sciences, Warszawa, Poland
e-mail: hryniewi@ibspan.waw.pl; K.Kaczmarek@ibspan.waw.pl

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_13

Maragah and Woodall (1992), Yashchin (1993), Schmid (1995) or Zhang (1998), propose to adjust design parameters of classical control charts (Shewhart, CUSUM, EWMA) in order to accommodate the impact of autocorrelation in data on chart's statistical properties. The origin of the second group of chapters is the chapter by Alwan and Roberts (1988) who proposed a control chart for residuals. In their approach a mathematical model of the observed process has to be identified using the methodology developed for the analysis of time series. The deterministic part of this model is used for the computation of predicted values of observations, and differences between predicted and observed values of the process, named residuals, are plotted on a control chart. Properties of different control charts for residuals have been investigated by many authors, such as, e.g., Wardell et al. (1994), Zhang (1997), Kramer and Schmid (2000). Both approaches have been compared in many chapters, such as, e.g., Lu and Reynolds (1999). It has to be noted, however, that the applicability of the charts for residuals in SPC was a matter of discussion (see, e.g., the chapter by Runger (2002)), but now this approach seems to be prevailing. Recently, more complicated procedures have been proposed. For example, the ARMA chart proposed by Jiang et al. (2000), the chart proposed by Chin and Apley (2006) based on second-order linear filters, the chart proposed by Apley and Chin (2007) based on general linear filters or the PCA-based procedure for the monitoring multidimensional processes proposed by De Ketelaere et al. (2015).

A proper design of a control chart for autocorrelated data requires the knowledge of the mathematical model of the monitored process. When series of observations (production runs) are long enough to determine an appropriate model of dependence several solutions have already been proposed for the calculation of such characteristics like the ARL. Even in this case, however, serious problems arise when we want to calculate chart's characteristics when the monitored process goes out of control. The situation is even worse when the amount of available data is not sufficient for the identification of the underlying model of dependence. In such a case only few analytical results exist (see, e.g., the chapter by Kramer and Schmid (2000) or the chapter by Apley and Lee (2008)). These difficulties stem mainly from the fact that for imprecisely (or wrongly) identified model of dependence not only observations, but residuals as well, are autocorrelated. Unfortunately, this happens in practice when, e.g., the monitored process is in its prototype phase or when we monitor patients in a health-care system. The latter example gives motivation for the research described in this chapter.

It seems to be rather unquestionable that proper identification of the dependence model is equivalent to finding a good predictor for future observations. When we do not have enough data for building a good model, i.e., when the available time series is too short, one can use methods developed by econometricians for prediction purposes in short economic time series. In such situations they prefer to use Bayesian methods combined with the Markov Chain Monte Carlo simulation methodology. A very good description of this approach can be found in the book by Geweke (2005). What is specific in this approach is the concept of model averaging. The Bayesian model in this approach contains not only prior knowledge about model parameters, but also prior knowledge about several possible models that can be used for prediction. In practice, non-informative priors are

used, and MCMC simulations are used for the evaluation of predictive posterior distributions. Hryniewicz and Katarzyna Kaczmarek-Majer (2016a) proposed to use some computational intelligence methods for the construction of the prior distribution on the pre-chosen set of models. Their algorithm appears to be highly competitive when compared to the best available algorithms used for the prediction in short time series. In this chapter we try to adopt a similar approach for the construction of Shewhart control charts for residuals.

This chapter is an extended and re-worked version of the chapter Hryniewicz and Katarzyna Kaczmarek-Majer (2016b) published in the proceedings of the international conference ISQC 2016 held in Hamburg. In particular, it contains results of new and extensive computer simulations that allow to evaluate properties of the proposed new control chart in more realistic, from a practical point of view, setting. The chapter is organized as follows. In the next section we describe the assumed mathematical model of the monitored process, and present the algorithm for the construction of the proposed XWAM chart. Section 3 is devoted to the description of methods that have been used for building alternative models of the monitored process. Simulation methods have been used for the evaluation of statistical properties of the proposed control chart. Comprehensive experiments have been performed, but due to the limited volume of this chapter only some representative results have been described in Sect. 4. The chapter is concluded in the last section where we also outline possible areas of future investigations.

2 Mathematical Model and the Design of an XWAM Control Chart

2.1 Introductory Remarks

Control charts perform well when they are designed using sufficient amount of data. In the case of classical control charts the amount of statistical data is sufficient for design purposes if it allows to estimate process parameters with good precision. The situation is much more difficult in the case of control charts for residuals. In this case the data is used for the estimation of the underlying model of the process, and the parameters of the probability distribution of residuals. In this section we propose an alternative design of the X chart for residuals that can be used when available samples are small.

2.2 Mathematical Model

Consider random observations described by a series of random variables $X_1, X_2, \ldots$. In the context of statistical quality control these random variables may describe individual observations or observed values of sample statistics, such as, e.g., averages plotted on a Shewhart $\bar{X}$-chart. The full mathematical description of such a

series can be done using a multivariate (possibly infinitely-dimensional) probability distribution. Unfortunately, in practice this usually cannot be done. Therefore, statisticians introduced simpler and easier tractable mathematical models, based on the notion of conditionality. In the most popular model of this kind the random variable representing the current observation is given as the sum of a deterministic part depending on the observed values of previous observations, and a random part whose probability distribution does not depend upon the previously observed values, i.e.,

$$X_i = f(x_1, \ldots . x_{i-1}) + \epsilon_i, i = 1, \ldots . \tag{1}$$

In the simplest version of (1) we usually assume that random variables $\epsilon_i, i = 1, \ldots$ are mutually independent and identically distributed. On the other hand, we often assume that the deterministic part $f(x_1, \ldots, x_{i-1})$ has a form that assures stationarity of the time series $X_1, X_2, \ldots$. In this chapter we make even stronger assumption that

$$X_i = a_1 x_{i-1} + \ldots + a_p x_{i-p} + \epsilon_i, \tag{2}$$

where $\epsilon_i, i = 1, \ldots$ are normally distributed independent random variables with the expected value equal to zero, and the same finite standard deviation. Thus, our assumed model describes a classical autoregressive stochastic process of the pth order $AR(p)$. The comprehensive description of the $AR(p)$ process can be found in every textbook devoted to the analysis of time series, e.g., in the seminal book by Box et al. (2008) or a popular textbook by Brockwell and Davis (2002). In these books one can find the description of more general models, such as, e.g., the $ARMA(p, q)$ which are also special cases of (1), and are widely used in the statistical analysis of time series.

Estimation of the model $AR(p)$, given by (2), is relatively simple when we know the order p of the considered model. In order to do this we have to calculate first p sample autocorrelations $r_1, r_2, \ldots, r_p$, defined as

$$r_i = \frac{n \sum_{t=1}^{n-i} (x_t - \hat{\mu})(x_{t+i} - \hat{\mu})}{(n-i) \sum_{t=1}^{n} (x_t - \hat{\mu})^2}, i = 1, \ldots, p, \tag{3}$$

where n is the number of observations (usually, it is assumed that $n \geq 4p$), and $\hat{\mu}$ is their average. Then, the parameters $a_1, \ldots, a_p$ of the $AR(p)$ model are calculated by solving the Yule-Walker equations (see, Brockwell and Davis 2002)

$$\begin{aligned} r_1 &= a_1 + a_2 r_1 + \ldots + a_p r_{p-1} \\ r_2 &= a_1 r_1 + a_2 + \ldots + a_p r_{p-2} \\ &\ldots \\ r_p &= a_1 r_{p-1} + a_2 r_{p-2} + \ldots + a_p \end{aligned} \tag{4}$$

In practice, however, we do not know the order of the autoregression process, so we need to estimate p from data. In order to do this let us first define a random variable,

called the residual.

$$Z_i = X_i - (a_1 x_{i-1} + \ldots + a_p x_{i-p}), i = p+1, \ldots, N. \tag{5}$$

The probability distribution of residuals is the same as the distribution of random variables $\epsilon_i, i = 1, \ldots$ in (2), and its variance can be used as a measure of the accuracy of predictions. For given sample data of size n the variance of residuals is decreasing with the increasing values of p. However, the estimates of p models parameters $a_1, \ldots, a_p$ become less precise, and thus the overall precision of prediction with future data deteriorates. As the remedy to this effect several optimization criteria with a penalty factor which discourages the fitting of models with too many parameters have been proposed. In this research we use the *BIC* criterion proposed by Akaike (1978), and defined as

$$BIC = (n-p)\ln[n\hat{\sigma}^2/(n-p)] + n(1+\ln\sqrt{2\pi}) + p\ln[(\sum_{t=1}^{n} x_t^2 - n\hat{\sigma}^2)/p], \tag{6}$$

where x_t are process observations transformed in such a way that their expected values are equal to zero, and $\hat{\sigma}^2$ is the observed variance of residuals. The fitted model, i.e., the estimated order p and parameters of the model $\hat{a}_1, \ldots, \hat{a}_p$ minimizes the value of *BIC* calculated according to (6).

It is a well known fact that the accuracy of prediction in time series strongly depends upon the number of available observations. In Sect. 4 we will present some numerical illustration of this effect. The problem begins, however, when the number of available observations is strongly limited. In the context of SPC this means that we have, e.g., to design a control chart for a short production run. In such a case the accuracy of the estimated model of a monitored process may be completely insufficient if we follow recommendations applicable in the case of a control chart for independent observations.

The problem mentioned above arises in many areas when only short time series are available, such as, e.g., in the case of economic data. In order to overcome this econometricians proposed an empirical (objective) Bayesian approach to the analysis of time series. One of the most important aspects of this approach is the averaging of models. According to Geweke (2005) we define a set $M = \{M_1, M_2, \ldots, M_J\}$ of multiple alternative probabilistic models of a considered process. Then, the posterior density of a vector of interest ω (e.g., some consecutive predicted values of a process) is defined as follows (Geweke 2005)

$$p(\omega|y, M) = \sum_{j=1}^{J} p(M_j|y, M)p(\omega|y, M_j), \tag{7}$$

where y is a series of observations, $p(\omega|y, M_j)$ is the posterior density of the vector of interest conditional on model M_j, and $p(M_j|y, M)$ are the prior model probability

distributions. In this chapter we will use the concept of model averaging for the construction of a control chart. Different $AR(p)$ models will be used as alternative probabilistic models of a monitored process, and their prior probabilities (weights) will be computed using, for example, a methodology described in Sect. 3 of this chapter.

2.3 Design of the XWAM Control Chart

SPC for processes with autocorrelated data using a control chart for residuals was firstly proposed by Alwan and Roberts (1988). Their methodology is applicable for any class of processes, so it is also applicable for the $AR(p)$ process considered in this chapter. According to the methodology proposed by Alwan and Roberts (1988) the deterministic part of (1) is estimated from sample data, and then used for the calculation of residuals. This methodology is also known under the name "filtering". In our case it is the deterministic part of the $AR(p)$ process estimated according to the methodology described in Sect. 2.2 from a sample of n elements. We denote this estimated model as M_0, and its parameters by a vector $(a_{1,0}, \ldots .a_{p_0,0})$. We assign to this estimated model a certain weight $w_0 \in [0, 1]$. We also consider k alternative models $M_j, j = 1, \ldots, k$, each described by a vector of parameters $(a^0_{1,j}, \ldots, a^0_{p_j,j})$. In general, any model with known parameters can be used as an alternative one, but in this chapter we restrict ourselves to the models chosen according to the algorithm described in Sect. 3. Let $w_1^{'}, \ldots, w_k^{'}$ denote the weights assigned to models $M_1, \ldots, M_k$ by the algorithm described in Sect. 3 when only alternative models are considered. Because the total weight of the chosen alternative models is $1 - w_0$, in the construction of our control chart, coined XWAM (X Weighted Average Model chart), to the estimated model we assign the weight w_0, and to each chosen alternative model we will assign a weight $w_j = (1 - w_0)w_j^{'}, j = 1, \ldots, k$.

When we model our process using $k + 1$ models (one estimated from data and k alternative) each process observation generates $k + 1$ residuals. In the case considered in this chapter they are calculated using the following formula

$$z_{i,j} = x_i - (a_{1,j}x_{i-1} + \ldots + a_{p_j,j}x_{i-p_j}), j = 0, \ldots, k; i = p_j + 1, \ldots. \tag{8}$$

In (8) we have assumed that for a model with $p_j, j = 0, \ldots, k$ parameters we need exactly p_j previous consecutive observations in order to calculate first residual. Therefore, we need $i_{min} = \max(p_0, \ldots, p_k) + 1$ observations for the calculation of all residuals in the sample. For the calculation of the parameters of the XWAM control chart we use $n - i_{min} + 1$ weighted residuals calculated from the formula

$$z_i^{\star} = \sum_{j=0}^{k} w_j z_{i,j}, i = i_{min}, \ldots, n. \tag{9}$$

Note, that the number of observed weighted residuals in the sample is smaller than the sample size. Thus, the effective sample size used for the design of a control chart is smaller than the number of available observations from the process. The central line of the chart is calculated as the mean of $z_i^\star$, and the control limits are equal to the mean plus/minus three standard deviations of $z_i^\star$, respectively.

The operation of the XWAM control chart is a classical one. First decision is made after i_{min} observations. The weighted residual for the considered observation is calculated according to (9), and compared to the control limits. An alarm is generated when the weighted residual falls beyond the control limits.

3 Similarity Measures of Series of Observations

3.1 Introductory Remarks

Finding one appropriate probabilistic model, and estimating its parameters, may become a very challenging task for short series of observations. In this section, we explain the proposed approach of selecting k alternative models that describe the monitored process. The selection is determined by distances learned between the monitored process and the training series from a template database. The training database consists of sample realizations of template predictive models. Within the proposed approach, distances between the monitored process and the training series are evaluated, and as a result of their aggregation, prior model probabilities (weights) are established for the chosen k alternative models. This combination is inspired by Bayesian averaging, as extensively described by Geweke (2005).

3.2 Similarity Measures of Series of Observations

The similarity of two time series is evaluated by calculating the distance between them. Within the proposed approach, the Dynamic Time Warping (DTW) algorithm for measuring the distance between two series as introduced by Berndt and Clifford (1994) is adapted. DTW is the classical elastic measure that enables to calculate the smallest distance between two series of observations independently of certain non-linear variations in the time dimension. Therefore, DTW calculates the best match between two given series allowing similar shapes to match even if they are out of phase in time axis. For the recent survey and the experimental comparison of various similarity measures for time series data, see e.g., Wang et al. (2013). Wang et al. (2013) conclude that especially on small data sets elastic measures like DTW can be significantly more accurate than the Euclidean distance or other lock-step measures because the elastic (non-linear) measures take into account the dilatation in time.

Let $X = \{x_1, x_2, \ldots, x_N\}$ and $Z = \{z_1, z_2, \ldots, z_M\}$ denote time series to be compared. To find the best alignment between X and Z, we first construct a N-by-M (cost) matrix where (ith, jth) element of it corresponds to the local cost function. The local cost function $d(i, j)$ is the distance between two points x_i and z_j of compared time series

$$d(i, j) = f(x_i, z_j) \geq 0 \tag{10}$$

The magnitude of the difference $d(i, j) = |x_i - z_j|$ (Manhattan) or square of the difference $d(i, j) = (x_i - z_j)^2$ (Euclidean) are some of the most common local cost functions considered in applications. In the experiments of this research, the Euclidean local cost distance is used. Then, to find the best match between two given series X and Z, we retrieve a path (the so called warping path) through the cost matrix that minimizes the cumulative distance. The following recursive relation defines the cumulative distance $g(i, j)$ for $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, M\}$

$$g(i, j) = d(i, j) + min[g(i - 1, j), g(i - 1, j - 1), g(i, j - 1)] \tag{11}$$

The cumulative distance $g(i, j)$ is the sum of the distance between current elements and the minimum of the cumulative distances of the neighboring points. Two points (x_i, z_j) and (x_{i*}, z_{j*}) on the N-by-M cost matrix are called neighboring if

$$(|i - i*| = 1 \; and \; |j - j*| = 0) \; or \; (|i - i*| = 0 \; and \; |j - j*| = 1) \tag{12}$$

The warping path is found using the dynamic programming and the algorithm's complexity is $O(NM)$. When the X and Z series are of the same length, then the value of $g(N, M)$ defines the DTW distance between them.

In Fig. 1, the performance of the Euclidean and DTW distances are compared for three exemplary short series of observations (five observations each) generated from three different autoregressive processes, namely $AR(-0.9)$, $AR(-0.5)$, and $AR(0.0)$.

In general, time series generated from white noise AR(0.0) should be more similar to time series generated from AR(−0.5) process than to time series generated from AR(−0.9) process because of their autoregressive characteristics. However, as observed, the Euclidean distance between series from $AR(0.0)$ and $AR(-0.9)$ amounts to 3.7, and between series from $AR(0.0)$ and $AR(-0.5)$ it results to 4.1, which is contradictory to intuition. At the same time, the DTW distance between series from $AR(0.0)$ and $AR(-0.9)$ amounts to 3.7, whereas the distance between series from $AR(0.0)$ and $AR(-0.5)$ is smaller and amounts to 3.5. It this context, the DTW similarity measure provides appropriate results that are in line with intuition. Time series with stronger autocorrelations have similar patterns even if they are dilated in time.

Further numerical experiments will be presented in Sect. 4. They confirm the good properties of the DTW measure, especially for time series with identified dilatation in time.

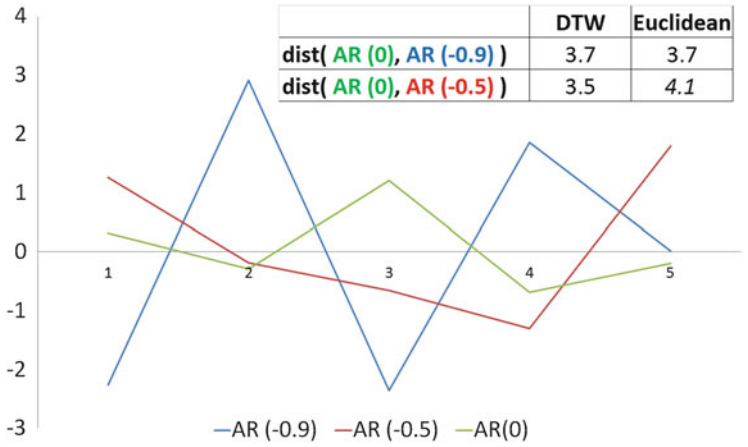


Fig. 1 Euclidean and DTW distances for exemplary series of observations

3.3 Construction of Prior Probabilities (Weights)

Having defined the distance between two time series, the proposed method of selecting k alternative models is explained. The input for the algorithm is the monitored process y, the desired number of alternative models k and definitions of the AR processes to be considered in the template database. We adapt stationary AR processes of different orders as template models M. It needs to be stated, that in numerical experiments the order is usually assumed less or equal 2.

The output of the algorithm is in form of definitions of alternative models $\{M_1 \ldots, M_k\}$ to be considered in predictions of the monitored process and their respective weights $\{w_1, \ldots, w_k\}$ such that $\sum_{h=1}^{k} w_h = 1$.

Algorithm 1 depicts a high-level description of the proposed approach. It consists of the following steps:

Step 1. Generation of the template database $Y_{J,s}$.
The template database consists of models $\{M_1, \ldots, M_J\}$ that are stationary AR processes of order less or equal p. For each of the J models (processes) its s realizations (training time series) are generated and considered for similarity calculations. For the clarity reasons, the length of generated series is the same as length of the considered monitored process.

Step 2. Calculating distances between the monitored process y and the training time series from the template database using the DTW distance.
For $m \in J$ and their realizations $i \in s$, the distance between the training time series and the considered monitored series of observations is calculated

$$dist_{m,i} = DTW(y_{m,i}, y) \tag{13}$$

Algorithm 1 Building alternative models of the monitored process (BAM)

▷ **Input**:
y - monitored process,
p - max order of the AR process considered to build template database,
s - number of sample time series from each of the template AR processes,
α - min difference between autoregressive coefficients of AR models in template database,
k - number of alternative models to be considered
▷ **Output**:
$M_1, \ldots, M_k$ - alternative models to be considered for the monitored process,
$w_1, \ldots, w_k$ - weights for the alternative models
procedure BAM(y, p, s, α, k)
    $l \leftarrow$ length(y)
    $J \leftarrow 0$
    for $order = 0$ **to** p **do** ▷ **Step 1**. Generation of template database
        for $\theta = -1 + \alpha$ **to** 1 add α **do**
            if generateAR(length=l, order, θ) is stationary **then**
                for $i = 1$ **to** s **do**
                    $Y_{i,order,\theta} \leftarrow$ generateAR(length=l, order, θ) ▷ θ is a list of autoregressive parameters for AR order greater or equal 2
                $J \leftarrow J + 1$
    for $m = 1$ **to** J **do** ▷ **Step 2**. Calculating similarity of monitored process y to time series from the template database
        for $i = 1$ **to** s **do**
            $dist_{m,i} \leftarrow$ distanceDTW(y, $y_{m,i}$)
        $dist_m \leftarrow$ meanDistance(M_m)
    for $m = 1$ **to** J **do** ▷ **Step 3**. Aggregating similarities to establish weights
        $M_1, \ldots, M_k \leftarrow$ selectAlternativeModels($dist_m$, k)
    $w_1, \ldots, w_k \leftarrow$ scaleWeights(M, k)
    return $M_1, \ldots, M_k, w_1, \ldots, w_k$

Step 3. Aggregating similarities to establish weights corresponding to models $\{M_1 \ldots, M_k\}$.

The mean aggregation operator is considered to construct weights for each model based on distances retrieved for each of the s sample time series. For model M_m where $m \in J$ having s realizations, the average distance between the training time series and the considered monitored series of observations is calculated as follows

$$dist_m = \frac{\sum_{i=1}^{s} dist_{m,i}}{s} \tag{14}$$

Having evaluated the average distance for each of the template models $\{M_1, \ldots, M_J\}$, the k models with smallest distance are selected. Then, the prior weights $\{w_1, \ldots, w_k\}$ are calculated

$$w_i = \frac{dist_i}{\sum_{h=1}^{k} dist_h}. \tag{15}$$

4 Numerical Experiments

4.1 Properties of X Charts and X Charts for Residuals

Let consider Shewhart X control charts, both for raw data and for residuals, whose parameters are designed using information from relatively small samples. In the case of raw data and independent observations the basic characteristic of a chart with estimated control limits, the average run length (ARL), can be computed using the approach proposed by Chakraborti (2000). Previously, the effect of parameter estimation on properties of the classical Shewhart control chart has been investigated by many authors using mainly Monte Carlo simulations. In the case of independent observations they found that estimated control limits, in general, are too wide. Thus, the values of the ARL are larger than expected, and special corrections are needed, such as, e.g., proposed by Albers and Kallenberg (2004). The same effect has been observed in the case of autocorrelated data. When we use a Shewhart control chart for residuals, and we have enough data to estimate the underlying model of the process, and the variance of residuals, sufficiently precisely, then the chart for residuals behaves like a classical Shewhart control chart. However, when we do not have enough data, and this is a usual case in practice, the value of ARL of the chart for residuals is, as it was proved by Kramer and Schmid (2000), smaller than in the case of the classical Shewhart control chart applied for original (raw) observations. In order to illustrate these well known features we have performed a simulation experiment in which $N = 50{,}000$ (200,000 in the case of independent observations) charts were designed, and for each of them $N_R = 5000$ process runs of maximum $M_R = 500{,}000$ observations (curtailment value) were simulated. We have performed this experiment for the ordinary Shewhart X chart for individual observations, and for the Shewhart X-chart for residuals. The charts of both types have been designed using the information coming from the simulated sample of n items. Note, that in the case of a control chart for residuals the underlying model was estimated using a methodology described in Sect. 2. For each of the considered charts we have calculated the average run length (ARL), and the median run length (MRL). Then, in order to compare both types of charts we have computed the following characteristics of the respective distributions: average of the distribution of ARL's (AvgARL), standard deviation of the distribution of ARL's (StdARL), median the distribution of ARL's (MedARL), skewness of the distribution of ARL's (SkewARL), average of the distribution of MRL's (AvgMRL), standard deviation of the distribution of MRL's (StdMRL), median of the distribution of MRL's (MedMRL), and skewness of the distribution of MRL's (SkewMRL). In the case of independent consecutive observations (both in the sample used for design purposes, and the monitored process) the characteristics of the distribution of ARL's are presented in Table 1, and the characteristics of the distribution of MRL's are presented in Table 2.

The results of simulations presented in Tables 1 and 2 confirm many of well known facts. First, consider the case of the X chart for direct, and independent,

Table 1 Characteristics of the ARL distributions for X charts, and X charts for residuals—independent observations

	X-chart				X-chart (residuals)			
n	AvgARL	StdARL	MedARL	SkewARL	AvgARL	StdARL	MedARL	SkewARL
20	1554.7	9228.1	256.9	23.0	485.6	5106.4	66.5	47.5
30	863.7	3326.9	287.6	37.2	369.9	1924.4	114.4	66.2
40	674.0	1730.8	306.3	30.5	342.9	801.6	150.2	14.4
50	578.1	988.5	315.4	10.9	341.2	637.1	180.0	13.8
100	455.9	401.2	342.3	4.1	345.8	315.1	258.1	5.7
200	408.6	225.0	355.6	2.1	356.2	194.6	309.5	2.0
500	385.0	125.3	364.9	1.2	365.9	118.8	346.4	1.2
1000	377.2	85.1	367.0	0.8	369.4	83.1	359.2	0.8
2000	374.2	59.1	369.2	0.5	371.4	58.3	366.6	0.5

Table 2 Characteristics of the MRL distributions for X charts, and X charts for residuals—independent observations

	X-chart				X-chart (residuals)			
n	AvgMRL	StdMRL	MedMRL	SkewMRL	AvgMRL	StdMRL	MedMRL	SkewMRL
20	1134.3	8650.3	178.0	35.6	353.9	4792.8	47.0	70.1
30	601.2	2571.7	200.0	63.9	258.9	1665.9	80.0	154.3
40	467.5	1217.0	212.5	33.7	238.8	557.4	105.0	14.5
50	401.0	685.4	219.0	10.8	237.6	441.5	126.0	13.8
100	316.2	278.2	237.0	4.1	240.7	218.4	180.0	5.7
200	283.4	156.1	246.5	2.2	248.0	135.0	216.0	2.0
500	267.0	87.0	253.0	1.2	254.8	82.5	241.0	1.2
1000	261.6	59.1	254.5	0.8	257.2	57.7	250.0	0.8
2000	259.5	41.2	256.0	0.5	258.4	40.6	255.0	0.5

observations (columns 2–5). The distribution of ARL's (over a set of possible control charts) for small samples is in this case extremely positively skewed. Averaging of ARL's and MRL's yields for small samples strongly positively biased estimators of the theoretical values of these characteristics (370.4 and 256.4, respectively). On the other hand, medians of ARL's and MRL's are negatively biased, but this bias seems to be visibly smaller. In both cases the bias results from imprecise estimation of control limits. When we consider the X chart for residuals (columns 6–9) the situation is different. In this case the uncertainty related to imprecisely calculated control limits (positive bias) is combined with the uncertainty related to the computation of residuals (negative, as it was proved in Kramer and Schmid (2000)). Paradoxically, a false assumption of dependence that leads to the usage of a control chart for residuals, for small and medium sample sizes leads to better characteristics of this chart in comparison to the X chart for individual observations designed under the assumption of independent observations. Only for large sample sizes control charts designed using both approaches have similar characteristics, as it is expected in theory. Note, that in the case of a control chart for residuals

Table 3 Properties of the X chart for residuals with dependent observations—negative autocorrelation

	$\rho = -0.9$				$\rho = -0.5$			
n	AvgARL	MedARL	MedMRL	SkewARL	AvgARL	MedARL	MedMRL	SkewARL
20	2212.6	165.7	115.0	19.3	763.5	84.4	59.5	34.4
30	928.1	203.0	141.0	42.7	503.5	132.3	92.0	57.1
40	589.6	225.3	157.0	91.6	419.8	171.7	120.0	30.3
50	524.4	244.4	204.0	97.5	388.5	199.7	139.0	18.5
100	395.1	293.8	204.0	3.9	368.7	273.8	190.0	3.9
200	377.5	327.3	227.0	2.3	369.6	320.7	223.0	2.1
500	371.4	351.0	244.0	1.1	370.8	351.2	244.0	1.2
1000	370.7	360.6	250.0	0.8	370.9	360.8	251.0	0.8
2000	371.0	366.7	254.0	0.5	371.1	366.2	254.0	0.5

Table 4 Properties of the X chart for residuals with dependent observations—positive autocorrelation

	$\rho = 0.9$				$\rho = 0.5$			
n	AvgARL	MedARL	MedMRL	SkewARL	AvgARL	MedARL	MedMRL	SkewARL
20	1304.8	70.5	49.0	24.7	629.3	67.7	47.3	40.5
30	654.3	116.0	81.0	48.3	457.3	119.1	83.0	70.5
40	461.7	155.3	108.0	83.8	398.2	160.4	112.0	31.8
50	390.9	184.4	128.0	183.4	374.8	191.2	133.0	19.3
100	361.4	266.3	185.0	4.3	364.9	270.5	188.0	3.9
200	365.4	318.2	221.0	2.2	368.4	319.6	222.0	2.1
500	369.2	349.7	243.0	1.2	370.6	350.5	243.5	1.2
1000	370.5	360.0	250.0	0.8	370.9	360.7	250.5	0.8
2000	371.3	366.3	254.0	0.5	371.1	366.2	254.0	0.5

the combined bias of the estimators of ARL and MRL, based on averaging, is not a monotonic function of the sample size n, and attains its minimum at n approximately equal to 40. On the other hand, when we use estimators based on the medians of ARL's and MRL's the negative bias is monotonically decreasing with the increase of sample sizes.

In Tables 3 and 4 we present the results of similar simulation experiments for autocorrelated data when the autocorrelation is described by the autoregression model of the first order—AR(1) model. We consider four cases of the strength of dependence, described by the autocorrelation coefficients equal to -0.9, -0.5, 0.5, and 0.9, respectively.

The interpretation of the results presented in Tables 3 and 4 is similar to that in the case of independent data, and confirms findings of many other researchers. In general, the estimators of ARL's and MRL's, based on averaging, are positively biased, and those based on medians are biased negatively. The estimators based on averaging are more sensitive to the strength of dependence in the case of negative

dependence. In case of the positive dependence the observed bias practically does not depend on the strength of dependence (except for very small sample sizes).

Extreme skewness of the distributions of ARL's and MRL's has a very negative impact on the investigations based on computer simulations. If we use averages (over a set of simulated control charts) for the estimation purposes even in the case of thousands of simulated charts few outlying cases, that make the value of skewness so high, may dramatically change the results of estimation. Therefore, one would prefer to use the median as the more robust estimator of ARL's and MRL's. However, in the case of averages we have a commonly accepted benchmark value, the ARL for an in-control state equal to 370.4, but for the median of ARL's such a benchmark does not exist. Therefore, in this chapter we will focus on the approach in which we use the average of ARL's, noting that in future research the approach with the median will be more appropriate.

Let us look at the problem described above from a different point of view. When sample sizes are small it is always possible to design a control chart with too wide control lines. In such a case even large shifts of the process level will not be detected. Therefore, if we observe a long sequence of observations between the control lines we can only say that the process is either in the in-control state or the control limits are too wide. In such situation one can think about an additional decision rule: to stop monitoring after M_R observations, and to redesign the chart. In this research we consider the case when the curtailment value is set to 1000 observations, and applied only in the case of samples not greater than 100. The impact of this curtailment on properties of the distributions of ARL's and MRL's can be inferred from the data contained in Tables 5, 6, and 7

The comparison of the results presented in Tables 1, 2, 3, 4 and Tables 5, 6, 7 shows that all considered characteristics for curtailed experiments are, especially for small sample sizes, significantly smaller than those for (practically) uncurtailed ones. It is worth noting than in the case of a classical Shewhart control charts with known process parameters the curtailment at 1000 observations decreases the ARL from 370 to 345. In the considered in this chapter case of the X chart for residuals with estimated control lines this decrease seems to be more significant. Therefore, we cannot use the value of 370 as the target value of the ARL. The choice of such target value, especially when we estimate control lines from small samples, seems to be a still open question.

Table 5 Characteristics of the ARL and MRL distributions for X charts for residuals—independent observations, runs curtailed at 1000

	ARL				MRL			
n	AvgARL	StdARL	MedARL	SkewARL	AvgMRL	StdMRL	MedMRL	SkewMRL
20	159.1	211.8	66.2	2.0	142.8	237.3	47.0	2.6
30	199.3	209.1	113.8	1.6	175.8	237.8	80.0	2.3
40	224.4	201.2	151.0	1.4	195.0	231.8	106.0	2.2
50	243.0	193.6	178.4	1.2	208.8	224.4	125.0	2.1
100	288.3	159.7	253.4	0.9	236.8	183.9	180.0	2.0

Table 6 Properties of the X chart for residuals with dependent observations—negative autocorrelation, runs curtailed at 1000

	$\rho = -0.9$				$\rho = -0.5$			
n	AvgARL	MedARL	MedMRL	SkewARL	AvgARL	MedARL	MedMRL	SkewARL
20	295.1	165.5	115.0	0.97	196.2	83.6	59.0	1.7
30	294.8	201.1	141.0	0.95	227.4	132.3	92.0	1.4
40	296.7	222.7	157.0	0.93	249.0	149.6	118.5	1.2
50	301.2	241.2	170.5	0.88	264.5	200.0	140.0	1.1
100	315.3	284.2	204.0	0.75	301.3	267.6	191.0	0.82

Table 7 Properties of the X chart for residuals with dependent observations—positive autocorrelation, runs curtailed at 1000

	$\rho = 0.9$				$\rho = 0.5$			
n	AvgARL	MedARL	MedMRL	SkewARL	AvgARL	MedARL	MedMRL	SkewARL
20	190.8	70.0	49.0	1.7	172.5	67.8	48.0	1.9
30	220.0	117.4	82.0	1.4	214.1	119.2	83.0	1.5
40	241.0	156.8	109.0	1.3	249.1	169.3	113.0	1.3
50	255.0	187.4	131.0	1.1	259.7	193.7	135.0	1.1
100	296.8	262.8	187.0	0.82	299.5	265.8	189.0	0.83

4.2 Properties of XWAM Charts for Residuals

The results of experiments presented in the previous section can be used for comparison purposes when we investigate properties of the newly proposed XWAM chart for residuals. The properties of the XWAM chart, described in Sect. 2, have been analyzed using extensive simulation experiments. The outer loop of the experiment consisted of the generation of N_C XWAM control charts, and for each chart N_R process runs have been generated in the inner loop of the experiment. All runs were curtailed at a predefined value M_R. Then, four characteristics have been calculated: average ARL (AvgARL), median ARL (MedARL), average MRL (AvgMRL), and median MRL (MedMRL), where MRL is the median of observed run lengths.

In order to illustrate the design of the proposed XWAM chart consider the case when a chart has to be designed basing on 20 observations from a monitored process. The data presented below have been generated from an autoregressive process of the second order, AR(2), with the parameters 0.7 and −0.9, respectively.

$$-0.94,\ -1.35,\ -1.4,\ -0.34,\ 0.71,\ 0.14,\ -0.49,\ -1.08,\ 1.41,\ 2.25$$
$$-0.05,\ -2.74,\ -1.02,\ 2.95,\ 4.03,\ 0.43,\ -2.67,\ -2.45,\ 1.98,\ 4.25,$$

The autoregression model estimated from these data using the BIC criterion is the AR(2) model with the parameters (0.6094, −0.8236). Using the algorithm described in Sect. 3 we have found five best alternative models. All of them are

Table 8 Chart in-control characteristics for different weights assigned to the estimated model

w_0	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
ARL	97.1	184.5	279.7	352.4	387.7	395.2	383.3	357.7	327.9	292.8	260.9
MRL	68.0	132.0	202.0	260.0	296.0	305.0	292.0	262.0	235.0	207.0	186.0

AR(2) models with parameters $(-0.1, -0.5)$, $(0.1, -0.4)$, $(0.4, -0.7)$, $(-0.5, -0.5)$, and $(0.3, -0.7)$, respectively. The weights assigned to these alternative models were approximately the same ($w_i^{'} = 0.2, i = 1, \ldots, 5$). The estimated model is different from the original model used in simulations, but not too much. However, the alternative models are not very close to the original one, as one could expect.

For the control chart designed using this sample and respective models of the process we have generated, from the original $(0.7, -0.9)$ model, 5000 runs of the process, curtailed at $M_R = 1000$. The values of its characteristics, ARL and MRL, are presented in Table 8 for different values of the weight w_0 assigned to the estimated model.

The results presented in Table 8 illustrate the role of alternative models. Their inclusion widens the control limits, and thus increases the values of chart's characteristics such as ARL and MRL. This feature is easily explained if we notice that residuals calculated using the estimated model are minimal or close to minimal possible (minimal possible residuals can be obtained when Burg's algorithm, not considered in this chapter, is used for the estimation of the process model). Thus, sample residuals calculated using any other model (even the true one!) are usually larger, and their inclusion leads to the widening of control lines, and thus to the increase of ARL's and MRL's. Therefore, the inclusion of alternative models is beneficiary only when the run lengths of a chart designed using an estimated model are shorter than expected. In this particular case the optimal weight of the estimated model seems to be close to 0.7. For bigger and smaller weights the number of false alarms is too high.

The model's parameters describing the sample considered above are not so much different from the parameters used in simulations. However, when sample sizes are small, it may not be the case. Consider, for example, the following sample that has been generated using the same model.

$$\begin{array}{cccccccccc} -3.48 & 0.82 & 2.70 & 1.44 & -1.72 & -3.19 & -0.91 & 3.01 & 2.07 & -0.74 \\ -1.12 & -0.12 & -1.73 & -0.25 & -1.32 & -2.3 & 1.48 & 6.52 & 3.13 & -3.83 \end{array}$$

The autoregression model estimated from these data using the BIC criterion is the AR(2) model with the parameters $(0.379, -0.6094)$. Using the algorithm described in Sect. 3 we have found five best alternative models, and all of them are AR(2) models with parameters $(0.8, -0.8)$, $(0.6, -0.8)$, $(0.9, -0.6)$, $(0.4, -0.7)$, and $(-0.7, -0.4)$, respectively. The weights assigned to these alternative models were approximately the same ($w_i^{'} = 0.2, i = 1, \ldots, 5$). It has to be noted that in the case of this particular sample the estimated model differs from the original one.

Table 9 Chart in-control characteristics for different weights assigned to the estimated model—extreme sample

w_0	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
ARL	947.4	966.7	980.3	988.4	991.9	994.2	997.4	998.7	999.4	999.7	999.8
MRL	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0

Table 10 Average in-control ARL for different weights assigned to the estimated model, n = 20

Model/w_0:	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
AR(−0.9)	256.6	290.7	322.6	348.1	368.1	383.4	395.3	404.4	411.2	416.0	419.3
AR(−0.5)	200.2	212.7	226.1	240.4	255.3	270.5	285.3	299.3	312.0	322.9	331.7
AR(0)	177.4	185.3	194.3	204.5	216.0	229.2	244.4	261.7	281.4	303.2	326.3
AR(0.5)	165.8	185.0	204.2	222.7	239.7	254.5	267.1	275.0	284.8	289.5	291.4
AR(0.9)	155.5	173.5	181.3	182.0	178.4	172.3	164.9	157.1	149.2	141.7	134.6
AR(0.7 , −0.9)	201.2	381.6	466.0	507.2	526.8	533.3	531.5	524.5	514.0	501.2	487.3

Table 11 Median in-control ARL for different weights assigned to the estimated model, n = 20

Model/w_0:	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
AR(−0.9)	131.7	161.9	187.0	214.7	230.6	243.6	252.8	260.4	271.9	285.2	419.3
AR(−0.5)	87.1	100.7	111.5	124.6	139.0	155.1	173.1	187.7	198.7	218.8	229.4
AR(0)	69.8	76.2	82.8	90.7	99.2	112.4	123.2	140.5	164.5	187.8	213.1
AR(0.5)	66.5	76.9	87.5	104.8	118.3	132.5	140.5	150.7	158.7	168.6	173.2
AR(0.9)	55.2	61.5	60.5	57.5	55.0	49.9	48.0	45.6	43.1	40.7	38.5
AR(0.7 , −0.9)	96.6	233.0	370.5	478.0	519.3	539.5	517.2	487.7	450.8	416.8	377.4

In Table 9 we present the results of a similar simulation experiment as in the case of the first considered sample. The residuals calculated from the estimated model are, in this case, large, and thus the values of ARL and MRL are much larger from expected. Therefore, by adding alternative models (even if they are closer to the original one) we do not improve the situation, and the values of ARL and MRL remain too high.

The results presented in Tables 8 and 9 illustrate the operation of the proposed algorithm for particular samples. More general properties of the proposed XWAM control chart have been investigated in numerous simulation experiments for different models describing autocorrelation. In Tables 10 and 11 we present the average and median values of ARL's evaluated from 1000 generated control charts, and 5000 process runs generated for each control chart, i.e., from all together 5,000,000 simulated process runs. In all these runs the simulated process was in the in-control state, and the length of one run was curtailed at 1000 process observations.

The results presented in Tables 10 and 11 are very interesting from many points of view. First, let us notice that the values of averages ARL's in the in-control state (Table 10) are significantly different from the respective values of medians of averages (Table 11), and this difference is greater than in the cases described

in Tables 5, 6, and 7, where we simulated 50,000 charts. Moreover, the values of average and median ARLs obtained in the case when only a simulated sample was used for the design of the chart (classical X chart for residuals, $w = 1$) are lower than their counterparts from Tables 5, 6, and 7. These differences tell us that the number of samples (charts) used for the evaluation of properties of the XWAM chart (1000), and the sample sizes used, are not sufficient for precise estimation of ARL's. Unfortunately, the simulation of XWAM charts is time consuming (due to the time used for finding alternative models), and simulation of a much larger number of considered charts is, unfortunately, infeasible. Therefore, the results presented in the following part of this section have, as for now, rather qualitative character. The second, and the most important, feature of the XWAM chart that can be inferred from Tables 10 and 11 is the following: by taking into account alternative models we usually increase the values of ARL's in the in-control state. However, for strong positive correlations ($\rho = 0.9$) or more complicated models (AR(0.7, −0.9)) this behavior is slightly different. With a decreasing value of w_0 the value of ARL increases, then attains a maximum and for smaller values of w_0 decreases. This phenomenon is possibly due to imprecise estimation of model's parameters for very small samples ($n = 20$), and processes of these types. When we use a classical X control chart for residuals ($w_0 = 1$) the averages (and medians) of ARL's are rather small, and this means that the rate of false alarms may be too high. By using the XWAM chart with the parameter $w_0 < 1.0$ we significantly decrease the rate of false alarms. The question arises then if the discriminative power of such charts is sufficiently good. The answer to this question is presented in the following part of this section.

The results presented in Tables 10 and 11 show how the concept of the XWAM control chart works in practice when monitored processes are in the in-control state. What is equally important, however, it is the ability of a chart to detect shifts of a monitored process. In this chapter we consider only shifts of the average value, measured in units of standard deviation. Let us begin this analysis from considering the case of independent sample and process observations. The respective values of the averaged over 1000 considered charts values of the ARL are presented in Table 12.

Table 12 Average ARL for different weights assigned to the estimated model and different shifts of the process level, $\rho = 0$, n = 20

Shift/w_0:	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
−3	4.9	4.8	4.8	4.7	4.7	4.7	4.7	4.8	4.8	5.0	5.2
−2	9.8	9.7	9.6	9.6	9.6	9.8	10.0	10.4	11.0	12.0	13.5
−1	46.9	48.3	50.0	52.3	55.0	58.6	62.9	68.4	75.2	84.0	95.4
0	177.4	185.3	204.6	220.2	237.3	255.9	276.1	297.3	318.6	339.2	358.0
1	50.0	51.3	53.0	55.0	57.6	60.8	64.8	70.0	76.6	85.1	96.4
2	10.4	10.3	10.2	10.2	10.2	10.4	10.6	11.0	11.6	12.5	13.9
3	5.0	4.9	4.8	4.8	4.8	4.8	4.8	4.8	4.9	5.0	5.2

Table 12 shows a very interesting feature of the XWAM chart for residuals. When we decrease the weight w_0 assigned to the estimated model the respective values of the average ARL's are increasing when shifts of the process mean are either not present (the in-control state) or are small (e.g., the shift of one standard deviation). For larger shifts these values for the XWAM chart are similar or even smaller (!) than in the case of the classical X chart for residuals ($w_0 = 1$). It means that for the XWAM chart for residuals the rate of false alarms is smaller than in the case of the X chart for residuals. However, the rate of (expected) alarms is similar or even smaller for large shifts of the process level. Thus, the newly proposed chart seems to be more effective than the classical control chart for residuals.

Let us consider now the same problem when consecutive observations are autocorrelated. In Table 13 we show the average values of ARL for different shifts when we use a sample of 20 elements, and the observations are positively, but not very strongly, correlated ($\rho = 0.5$).

From Table 13 it can be seen quite clearly that the behavior of the XWAM chart in this case is nearly the same as in the case of independent observations. The XWAM chart has better discriminative power, calculated as the quotient of the ARL in the out-of-control state (shifted process) and the ARL in the in-control state. Respective values of the coefficient of discriminative power are presented in Table 14. It is interesting that in this case an "optimal" behaviour is to *neglect* the estimated model ($w_0 = 0$). Observed sample is used in this case only for finding good alternative

Table 13 Average ARL for different weights assigned to the estimated model and different shifts of the process level, $\rho = 0.5$, n = 20

Shift/w_0:	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
−3	16.7	16.3	15.7	15.0	14.2	13.5	12.8	12.2	11.8	11.4	11.2
−2	43.0	47.0	45.8	46.4	46.6	46.4	45.9	45.1	44.1	43.1	42.2
−1	107.2	117.2	126.4	135.0	142.7	149.2	154.3	158.0	160.3	161.4	161.5
0	165.8	185.0	204.2	222.7	239.7	254.5	267.1	275.0	284.8	289.5	291.4
1	105.2	115.3	124.6	132.7	139.5	144.6	148.1	150.2	151.3	151.5	151.0
2	42.0	44.0	45.3	46.0	46.3	46.1	45.7	44.9	43.9	42.6	41.4
3	16.2	15.9	15.4	14.7	14.0	13.3	12.5	11.9	11.4	11.1	10.8

Table 14 Discriminative power of the XWAM chart for different shifts of the process level, $\rho = 0.5$, n = 20

Shift/w_0:	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
−3	0.101	0.088	0.077	0.067	0.059	0.053	0.048	0.044	0.041	0.039	0.038
−2	0.259	0.254	0.224	0.208	0.194	0.182	0.172	0.164	0.155	0.149	0.145
−1	0.647	0.634	0.619	0.606	0.552	0.586	0.578	0.575	0.563	0.558	0.554
0	1	1	1	1	1	1	1	1	1	1	1
1	0.634	0.623	0.610	0.596	0.582	0.568	0.554	0.546	0.531	0.523	0.518
2	0.253	0.238	0.222	0.207	0.193	0.181	0.171	0.163	0.154	0.147	0.142
3	0.100	0.086	0.075	0.066	0.058	0.052	0.047	0.043	0.040	0.038	0.037

Table 15 Average ARL for different weights assigned to the estimated model and different shifts of the process level, $\rho = -0.5$, n = 20

Shift/w_0:	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
−3	3.1	3.1	3.1	3.1	3.2	3.2	3.2	3.2	3.2	3.3	3.4
−2	4.3	4.3	4.4	4.5	4.6	4.7	4.8	5.0	5.2	5.4	5.8
−1	23.5	24.8	26.2	27.8	29.5	31.4	33.5	35.9	38.6	41.7	45.2
0	200.2	212.7	226.1	240.4	255.3	270.5	285.3	299.3	312.0	322.9	331.7
1	23.0	24.3	25.7	27.3	29.0	30.9	33.0	35.5	38.3	41.6	45.6
2	4.3	4.4	4.4	4.5	4.6	4.7	4.9	5.2	5.2	5.5	5.8
3	3.1	3.1	3.1	3.2	3.2	3.2	3.2	3.2	3.3	3.3	3.4

Table 16 Discriminative power of the XWAM chart for different shifts of the process level, $\rho = -0.5$, n = 20

Shift/w_0:	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
−3	0.016	0.015	0.014	0.013	0.012	0.012	0.011	0.011	0.010	0.010	0.010
−2	0.021	0.020	0.019	0.019	0.018	0.017	0.017	0.017	0.017	0.017	0.017
−1	0.117	0.117	0.116	0.116	0.116	0.116	0.117	0.120	0.124	0.129	0.136
0	1	1	1	1	1	1	1	1	1	1	1
1	0.115	0.114	0.114	0.114	0.114	0.114	0.117	0.119	0.123	0.129	0.137
2	0.021	0.021	0.020	0.019	0.018	0.017	0.017	0.017	0.017	0.017	0.017
3	0.016	0.015	0.014	0.013	0.012	0.012	0.012	0.011	0.011	0.010	0.010

models. One should also note that a simple widening of control limits for a classical chart for residuals will increase the in-control ARL to the required value, but—contrary to the case of the XWAM chart—also automatically will increase the value of ARLs for shifted processes. In such a case, the average time to alarm signal for large shifts will be much greater than the respective time for the proposed XWAM chart. The results of similar experiment performed for other autoregression models show a similar behavior of the XWAM chart.

Interesting case is presented in Tables 15 and 16. In these tables we consider the case of negative dependence of medium strength ($\rho = -0.5$). The value of w_0 for which the discriminative power is the best is in this case about 0.5. Therefore, if we need to balance somehow the requirement for small rate of false alarms and good discrimination properties we should use an "optimal" value of w_0 from an interval $[0, 0.5]$. It means that in this case for the construction of a control chart we should use appropriately weighted modes, both estimated and alternative.

Finally, let's consider the influence of a sample size on the performance of XWAM chart. This problem is rather seldom considered in literature (see Köksal et al. (2008) for more information). In Table 17 we present the comparison between the values of ARL's for two sample sizes, 20 and 50. The process used for comparisons is the autoregressive process of the first order $AR(0.9)$. We have deliberately chosen this process, as in this case the performance of the classical

Table 17 Values of ARL for different sample sizes, $\rho = 0.9$

	$w_0 = 1.0$		$w_0 = 0.8$		$w_0 = 0.6$		$w_0 = 0.4$		$w_0 = 0.2$		$w_0 = 0.0$	
Shift/n	20	50	20	50	20	50	20	50	20	50	20	50
−3.0	107.8	167.0	114.2	177.4	107.9	169.4	99.2	155.2	91.0	140.2	83.5	126.1
−2.0	130.5	206.5	144.0	230.0	138.9	226.9	128.3	212.9	117.3	195.8	107.2	178.4
−1.0	148.5	239.7	169.5	276.5	165.9	279.8	153.5	267.2	139.3	248.7	126.2	228.7
0.0	155.5	254.8	181.3	298.1	178.4	305.0	164.9	293.4	149.2	274.1	134.6	252.3
1.0	148.2	243.5	174.2	281.1	169.4	285.5	155.5	273.0	140.2	254.1	126.3	233.2
2.0	129.5	210.3	151.1	233.7	143.8	231.4	129.9	217.5	116.5	199.8	104.8	181.8
3.0	105.9	168.8	121.1	178.1	112.6	170.1	100.3	156.0	89.3	140.6	80.1	126.4

Table 18 Discriminative power of the XWAM chart for different sample sizes, $\rho = 0.9$

	$w_0 = 1.0$		$w_0 = 0.8$		$w_0 = 0.6$		$w_0 = 0.4$		$w_0 = 0.2$		$w_0 = 0.0$	
Shift/n	20	50	20	50	20	50	20	50	20	50	20	50
−3.0	0.69	0.66	0.63	0.60	0.60	0.56	0.60	0.53	0.61	0.51	0.62	0.50
−2.0	0.84	0.81	0.79	0.77	0.78	0.74	0.78	0.73	0.79	0.71	0.80	0.71
−1.0	0.95	0.94	0.93	0.93	0.93	0.92	0.93	0.91	0.93	0.91	0.94	0.91
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	0.95	0.96	0.96	0.94	0.95	0.94	0.94	0.93	0.94	0.93	0.94	0.92
2.0	0.84	0.83	0.83	0.78	0.81	0.76	0.79	0.44	0.78	0.73	0.78	0.72
3.0	0.68	0.66	0.67	0.60	0.63	0.56	0.61	0.53	0.60	0.51	0.60	0.50

X chart for residuals is, as it was already noticed by many authors, very poor. Therefore, the question is if the usage of the XWAM chart helps in this difficult case.

The results of simulations presented in columns 2 and 3 of Table 17 confirm already known results that classical X charts for residuals perform very badly. The rate of false alarms is extremely high, and, on the other hand, average times to alarm are also very high, even for very large shifts of process levels. This is also confirmed in Table 18 where respective coefficients of discriminative power are displayed.

The performance of a classical control chart for residuals ($w_0 = 1$) for a very small sample size ($n = 20$) is very bad. The rate of false alarms is high, and discrimination rate is unsatisfactory (even large shifts of the process level are detected with a large delay). This bad behavior is due to the correlation of residuals which affects the estimated standard deviation. When we increase the sample size to $n = 50$ the false alarm rate is, as expected, significantly decreased. However, the discrimination rate is only slightly better. The performance of respective XWAM charts is slightly better if we take the value of w_0 from an interval $[0.4, 0.6]$. The false alarm rate is for both considered sample sizes lower, and the discrimination rate slightly better. Thus, similarly to previously considered cases, the performance of the newly proposed XWAM chart is better than the performance of the classical X chart for residuals.

5 Conclusions

In the chapter we have proposed a new method for the construction of the Shewhart X control chart for residuals. The inspiration of the proposed methodology comes from the concept of the Bayesian model averaging, already successfully applied by econometricians in the analysis of economic short time series. The novelty of the proposed approach consists in the new method for the calculation of weights. Following our previous experience with prediction models for short time series, we propose to compute these weights using methods of data mining. In this particular research we use the methodology of Dynamic Time Warping (DTW) for finding alternative models for the considered sequence of observations. We use artificially generated template time series, and find these series (and in consequence their models) our data are similar to. Then, the degrees of similarity are used for the computation of model weights. In this research the template time series have been generated from simple autoregressive models. However, the proposed approach is more general, and allows to use as a template any well identified time series.

In order to evaluate the proposed methodology we have performed many simulation experiments. In this chapter, due to a limit for its volume, we have presented the results of only some of them. The presented results can be regarded as a positive "proof of concept". Control charts designed according to the proposed methodology have better properties than traditionally designed Shewhart X control charts for residuals. However, the properties of these improved charts are often unsatisfactory from a practical point of view. Therefore, there is a need to apply the proposed methodology for such control charts for residuals as EWMA or CUSUM, which have been proved to perform better than the X chart. Moreover, further research is needed with the aim to find good alternative models using less computational effort.

References

Akaike, H. (1978). Time series analysis and control through parametric model. In D. F. Findley (ed.), *Applied Time Series Analysis*. New York: Academic Press

Albers, W., & Kallenberg, C. M. (2004). Estimation in Shewhart control charts: Effects and corrections. *Metrika, 59*, 207–234.

Alwan, L. C., & Roberts, H. V. (1988). Time-Series Modeling for statistical process control. *Journal of Business & Economic Statistics, 6*, 87–95.

Apley, D. W., & Chin, C. (2007). An optimal filter design approach to statistical process control. *Journal of Quality Technology, 39*, 93–117.

Apley, D. W., & Lee, H. C. (2008). Robustness comparison of exponentially weighted moving-average charts on autocorrelated data and on residuals. *Journal of Quality Technology, 40*, 428–447.

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *AAAI-94 Workshop on Knowledge Discovery in Databases* (pp. 359–370).

Box, G. E. P., Jenkins, G. M., & MacGregor, J. F. (1974). Some recent advances in forecasting and control, part II. *Journal of the Royal Statistical Society, Series C, 23*, 158–179.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time Series Analysis. Forecasting and Control* (4th ed.). Hoboken, NJ: J.Wiley.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to Time Series and Forecasting* (2nd ed.). New York: Springer.

Chakraborti, S. (2000). Run length, average run length and false alarm length of Shewhart X-bar chart: Exact derivations by conditioning. *Communications in Statistics – Simulations and Computations, 29*, 61–81.

Chin, C.-H., & Apley, D. W. (2006). Optimal design of second-order linear filters for control charting. *Technometrics, 48*, 337–348.

De Ketelaere, B., Hubert, M., & Schmitt, E. (2015). Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. *Journal of Quality Technology,47*, 318–335.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. Hoboken, NJ: J. Wiley

Hryniewicz, O., & Katarzyna Kaczmarek-Majer (2016a). Bayesian analysis of time series using granular computing approach. *Applied Soft Computing Journal, 47*, 644–652.

Hryniewicz, O., & Katarzyna Kaczmarek-Majer (2016b). Monitoring of short series of dependent observations using a control chart approach and data mining techniques. In *Proceedings of the International Workshop ISQC 2016, Helmut Schmidt Universität, Hamburg* (pp. 143–161).

Jiang, W., Tsui, K., & Woodall, W. H. (2000). A new SPC monitoring method: The ARMA chart. *Technometrics,42*, 399–410.

Köksal, G., Kantar, B., Ula, T. A., & Testik, M. C. (2008). The effect of Phase I sample size on the run length performance of control charts for autocorrelated data. *Journal of Applied Statistics, 35*, 67–87.

Kramer, H., & Schmid, W. (2000). The influence of parameter estimation on the ARL of Shewhart type charts for time series. *Statistical Papers, 41*, 173–196.

Lu, C. W., & Reynolds, M. R. Jr. (1999). Control charts for monitoring the mean and variance of autocorrelated processes. *Journal of Quality Technology, 31*, 259–274.

Maragah, H. D., & Woodall, W. H. (1992). The effect of autocorrelation on the retrospective *X*-chart. *Journal of Statistical Computation and Simulation, 40*, 29–42.

Montgomery, D. C., & Mastrangelo, C. M. (1991). Some statistical process control methods for autocorrelated data (with discussion). *Journal of Quality Technology, 23*, 179–204.

Runger, G. C. (2002). Assignable causes and autocorrelation: Control charts for observations or residuals? *Journal of Quality Technology, 34*, 165–170.

Schmid, W. (1995). On the run length of Shewhart chart for correlated data. *Statistical Papers, 36*, 111–130.

Vasilopoulos, A. V., & Stamboulis, A. P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Technology, 10*, 20–30.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery, 26*, 275–309.

Wardell, D. G., Moskowitz, H., & Plante, R. D. (1994). Run-length distributions of special-cause control charts for correlated processes (with disscussion). *Technometrics, 36*, 3–27.

Yashchin, E. (1993). Performance of CUSUM control schemes for serially correlated observations. *Technometrics, 35*, 37–52.

Zhang, N. F. (1997). Detection capability of residual chart for autocorrelated data. *Journal of Applied Statistics, 24*, 475–492.

Zhang, N. F. (1998). A statistical control chart for stationary process data. *Technometrics, 40*, 24–38.

Challenges in Monitoring Non-stationary Time Series

Taras Lazariv and Wolfgang Schmid

Abstract Different approaches for monitoring non-stationary processes are discussed. Besides the transformation method, a more general procedure is described which makes use of the probability structure of the underlying in-control process. Here, the in-control process is assumed to be a multivariate state-space process. The out-of-control state is described by a general change point model which covers shifts and drifts in the components. Control charts with a reference vector are derived using the likelihood ratio, the sequential probability ratio, and the Shiryaev–Roberts approach. Moreover, the generalized likelihood ratio, the generalized sequential probability ratio, and the generalized modified Shiryaev–Roberts approach are used to obtain charts without reference parameters. All the introduced schemes are compared with each other assuming that a univariate unit root process with drift is present. We make use of several measures of the performance of control charts, such as the average run length, the worst average delay, and the limit average delay. This chapter also analyses how the charts with a reference parameter depend on the choice of this quantity.

Keywords Control charts · Statistical process control · Change-point detection · Time series · State-space model

1 Introduction

In the last 30 years, monitoring problems have been discussed in many areas, e.g., in economics (Frisén 2008), medicine (Kass-Hout and Zhang 2010), and the environmental sciences (Chou 2004). It has turned out that there are many further applications beyond engineering, the original field of its application (e.g., Mont-

T. Lazariv · W. Schmid (✉)
Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany
e-mail: lazariv@europa-uni.de; schmid@europa-uni.de

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_14

gomery 2009). In order to apply control charts to these new areas it was necessary to adapt the idea behind control schemes to these processes and sometimes to extend and modify the original approaches. In many situations the underlying processes are time series, the data have a memory and the variables are no longer independent.

In nearly all of the chapters just referred to, the underlying time series is assumed to be (weakly) stationary in the in-control state. Nowadays, the literature often distinguishes between residual charts and modified schemes. Residual charts are based on the idea of transforming the original data so that the transformed variables are independent. Then the well-known approaches of statistical process control for independent and identically distributed random variables can be applied to the transformed quantities. In contrast, modified schemes make use of the original observations. They are obtained by taking into account the probability structure of the underlying time series process. Residual charts have been discussed by Alwan and Roberts (1988), Wardell et al. (1994a,b), and Lu and Reynolds (1999), among others, while modified charts have been treated by, e.g., Nikiforov (1975) and Schmid (1995, 1997a,b).

In many applications, however, especially in economics, the process of interest frequently turns out to be close to non-stationarity or even is non-stationary: either it is not oscillating around a common mean or its variance and autocovariances are changing over time. The existing techniques fail at monitoring such processes. Therefore, it is important to have tools that can correctly detect changes in non-stationary processes. Monitoring non-stationary processes is a new field and it has not yet received much attention. Of course it is impossible to distinguish between a non-stationary process and a non-stationary process with change if no information on the probability structure of the underlying in-control non-stationary process is given.

Schmid and Steland (2000) applied nonparametric kernel control charts to a non-stationary process to analyse whether its derivative has significantly changed. Nonparametric procedures for monitoring time series have been proposed by, e.g., Steland (2002, 2005, 2007, 2010). Triantafyllopoulos and Bersimis (2016) proposed a Bayesian approach to monitor a possibly non-stationary process. A parametric approach was chosen by Lazariv and Schmid (2015). They used state-space models to model the underlying in-control process. These processes are very flexible and allow modeling a large family of non-stationary processes. Several control charts for detecting a mean shift were derived.

In the current chapter we want to discuss various techniques for deriving control charts for non-stationary processes. Nonparametric techniques are not employed. One approach is based on differencing, i.e. the original data are transformed by successively calculating the differences between two successive observations. This procedure is applied until the resulting process is stationary. Such an approach is frequently applied in econometrics. In relation to monitoring, it has been studied by Steland (2005, 2007), among others, to detect a change in a unit root process (see, e.g., Hayashi 2000). Similar to residual charts for stationary processes, this technique is based on suitably transforming the original data, here to a stationary process. Another attempt is to directly describe the possibly non-stationary in-

control process by a stochastic model and to derive charts by making use of the probability structure of the underlying process using the likelihood approach, the Shiryaev–Roberts method, etc. In the present chapter as well we will employ this approach. As in Lazariv and Schmid (2015), state-space models will be used to model the in-control process, but instead of a mean shift model we consider a more general out-of-control situation covering, e.g., mean drifts as well. In this chapter, the resulting charts are compared with the charts obtained by differencing. The underlying process is a random walk with drift.

2 Handling Non-stationary Processes

In practice there are different approaches to handling non-stationary processes. In economics, a popular approach is to transform the original process in a suitable way. If the underlying process is a unit root process, differencing is a widely applied procedure (e.g., Hayashi 2000). That approach will be briefly described in the next section. Another possibility is of course to directly model the non-stationary process by a suitable model. In this regard, state-space models are frequently applied, since they are on the one hand very flexible and on the other hand there are computational techniques that enable carrying out the statistical analysis of these processes very quickly.

2.1 Unit Root Problems

One of the major issues in finance is how to model the probability distribution of stock prices. In many areas of finance the standard model is the random walk in discrete time or its counterpart in continuous time, the Brownian motion (Ruppert 2004). In discrete time, this means

$$Y_t = Y_{t-1} + \varepsilon_t, \quad t \geq 1 \tag{1}$$

with $Y_0 = y_0$. The differences between two successive observations, here $\{\varepsilon_t\}$, are usually assumed to follow a white noise process. In the following, however, $\{\varepsilon_t\}$ may be a (weakly) stationary process. Now it may happen that after some time the process drifts away. In finance it is important to detect such a drift as early as possible. This situation can be described by the following change point model.

$$X_t = \begin{cases} Y_t & \text{for} \quad 1 \leq t < \tau \\ Y_t + (t - \tau + 1)a & \text{for} \quad t \geq \tau \end{cases}. \tag{2}$$

Fig. 1 Facebook share prices with estimated trend ($\hat{a} = 0.1171$ in the out-of-control period)

Here $\{X_t\}$ denotes the observed process and $\{Y_t\}$ the target (in-control) process, while τ is the unknown position of the change point. In the in-control state, i.e. for $\tau = \infty$, the observed process is a random walk, but in the out-of-control situation it is a random walk with drift.

Now the target process may have a deterministic trend as well. In that case,

$$Y_t = Y_{t-1} + \beta t + \varepsilon_t, \quad t \geq 1 \tag{3}$$

with $Y_0 = y_0$. This is a random walk with deterministic trend. Applying (2), the out-of-control process describes a random walk with deterministic trend and drift.

An example of such behaviour is presented in Fig. 1, where the daily closing prices of Facebook from May 18, 2012 to May 29, 2016 are plotted. The period from May 18, 2012 to August 1, 2013 shows the possible in-control behaviour.

2.2 State-Space Models

State-space models have been widely used in engineering. In recent years, more applications in economics have been found (Durbin and Koopman 2012). State-space models are quite flexible and cover a huge variety of processes.

We suppose that the in-control process $\{\boldsymbol{Y}_t\}$ is a p-dimensional time series following a state-space model, i.e. $\{\boldsymbol{Y}_t\}$ satisfies the set of equations

$$\boldsymbol{Y}_t = \boldsymbol{G}_t \boldsymbol{S}_t + \boldsymbol{W}_t, \quad t = 1, 2, \ldots, \qquad \text{where} \tag{4a}$$

$$\boldsymbol{S}_{t+1} = \boldsymbol{F}_t \boldsymbol{S}_t + \boldsymbol{V}_t, \quad t = 1, 2, \ldots. \tag{4b}$$

Equation (4a) is called the observation equation. The process $\{\boldsymbol{Y}_t\}$ is obtained from $\{\boldsymbol{S}_t\}$ by applying a linear transformation and adding a random noise variable $\boldsymbol{W}_t$. The state equation (4b) is q-dimensional and it describes the evolution of the state $\boldsymbol{S}_t$ over time.

In the next sections we will assume the following.

(A1) Suppose that, for all $t \geq 1$,

$$E\begin{pmatrix} \boldsymbol{V}_t \\ \boldsymbol{W}_t \end{pmatrix} = \boldsymbol{0}, \quad E(\boldsymbol{V}_t \boldsymbol{V}_t') = \boldsymbol{Q}_t, \quad E(\boldsymbol{W}_t \boldsymbol{W}_t') = \boldsymbol{R}_t, \quad E(\boldsymbol{V}_t \boldsymbol{W}_t') = \boldsymbol{U}_t.$$

Furthermore, $\{\boldsymbol{Q}_t\}$, $\{\boldsymbol{R}_t\}$, and $\{\boldsymbol{U}_t\}$ are specified sequences of $q \times q$, $p \times p$ and $q \times p$ matrices, respectively.

(A2) Assume that $\boldsymbol{S}_1, (\boldsymbol{V}_1', \boldsymbol{W}_1')', (\boldsymbol{V}_2', \boldsymbol{W}_2')', \ldots$ are uncorrelated.

(A3) Suppose $E(\boldsymbol{Y}_0 \boldsymbol{V}_t') = \boldsymbol{0}$ and $E(\boldsymbol{Y}_0 \boldsymbol{W}_t') = \boldsymbol{0}$ for all $t \geq 1$.

The parameter matrices $\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{Q}_t, \boldsymbol{R}_t, \boldsymbol{U}_t$ are defined very generally. However, in many applications they are not time-varying and many notations can be simplified and the index t in that case is omitted.

The best one-step ahead linear predictor $\hat{\boldsymbol{S}}_t$ of $\boldsymbol{S}_t$ given $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_{t-1}$ and the corresponding error covariance matrices $\boldsymbol{\Omega}_t = E\left((\boldsymbol{S}_t - \hat{\boldsymbol{S}}_t)(\boldsymbol{S}_t - \hat{\boldsymbol{S}}_t)'\right)$ for model (4) can be calculated using the Kalman recursions (Brockwell and Davis 2009) as

$$\hat{\boldsymbol{S}}_{t+1} = \boldsymbol{F}_t \hat{\boldsymbol{S}}_t + \boldsymbol{\Theta}_t \boldsymbol{\Delta}_t^{-1} (\boldsymbol{Y}_t - \boldsymbol{G}_t \hat{\boldsymbol{S}}_t) \tag{5}$$

with

$$\begin{cases} \boldsymbol{\Delta}_t = \boldsymbol{G}_t \boldsymbol{\Omega}_t \boldsymbol{G}_t' + \boldsymbol{R}_t \\ \boldsymbol{\Theta}_t = \boldsymbol{F}_t \boldsymbol{\Omega}_t \boldsymbol{G}_t' + \boldsymbol{U}_t \\ \boldsymbol{\Omega}_{t+1} = \boldsymbol{F}_t \boldsymbol{\Omega}_t \boldsymbol{F}_t' + \boldsymbol{Q}_t - \boldsymbol{\Theta}_t \boldsymbol{\Delta}_t^{-1} \boldsymbol{\Theta}_t' \end{cases} \tag{6}$$

for $t \geq 1$ and the starting conditions

$$\hat{\boldsymbol{S}}_1 = P(\boldsymbol{S}_1 | \boldsymbol{Y}_0), \ \boldsymbol{\Omega}_1 = E(\boldsymbol{S}_1 \boldsymbol{S}_1') - E(\hat{\boldsymbol{S}}_1 \hat{\boldsymbol{S}}_1').$$

Here $P(\boldsymbol{S}_1 | \boldsymbol{Y}_0)$ denotes the projection of the ith component S_{1i} of $\boldsymbol{S}_1$ on the span of $\boldsymbol{Y}_0$.

In order to start the Kalman filter it is necessary to know the mean and the covariance matrix of $\boldsymbol{S}_1$. In our simulation study we fix these values. In practice, however, they are unknown and have to be suitably determined. Various proposals have been made to do this (see, e.g., Koopman 1997; Durbin and Koopman 2012).

$\hat{\boldsymbol{S}}_{t+1}$ can be rewritten as a linear combination of $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t$

$$\hat{\boldsymbol{S}}_{t+1} = \sum_{j=1}^{t} \boldsymbol{A}_{t+1,j} \boldsymbol{Y}_j + \boldsymbol{a}_{t+1}(\boldsymbol{Y}_0) \tag{7}$$

with $\boldsymbol{A}_{t+1,j} = \left(\boldsymbol{E}_t \cdots \boldsymbol{E}_{j+1}\right) \boldsymbol{\Theta}_j \boldsymbol{\Delta}_j^{-1}$ and $\boldsymbol{E}_t = \boldsymbol{F}_t - \boldsymbol{\Theta}_t \boldsymbol{\Delta}_t^{-1} \boldsymbol{G}_t$. $\boldsymbol{a}_{t+1}(\boldsymbol{Y}_0) = (\boldsymbol{E}_t \cdots \boldsymbol{E}_1)\, \hat{\boldsymbol{S}}_1$ is a function of $\boldsymbol{Y}_0$.

Similarly, there can be obtained the best linear predictor $\hat{\boldsymbol{Y}}_t$ of $\boldsymbol{Y}_t$ given $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_{t-1}$, using the presentation (7)

$$\hat{\boldsymbol{Y}}_t = \boldsymbol{G}_t \hat{\boldsymbol{S}}_t = \sum_{j=1}^{t-1} \boldsymbol{B}_{t,j} \boldsymbol{Y}_j + \boldsymbol{b}_t(\boldsymbol{Y}_0) \tag{8}$$

for $t \geq 1$ with $\boldsymbol{B}_{t,j} = \boldsymbol{G}_t \boldsymbol{A}_{t,j}$ and $\boldsymbol{b}_t(\boldsymbol{Y}_0) = \boldsymbol{G}_t \boldsymbol{a}_t(\boldsymbol{Y}_0)$.

Let $\boldsymbol{\Sigma}_t$ denote the covariance matrix of $\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_t$. Then for $t \geq 1$,

$$\boldsymbol{\Sigma}_t = \boldsymbol{G}_t \boldsymbol{\Omega}_t \boldsymbol{G}_t' + \boldsymbol{R}_t.$$

In this chapter we assume that the parameters of the target process are known. In practice, however, they should be estimated using historical data. The influence of the parameter estimation is an important question, but we will not discuss it in the present chapter.

2.3 Modeling the Out-of-Control Process

In the following we want to consider a more general change-point model

$$\boldsymbol{X}_t = \begin{cases} \boldsymbol{Y}_t & \text{for} \quad 1 \leq t < \tau \\ \boldsymbol{Y}_t + \boldsymbol{D}_{t,\tau} \boldsymbol{a} & \text{for} \quad t \geq \tau \end{cases}, \tag{9}$$

where $\boldsymbol{D}_{t,\tau}$ denotes a known $p \times p$ matrix and $\boldsymbol{a} \in \mathbb{R}^p$ an unknown parameter vector. Choosing $\boldsymbol{D}_{t,\tau} = (t - \tau + 1)\boldsymbol{I}$ we obtain the above drift model and setting $\boldsymbol{D}_{t,\tau} = \boldsymbol{I}$ a mean shift model is obtained. Here $\boldsymbol{I}$ stands for the $p \times p$ identity matrix. Of course it is also possible to take the standard deviation of the process into account. Then, e.g., we have to choose $\boldsymbol{D}_{t,\tau} = diag(\sqrt{Var(Y_{t1})}, \ldots \sqrt{Var(Y_{tp})})$ for the shift model (see Lazariv and Schmid 2015). Of course it is also possible that there are some components with a drift and others with a shift. This can be handled by the presented approach as well.

3 Control Charts for Non-stationary Processes

There are different approaches to derive control charts for non-stationary processes. The easiest one is to transform the original data so that the transformed data follow a stationary process. Then all the well-known procedures for stationary processes can be applied to the transformed quantities. This method is briefly described in the next section. In Sects. 3.2 and 3.3 we introduce control charts for the generalized change point model assuming the in-control process is a state-space model. In Sect. 3.2 the charts are obtained by applying the likelihood ratio approach, the sequential probability ratio method, and the Shiryaev–Roberts procedures. These charts depend on the unknown parameter $\boldsymbol{a}$, which has to be replaced in practice by a suitable reference value. In Sect. 3.3 some generalized procedures are considered, where the corresponding probability density is maximized over $\boldsymbol{a}$ so that the resulting chart does not depend on $\boldsymbol{a}$.

3.1 The Transformation Approach

The transformation method works similar to the residual approach for stationary processes. This method works well for unit root problems. In that case, the differences of two successive observations are calculated until the resulting process is stationary. If, e.g., the in-control process is a simple univariate unit root process as in (1) and the out-of-control process is a drift model as in (2), then

$$X_t^* = X_t - X_{t-1} = \begin{cases} \varepsilon_t & \text{for} \quad 1 \le t < \tau \\ \varepsilon_t + a & \text{for} \quad t \ge \tau \end{cases}, \quad X_0 = 0.$$

Thus differencing leads to the problem of detecting a shift in a stationary problem, which has been intensively discussed in the literature (e.g., Hayashi 2000). This procedure can be applied to more unit root problems as well. However, it is restricted to a certain limited family of time series.

3.2 Control Charts with Reference Parameters for State-Space Models

Here we want to consider the problem of monitoring for more general non-stationary processes. We consider processes which can be described by a state-space process in the in-control state. This model class has been chosen because it is able to describe many types of non-stationary processes, including unit root processes. Moreover, recursive procedures are available for the statistical analysis of these processes, which makes them quite attractive in practice since the computational calculations can be done in a reasonable amount of time.

Lazariv and Schmid (2015) introduced several control charts for state-space models if a mean shift is present. Using the (generalized) likelihood ratio approach, the (generalized) sequential probability ratio test, and the (generalized) Shiryaev–Roberts procedure, they obtained control schemes with and without reference parameters. In an extensive simulation study, all these charts were compared with each other.

Here we extend their approach to the change point model (9). Replacing the matrix $\boldsymbol{D}_t = diag(\sqrt{Var(Y_{t1})}, \ldots \sqrt{Var(Y_{tp})})$ in Lazariv and Schmid (2015) by an arbitrary known matrix $\boldsymbol{D}_{t,\tau}$, $t \geq \tau$, it is possible to obtain control charts for further out-of-control situations, such as, e.g., drifts, drifts and shifts, etc.. This approach is briefly sketched in the following. In order to determine the likelihood function we need additional assumptions. It is demanded that (A1) and (A3) are fulfilled and that

(A2*) Let $\boldsymbol{S}_1$, $(\boldsymbol{V}_1', \boldsymbol{W}_1')'$, $(\boldsymbol{V}_2', \boldsymbol{W}_2')'$,... be independent.
(A4) Assume that $\boldsymbol{S}_1$, $(\boldsymbol{V}_1', \boldsymbol{W}_1')'$, $(\boldsymbol{V}_2', \boldsymbol{W}_2')'$,... are normally distributed.
(A5) Assume that $\boldsymbol{\Sigma}_t$ have full rank for all $t \geq 1$.

For more details we refer to Lazariv and Schmid (2015).

Let us rewrite the densities of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ in the in-control (f_0) and in the out-of-control (f_τ) states. Then it holds that

$$f_0(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = (2\pi)^{-np/2} \left(\prod_{t=1}^{n} \det \boldsymbol{\Sigma}_t \right)^{-1/2} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{n} (\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t)' \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t) \right\}, \tag{10}$$

where $\boldsymbol{\Sigma}_t = \boldsymbol{G}_t \boldsymbol{\Omega}_t \boldsymbol{G}_t' + \boldsymbol{R}_t$ stands for the error covariance matrix and $\hat{\boldsymbol{X}}_t$ is the best linear one-step predictor

$$\hat{\boldsymbol{X}}_t = \boldsymbol{b}_t(\boldsymbol{X}_0) + \sum_{j=1}^{t-1} \boldsymbol{B}_{t,j} \boldsymbol{X}_j \tag{11}$$

for $t \geq 1$.

According to the change-point model defined in (2) the likelihood function is given by

$$\begin{aligned} f_\tau(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) &= f_0(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\tau-1}, \boldsymbol{X}_\tau - \boldsymbol{D}_{\tau,\tau}\boldsymbol{a}, \ldots, \boldsymbol{X}_n - \boldsymbol{D}_{n,\tau}\boldsymbol{a}) \\ &= (2\pi)^{-np/2} \left(\prod_{t=1}^{n} \det \boldsymbol{\Sigma}_t \right)^{-1/2} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{n} (\boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_t)' \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_t) \right\}, \end{aligned} \tag{12}$$

where

$$\boldsymbol{Z}_t = \begin{cases} \boldsymbol{X}_t & \text{for} \quad 1 \leq t < \tau \\ \boldsymbol{X}_t - \boldsymbol{D}_{t,\tau}\boldsymbol{a} & \text{for} \quad \tau \leq t \leq n \end{cases},$$

and

$$\hat{\boldsymbol{Z}}_t = \boldsymbol{b}_t(\boldsymbol{Z}_0) + \sum_{j=1}^{t-1} \boldsymbol{B}_{t,j}\boldsymbol{Z}_j = \boldsymbol{b}_t(\boldsymbol{X}_0) + \sum_{j=1}^{t-1} \boldsymbol{B}_{t,j}\boldsymbol{X}_j - \sum_{j=\tau}^{t-1} \boldsymbol{B}_{t,j}\boldsymbol{D}_{j,\tau}\boldsymbol{a}$$

$$= \hat{\boldsymbol{X}}_t - \sum_{j=\tau}^{t-1} \boldsymbol{B}_{t,j}\boldsymbol{D}_{j,\tau}\boldsymbol{a} = \hat{\boldsymbol{X}}_t - \boldsymbol{G}_t \sum_{j=\tau}^{t-1} \boldsymbol{A}_{t,j}\boldsymbol{D}_{j,\tau}\boldsymbol{a} = \hat{\boldsymbol{X}}_t - \boldsymbol{G}_t\boldsymbol{H}_{t,\tau}\boldsymbol{a} \text{ for } t \geq 1$$

with

$$\boldsymbol{H}_{t,\tau} = \begin{cases} 0 & \text{for} \quad 1 \leq t \leq \tau \\ \sum_{j=\tau}^{t-1} \boldsymbol{A}_{t,j}\boldsymbol{D}_{j,\tau} & \text{for} \quad \tau < t \leq n \end{cases}. \tag{13}$$

Thus we get with $\boldsymbol{M}_{t,\tau} = \boldsymbol{G}_t\boldsymbol{H}_{t,\tau} - \boldsymbol{D}_{t,\tau}$ that

$$\boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_t = \begin{cases} \boldsymbol{X}_t - \hat{\boldsymbol{X}}_t & \text{for} \quad 1 \leq t < \tau \\ \boldsymbol{X}_t - \hat{\boldsymbol{X}}_t + \boldsymbol{M}_{t,\tau}\boldsymbol{a} & \text{for} \quad \tau \leq t \leq n \end{cases}.$$

3.2.1 The Likelihood Ratio Chart

The likelihood ratio (LR) approach is often used to derive control statistics for different types of target processes. Schmid (1997a) constructed a mean chart and Lazariv et al. (2013) a variance chart for a univariate stationary process using the LR method. The idea behind it is to consider for some fixed sample size n the testing problem that under H_0 the process is in-control ($\tau > n$) while under the alternative hypothesis a change occurs at time position τ ($1 \leq \tau \leq n$).

A detailed derivation of the control statistic is similar to that in Lazariv and Schmid (2015). Here only the final results are presented. The run length of the LR chart is given by

$$N_{LR}(c; \boldsymbol{a}^*) = \inf\{n \in \mathbb{N} : \max\{0, -g_{n;LR}(\boldsymbol{a}^*)\} > c\}. \tag{14}$$

where

$$g_{n;LR}(\boldsymbol{a}) = \min_{1 \leq i \leq n} \left(\sum_{t=i}^{n} \left((\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t + \frac{1}{2}\boldsymbol{M}_{t,i}\boldsymbol{a})' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{M}_{t,i}\boldsymbol{a} \right) \right).$$

Here, $\boldsymbol{a}^*$ denotes a reference value for the unknown shift $\boldsymbol{a}$.

3.2.2 The Sequential Probability Ratio Chart

The sequential probability ratio test (SPRT) was introduced by Wald (1947). It was used in Page (1954) to derive a mean chart for independent samples. Lazariv and Schmid (2015) derived an SPRT chart for a mean shift assuming the in-control process to be a state-space model. Following Lazariv and Schmid (2015), we get that the run length of the SPRT chart is equal to

$$N_{SPRT}(c, \boldsymbol{a}^*) = \inf\{n \in \mathbb{N} : \max_{0 \le i \le n} \{g_{n;SPRT}(\boldsymbol{a}^*) - g_{i;SPRT}(\boldsymbol{a}^*)\} > c\} \tag{15}$$

where

$$g_{n;SPRT}(\boldsymbol{a}) = -\sum_{t=1}^{n} (\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t + \tfrac{1}{2}\boldsymbol{M}_{t,1}\boldsymbol{a})' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{M}_{t,1}\boldsymbol{a}.$$

and $g_{0;SPRT} = 0$. As before, $\boldsymbol{a}^*$ is a reference value of the unknown parameter $\boldsymbol{a}$.

Note that the control statistic can be recursively calculated, which dramatically simplifies its determination.

3.2.3 The Shiryaev–Roberts Chart

In this section, we present a control chart based on the Shiryaev–Roberts (SR) procedure (Shiryaev 1963; Roberts 1966) for detecting changes in state-space models (see also Lazariv and Schmid 2015). Its run length is equal to

$$N_{SR}(c, \boldsymbol{a}^*) = \inf\{n \in \mathbb{N} : g_{n;SR}(\boldsymbol{a}^*) > c\} \tag{16}$$

with

$$g_{n;SR}(\boldsymbol{a}) = \sum_{\tau=1}^{n} \exp\left\{ -\sum_{t=\tau}^{n} (\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t + \tfrac{1}{2}\boldsymbol{M}_{t,\tau}\boldsymbol{a})' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{M}_{t,\tau}\boldsymbol{a} \right\}.$$

3.3 Control Charts without Reference Parameters for State-Space Processes

One of the main problems of the control charts with a reference or smoothing parameter concerns the prior choice of these quantities. The optimal choice depends on the unknown quantities of the out-of-control model, such as, e.g., the size of the shift. Frequently, such information is not available, so the choice of the reference value is sometimes like a lottery. For that reason, statements about the robustness of the charts with respect to the choice of the reference parameter are important.

Another possibility is the choice of the generalized likelihood function where the maximum over the unknown shift a is taken as well.

Consequently the quantity

$$\sup_{a \neq 0} \log f_\tau(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \longrightarrow \max,$$

is employed, where the likelihood is maximized over all possible sizes of the shift.

This is the idea behind the Generalized LR (GLR), the Generalized SPRT (GSPRT), and the Generalized Modified SR (GMSR) schemes. The details are presented below. The derivation of the charts follows using the same arguments as in Lazariv and Schmid (2015) where, however, the quantity $\boldsymbol{D}_t$ must be replaced by $\boldsymbol{D}_{t,\tau}$. For that reason we do not want to focus on the derivation of the results and we will directly give the final result.

3.3.1 The GLR Chart

The run length of GLR chart is given by

$$N_{GLR}(c) = \inf\{n \in \mathbb{N} : \max\{0, \max_{1 \leq i \leq n} \left(-\sum_{t=i}^{n} (\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t + \frac{1}{2}\boldsymbol{M}_{t,i}\tilde{\boldsymbol{a}}_{i,n})' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{M}_{t,i}\tilde{\boldsymbol{a}}_{i,n} \right) \} > c \}, \tag{17}$$

where $\tilde{\boldsymbol{a}}_{\tau,n}$ is the solution of the equation

$$\left(\sum_{t=\tau}^{n} \boldsymbol{M}'_{t,\tau} \boldsymbol{\Sigma}_t^{-1} \boldsymbol{M}_{t,\tau} \right) \tilde{\boldsymbol{a}}_{\tau,n} = \sum_{t=\tau}^{n} \boldsymbol{M}'_{t,\tau} \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t).$$

3.3.2 GSPRT Chart

In this case, the run length is obtained as

$$N_{GSPRT}(c) = \inf \left\{ n \in \mathbb{N} : \max_{0 \leq i \leq n} \left(g_{n;GSPRT} - g_{i;GSPRT} \right) > c \right\} \tag{18}$$

where

$$g_{n;GSPRT} = -\sum_{t=1}^{n} (\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t + \frac{1}{2}\boldsymbol{M}_{t,1}\tilde{\boldsymbol{a}}_n)' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{M}_{t,1}\tilde{\boldsymbol{a}}_n$$

and $\tilde{\boldsymbol{a}}_n = \tilde{\boldsymbol{a}}_{1,n}$.

3.3.3 GMSR Chart

The generalization of the SR chart leads to the problem that the maximum over some exponential sums must be calculated. In order to avoid this problem, we employ the sum over the individual likelihoods. This leads to

$$N_{GMSR}(c) = \inf\left\{n \in \mathbb{N} : \boldsymbol{a}_n^{*\prime}\ddot{\boldsymbol{S}}_n\boldsymbol{a}_n^* > c\right\} \tag{19}$$

where $\boldsymbol{a}_n^*$ is any solution of the equation $\ddot{\boldsymbol{S}}_n\boldsymbol{a} = -\dot{\boldsymbol{S}}_n$ and

$$\dot{\boldsymbol{S}}_n' = \sum_{i=1}^{n}\sum_{t=i}^{n}(\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t)'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{M}_{t,i}, \qquad \ddot{\boldsymbol{S}}_n = \sum_{i=1}^{n}\sum_{t=i}^{n}\boldsymbol{M}_{t,i}'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{M}_{t,i}.$$

4 Comparison Study

In this section we want to compare the above discussed control charts. We focus on a univariate in-control process. Here we present our results for a unit root process as defined in (1) with $y_0 = 0$ and $\{\varepsilon_t\}$ independent and standard normally distributed. The out-of-control process is given by the drift model (2).

4.1 Comparison Study Based on the Average Run Length

First, the average run length (ARL) is used as a performance measure. The in-control ARL is set equal to 500. The control limits for all charts were determined such that this calibration is fulfilled. After that the out-of-control ARLs of all charts are compared with each other. In our study, the reference value a^* takes values within the set $\{0.5, 1.0, \ldots, 3.0\}$. Moreover, an EWMA chart is applied to the first differences. The possible values of the smoothing parameter are taken from the set $\{0.1, 0.2, \ldots, 1.0\}$. Since there is no available explicit formula for the ARL, it is estimated by means of a simulation study based on 10^5 independent samples.

The results of our simulation study are given in the following table. The table shows the smallest out-of-control ARL over all reference values and smoothing parameters chart for a fixed drift size (Table 1).

The overall best scheme is the GSPRT scheme. It dominates all other schemes. Among the other charts the difference chart behaves the best for small drifts while for larger drifts the LR and the SPRT scheme dominate. The SR scheme is slightly worse than the LR and SPRT approaches, but a little bit better than the EWMA chart applied to the differences if the drift is large. It is interesting that the best LR

Table 1 ARLs for all control charts

a	LR	SPRT	SR	GLR	GSPRT	GMSR	EWMA
0.5	25.90(0.5)	25.76(0.5)	29.00(0.5)	36.34	17.77	64.71	24.31(0.1)
1.0	9.13(1.0)	9.15(1.0)	9.73(1.5)	11.90	6.26	33.36	8.86(0.2)
1.5	4.83(1.5)	4.84(1.5)	5.02(2.0)	6.21	3.44	22.48	4.80(0.3)
2.0	3.08(2.0)	3.06(2.0)	3.13(2.0)	4.01	2.30	16.92	3.13(0.4)
2.5	2.16(2.5)	2.15(2.5)	2.20(3.0)	2.80	1.72	13.55	2.23(0.6)
3.0	1.63(3.0)	1.64(3.0)	1.64(3.0)	2.15	1.39	11.30	1.68(0.7)

Table 2 Average delays of all control charts

a		LR	SPRT	SR	GLR	GSPRT	GMSR	EWMA
0.5	$\tau = 1$	25.90	25.84	29.00	36.34	17.77	64.71	24.31
	$\tau = 50$	22.03	21.95	21.18	31.59	31.08	42.77	22.82
1.5	$\tau = 1$	4.83	4.84	5.02	6.21	3.44	22.48	4.80
	$\tau = 50$	3.58	3.60	3.50	4.76	8.54	12.11	3.72
3.0	$\tau = 1$	1.63	1.64	1.64	2.15	1.39	11.30	1.68
	$\tau = 50$	1.62	1.62	1.62	2.08	3.58	5.48	1.68

and SPRT chart is the chart where the reference value is equal to the true drift size. The GLR chart behaves worse than the other schemes. However, the overall worst scheme is the GMSR chart, whose out-of-control ARL is much larger than those of the other charts.

4.2 Comparison Study Based on the Average Delay

The disadvantage of the ARL consists in the fact that the change is assumed to occur already at the first time point ($\tau = 1$). This is rarely the case in practice. Therefore the average detection delay (AD) is frequently used as an alternative performance criterion. The average delay is equal to the average number of observations from the shift at position τ to the time point of the signal. In Table 2 the ARL and the average delay for $\tau = 50$ are given for all considered charts.

In the literature, usually it is the limit of the average delay as $\tau \to \infty$ and the worst average delay over all τ that are taken as performance measures. A further analysis shows that except for the GSPRT chart, the worst average delay over $1 \leq \tau \leq 50$ is always already attained at $\tau = 1$, i.e. it is equal to the ARL. For these schemes the average delay is decreasing in τ. Thus we get the same ranking as for the ARL. The GSPRT chart behaves completely differently, since the average delay is increasing with τ and the results are worse. The chart seems to favour changes at the beginning but has problems detecting changes at later time points. If we consider the value of the average delay at $\tau = 50$ the SR scheme turns out to be the best. It is slightly better than the LR and the SPRT schemes, which are slightly

better than the EWMA approach. The table shows as well that CUSUM procedures have a better worst-case performance than the SR approach. This behaviour was already described by Yashchin (1993). The results for the generalized charts are worse. The best generalized procedure for small changes is the GSPRT approach while for medium changes the GLR chart is better than the GSPRT attempt. The GMSR chart behaves much worse.

4.3 Robustness Study with Respect to the Choice of the Reference Value

Up to now we have always considered for a fixed change the minimal ARL and the minimal average delay over all reference values and smoothing parameters. However, in most cases the practitioner does not know the true magnitude of the change. How good are the charts if instead of the best reference value and best smoothing parameter another value is taken? Here a robustness study is of importance. In Table 3 we give the worst average delay if a^* is chosen smaller (above) or larger (below) than the value leading to the minimum ARL. Note that in our analysis we have chosen $a^* \in \{0.5, 1.0, \ldots, 3.0\}$ and $\lambda \in \{0.1, 0.2, .., 1.0\}$. For example, the optimal choice of a^* for the LR chart is $a^* = 1.5$ if the expected shift is $a = 1.5$. In this case we get an average run length of 4.83. The direct neighbours of $a^* = 1.5$ are 1.0 and 2.0. If one chooses $a^* = 1.0$, the ARL is 5.15 (6.63%, above). For the choice $a^* = 2.0$, we obtain $ARL = 5.04$ (4.50%, below).

The table shows that the charts react differently to the choice of a^*. Nevertheless, the out-of-control ARLS are in all cases smaller than those of the GLR and the GMSR charts. Thus a small deviation from the optimal choice leads to acceptable results and there is no need to apply a generalized chart.

If, however, we consider the worst average run length over all possible values of a^* and for a fixed value of a, the results of the EWMA, LR, SPRT and SR charts are very bad. Assuming $a = 0.5$, the worst ARL for the SPRT (EWMA) chart is 87.54 (115.11); it is attained at $a^* = 3.0$ ($\lambda = 1.0$). For $a = 1.5$ we get 6.90 (11.86) for the SPRT (EWMA) scheme. These values are much worse than those of the GLR chart, which must be favoured in this case.

Table 3 Influence of the wrong choice of reference parameter, for all control charts

a	LR	SPRT	SR	GLR	GSPRT	GMSR	EWMA
0.5	25.90(0.5)	25.84(0.5)	29.00(0.5)	36.34	17.77	64.71	24.31(0.1)
	+18.26%	+19.27%	+1.19%	–	–	–	+22.86%
1.5	+6.63%	+5.98%	+1.29%	–	–	–	+1.13%
	4.83(1.5)	4.84(1.5)	5.02(2.0)	6.21	3.44	22.48	4.80(0.3)
	+4.50%	+4.37%	+11.84%	–	–	–	+3.93%

4.4 Conclusions

Summarizing the above results we can give the following recommendations. We do not recommend the use of the GMSR chart since the results are in general much worse than those of the other schemes. The reason may be that instead of maximizing the sum of the likelihoods we maximized the sum of the logarithms of the likelihoods. This may have led to the deterioration. Moreover, the GSPRT scheme must be carefully applied since it favours changes at the beginning and it has huge problems detecting a change at a later time point.

If some information about the magnitude of the change is known, then either the LR chart or the SPRT chart should be applied. If no information about the change is known, the GLR chart provides the best results.

5 Challenges and Problems

The monitoring of non-stationary processes is a challenging task and it has to be done carefully, since there are many hidden problems. Lazariv and Schmid (2015) showed that for some processes and change-point models the expectation of the run length does not exist. This is a very important issue since the ARL is the most popular measure for the performance of control charts. We want to address this problem in this chapter and check whether the same issue arises for the present change-point model (9).

For this purpose we have calculated a table of frequencies, namely, the frequencies with which the in-control run length will fall into certain intervals (see Table 4). The table shows the relative frequencies (in percentages) of $P(N(c) = i)$ for $i = 1, \ldots, 5$, $P(1000 \cdot i \leq N(c) < 1000 \cdot (i + 1))$ for $i = 1, \ldots, 5$ and $P(5000 \leq N(c) \leq 10{,}000)$. The results are based on simulating 10^5 independent random samples of a unit root process.

Table 4 Distributions of the in-control run lengths of the considered charts

i	LR	SPRT	SR	GLR
1	0.00	0.00	0.00	0.01
2	0.00	0.01	0.00	0.04
3	0.05	0.03	0.00	0.05
4	0.09	0.07	0.00	0.09
5	0.07	0.11	0.02	0.05
...	...	...	...	...
[1000, 2000]	11.15	11.56	11.54	11.14
[2000, 3000]	1.57	1.53	1.76	1.10
[3000, 4000]	0.21	0.20	0.20	0.09
[4000, 5000]	0.01	0.03	0.00	0.00
[5000, 10,000]	0.00	0.00	0.00	0.00

The table shows that there is no evidence of heavy tails. The Hill plot and the Pickands plot (see, e.g., Resnick 2007) are presented for the run length of the SPRT chart in order to analyse the tail behaviour and to check for the existence of the expectation of the run length. The run length is estimated using a simulation study using 10^5 repetitions. Figure 2 shows that the tail index is definitely larger than 1, which implies that the expectation of the run length exists.

Note that this result is different from the findings of Lazariv and Schmid (2015), where it was found that the average run lengths do not exist. How can this be explained? Of course in the present chapter another out-of-control case is studied in the comparison study and for that reason other control statistics are used. Nevertheless, this result is a little bit surprising.

A closer look at the structure of the control statistics shows that the matrix $\boldsymbol{M}_{t,\tau}$ greatly influences the control statistics of all the control charts ((14), (15), (16), (17), (18) and (19)). The problem is that for the change-point model in Lazariv and Schmid (2015), the matrix $\boldsymbol{M}_{t,\tau}$ tends to $\mathbf{0}$ as a function of t and for fixed τ because of $\boldsymbol{D}_t$. The quantity $\boldsymbol{D}_t$ models the variance of the target process in the univariate case and it seems to tend to infinity for a target process as in Lazariv and Schmid (2015). In this chapter, however, $\boldsymbol{D}_{t,\tau}$ depends on τ as well and it hence $\boldsymbol{M}_{t,\tau} = -1$, i.e. it is constant.

6 Summary

In the present chapter we discussed different schemes for monitoring non-stationary processes. We considered the transformation method, where the original data are suitably transformed to a stationary process, e.g., by detrending or differencing. Then all well-known control charts for stationary processes can be applied to the transformed quantities. The problem with this procedure is that it only works for special type of processes, such as, e.g., unit root processes. Another approach (see, e.g., Lazariv and Schmid 2015) is to use the probability structure of the underlying process to derive the control charts. Here, the in-control process is assumed to be a multivariate state-space process. The considered change point is quite general, including drifts and shifts in the components. Using the likelihood ratio, the sequential probability ratio, and the Shiryaev–Roberts procedure, control charts with a reference vector have been derived. Applying the generalized likelihood ratio, the generalized sequential probability ratio, and the generalized modified Shiryaev–Roberts procedure, control schemes without reference values have been obtained.

All the charts have been compared with each other, assuming that the in-control process is a unit root process and that a linear drift in the process may occur. Different performance criteria have been used to evaluate the introduced charts. The average run length, the worst average delay and the limit of the average delay have been considered. Moreover, it has been analysed how the charts with a reference value react if the optimal reference value leading to the smallest ARL is not used, but another value, that is either close to the optimal one or further away. It has been

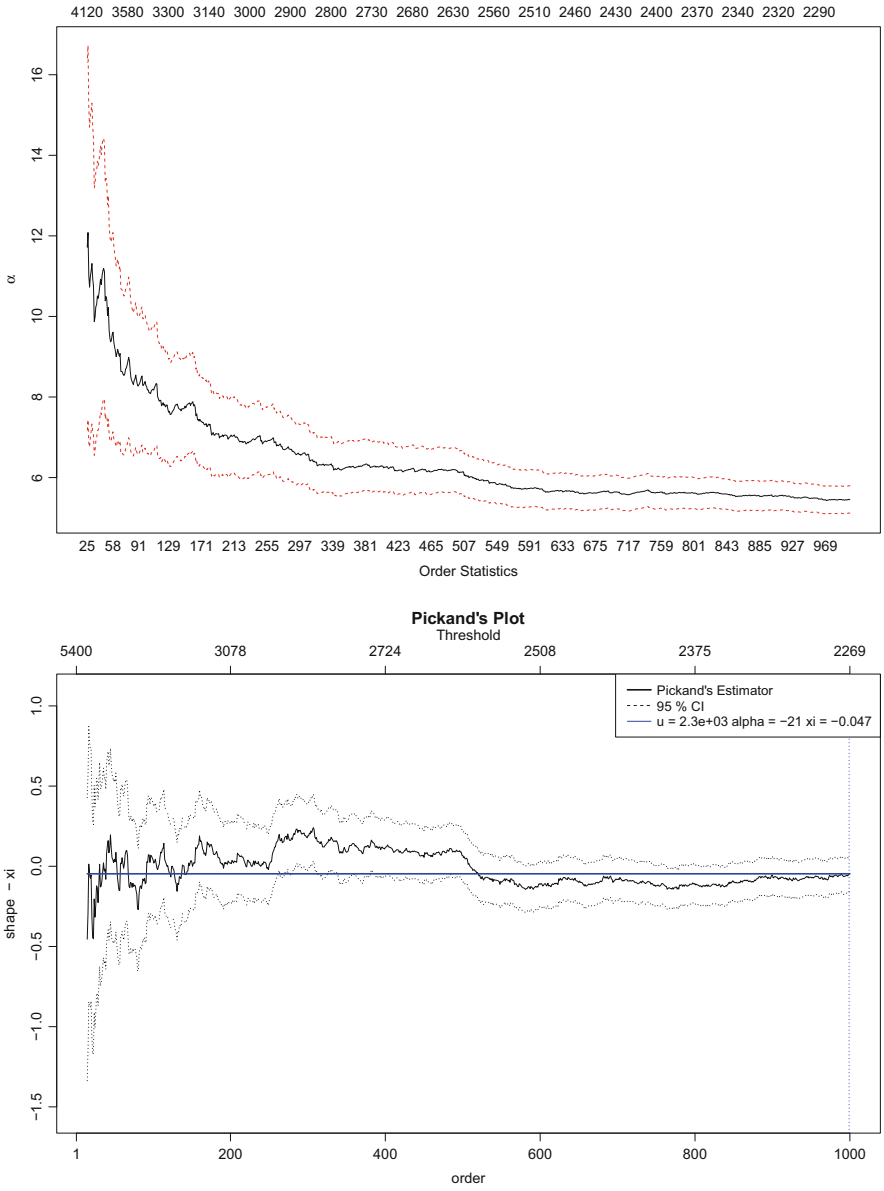


Fig. 2 Hill plot (above) and Pickands plot (below)

shown that the LR and the SPRT charts should be favoured if some knowledge of the expected drift is available. However, if no information about the drift is given, the GLR chart provides the best results.

In Lazariv and Schmid (2015), it was shown that the average run length of the introduced charts does not exist. Our approach is a generalization of that of Lazariv and Schmid (2015). Using the Hill plot and the Pickands plot, the tails of the run lengths of the introduced charts have been analysed and it has been concluded that in the present case the average run length exists. The reason for the different behaviour lies in the consideration of a different out-of-control model.

References

Alwan, L. C., & Roberts, H. V. (1988). Time-series modeling for statistical process control. *Journal of Business & Economic Statistics, 6*(1), 87–95.

Brockwell, P. J., & Davis, R. A. (2009). *Time Series: Theory and Methods*. Berlin: Springer.

Chou, C. J. (2004). Groundwater monitoring: Statistical methods for testing special background conditions. In: G. B. Wiersma (Ed.), *Environmental Monitoring*. Boca Raton, FL: CRC Press.

Durbin, J., & Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Frisén, M. (2008). *Financial Surveillance* (Vol. 71). Hoboken, NJ: John Wiley & Sons.

Hayashi, F. (2000). *Econometrics*. Princeton, NJ: Princeton University Press.

Kass-Hout, T., & Zhang, X. (2010). *Biosurveillance: Methods and Case Studies*. Boca Raton, FL: CRC Press.

Koopman, S. J. (1997). Exact initial Kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Association, 92*(440), 1630–1638.

Lazariv, T., & Schmid, W. (2015). Surveillance of non-stationary processes Discussion Paper

Lazariv, T., Schmid, W., & Zabolotska, S. (2013). On control charts for monitoring the variance of a time series. *Journal of Statistical Planning and Inference, 143*(9), 1512–1526.

Lu, C. W., & Reynolds, M. (1999). Control charts for monitoring the mean and variance of autocorrelated processes. *Journal of Quality Technology, 31*(3), 259–274

Montgomery, D. C. (2009). *Introduction to Statistical Quality Control* (6th ed.). Hoboken, NJ: John Wiley & Sons.

Nikiforov, I. (1975). Sequential analysis applied to autoregression processes. *Automation and Remote Control, 36*, 1365–1368

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

Resnick, S. I. (2007). *Extreme Values, Regular Variation, and Point Processes*. Berlin: Springer.

Roberts, S. (1966). A comparison of some control chart procedures. *Technometrics, 8*(3), 411–430.

Ruppert, D. (2004). *Statistics and Finance: An Introduction*. Berlin: Springer.

Schmid, W. (1995). On the run length of a Shewhart chart for correlated data. *Statistical Papers, 36*(1), 111–130.

Schmid, W. (1997a). CUSUM control schemes for Gaussian processes. *Statistical Papers, 38*(2), 191–217

Schmid, W. (1997b). On EWMA charts for time series. In *Frontiers in Statistical Quality Control* (Vol. 5, pp. 115–137). Berlin: Springer.

Schmid, W., & Steland, A. (2000). Sequential control of non-stationary processes by nonparametric kernel control charts. *Allgemeines Statistisches Archiv (Journal of the German Statistical Association), 84*, 315–336.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability & Its Applications, 8*(1), 22–46.

Steland, A. (2002). Nonparametric monitoring of financial time series by jump-preserving estimators. *Statistical Papers, 43*, 361–377

Steland, A. (2005). Random walks with drift – a sequential approach. *Journal of Time Series Analysis, 26*(6), 917–942.

Steland, A. (2007). Monitoring procedures to detect unit roots and stationarity. *Econometric Theory, 23*(06), 1108–1135.

Steland, A. (2010). A surveillance procedure for random walks based on local linear estimation. *Journal of Nonparametric Statistics, 22*(3), 345–361.

Triantafyllopoulos, K., & Bersimis, S. (2016). Phase II control charts for autocorrelated processes. *Quality Technology & Quantitative Management, 13*(1), 88–108.

Wald, A. (1947). *Sequential Analysis*. New York: Wiley.

Wardell, D. G., Moskowitz, H., & Plante, R. D. (1994a) Run-length distributions of residual control charts for autocorrelated processes. *Journal of Quality Technology, 26*(4), 308–317.

Wardell, D. G., Moskowitz, H., & Plante, R. D. (1994b). Run-length distributions of special-cause control charts for correlated processes. *Technometrics, 36*(1), 3–17.

Yashchin, E. (1993). Performance of CUSUM control schemes for serially correlated observations. *Technometrics, 35*(1), 37–52.

Part II
Design of Experiments

Design of Experiments: A Key to Successful Innovation

Douglas C. Montgomery and Rachel T. Silvestrini

Abstract An important theme in this chapter is that the use of statistical methodology, such as design of experiments, can aid innovation. Design of experiments is viewed as part of a process for enabling both breakthrough innovation and incremental innovation, without which Western society will fail to be competitive. Quality engineering technology in general is part of a broader approach to innovation and business improvement called statistical engineering. The most powerful statistical technique in statistical engineering is design of experiments. Several important developments in this field are reviewed, the role of designed experiments in innovation examined, and new developments and applications of the methods discussed.

Keywords Quality engineering · The scientific method · Optimal design · Computer experiments

1 Introduction

In June 2007 (http://www.bloomberg.com/news/articles/2007-06-10/at-3m-a-struggle-between-efficiency-and-creativity) Brian Hindo wrote an article in Bloomberg News entitled "At 3M, A struggle Between Efficiency and Creativity." The article strongly suggests that programs such as Six Sigma and Total Quality Management (TQM) stifle innovation if they become engrained within a company's

D. C. Montgomery
School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA
e-mail: doug.montgomery@asu.edu

R. T. Silvestrini (✉)
Department of Industrial and Systems Engineering, Rochester Institute of Technology, Rochester, NY, USA
e-mail: rtseie@rit.edu

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_15

culture. Hindo writes "Efficiency programs such as Six Sigma are designed to identify problems in work processes... When these types of initiatives become ingrained in a company's culture, as they did at 3M, creativity can easily get squelched. After all, a breakthrough innovation is something that challenges existing procedures and norms." In the article, this opinion seems to be shared with several other CEOs as well as a number of business school professors based on quotations throughout his piece. While Hindo presents many good points about invention and innovation needing room for unstructured discovery, we believe that programs such as Six Sigma and TQM, with toolboxes that include Design of Experiment, can still coexist with creativity, innovation, and invention.

In this chapter we explore the larger context of whether or not the use of statistically methodologies stifle innovation. Spoiler alert, we do not think that these methodologies suppress innovation. On the contrary, we illustrate their place and appropriate use and illustrate examples of success. Statisticians view one such statistical method, design of experiments, as part of a process for enabling both breakthrough and incremental innovation. Montgomery and Woodall (2008) provide an overview of the statistical methods, including design of experiments, used within Six Sigma, and the impact of Six Sigma in practice.

There are some notable authors who discuss the use of experimentation and the role experimentation plays in innovation. In his book entitled, "Experimentation Matters," Thomke (2003a) argues that "experimentation fuels the discovery and creation of knowledge and thereby leads to the development and improvement of products, processes, system, and organizations." Another article, appearing in the same year, Thomke (2003b) further adds that, "for hundreds, if not thousands, of years, systematic experimentation has been at the heart of all innovation." Bisgaard (2012) discusses how specific methodologies such as Design for Six Sigma (DFSS) can be applied to achieve quality via product innovation, which in turn should provide enhanced value to customers. Statistical programs such as Six Sigma, DFSS and TQM, as well as the formal process of design of experiments, generally fall under the Quality Engineering realm.

Quality engineering technology in general is part of a broader approach to innovation and business improvement called statistical engineering. Readers should reference Hoerl and Snee, who present two chapters (2010a, 2010b), which discuss aspects of statistical engineering and how best to use statistical methods for improved results. Also see the chapters by Anderson-Cook et al. (2012a,b), Box and Woodall (2012) and Hockman and Jensen (2016). Antony et al. (2011) discuss and illustrate how designed experiments can promote innovative solutions to complex problems in non-manufacturing and service organizations.

We believe that the most powerful statistical technique in statistical engineering is design of experiments. In this chapter we explore what innovation is, how it is different from invention, and its place within research and development. We also discuss design of experiments and its relationship with the scientific method. Finally, we present important developments in this field of experimental design, the role of designed experiments in innovation, and applications of the methods illustrated.

2 Innovation and Invention

Innovation is the successful exploitation of new ideas for products, services or processes. This includes both radical new ideas (breakthrough innovation) and changes to existing ones (incremental innovation). Successful innovation is a key factor in higher and more sustainable profitability, staying ahead of your competition and providing higher value to customers. Thus, all businesses should innovate in order to thrive. Innovation offers a way of meeting challenges both inside and outside a business and allows businesses to compete effectively in the increasingly competitive global environment.

What is the difference between innovation and invention? The two are listed as synonyms of each other. An invention is described as a unique or novel device or discovery. Like innovation this can be in the form of a breakthrough or built on a pre-existing idea. An invention that is not derived from an existing model or idea, or that achieves a completely unique function, discovery, or result, may be a breakthrough. An invention may also be an improvement upon something that already exists. The difference between innovation and invention is subtle. A 2015 Wired article, entitled "Innovation vs. Invention: Make the Leap and Reap the Rewards," by Bill Walker discusses these subtleties (http://www.wired.com/insights/2015/01/innovation-vs-invention/). Walker emphasizes that innovation deals with the concepts of *use* while invention pertains to a *thing*.

In his article on efficiency and creativity, Hindo cites three examples of innovation within 3M in his 2007 article: masking tape, Thinsulate, and the Post-it note. We believe these are both inventions and innovations. All of these products provided a fundamentally new product—a thing—to the market and fulfilled an unmet need—a use. All three of these products can be classified in the breakthrough category. Interestingly, Post-it notes were an innovative idea founded on a failed invention. Dr. Spencer Silver is credited with the development of the adhesive chemical used in Post-it notes, however it was Art Fry, a colleague of Silver's, who came up with the idea of using the product in the post-it style. Originally, Dr. Silver was trying to develop a super-strong adhesive product, but accidentally created a reusable light adhesive product.

Forbes regularly publishes a list of the "Most Innovative Companies." Among that list in the past 10 years include companies such as Apple, Google, 3M, Toyota, Microsoft, GE, P&G, Nokia, Starbucks, and IBM. In 2015, Tesla ranked number 1 (http://www.forbes.com/innovative-companies/list/#tab:rank). A brief survey of the list reveals a list of companies that are both innovative and inventive and thus have an edge in the market. Many of these companies have strong, well-known activities that embrace statistics and statistical engineering, including the use of designed experiments. A large portion of the innovation and invention activities in many organizations takes place within Research and Development (R&D).

Type *research and development* into your web browser and the first thing that pops up is a definition. The definition is "(in industry) work directed towards the innovation, introduction, and improvement of produces and processes." While

R&D is listed as an umbrella term, we feel that it is important to distinguish the two. Research is the area of a company that is directed to take risks and allow failures. Surprises are both rewarded and celebrated, especially when they result in a novel discovery. In contrast, Development would like no surprises as they can lead to catastrophic failure. The customer of the development department is generally manufacturing or the fulfillment process, where consistency and lack of variability are key quality metrics.

Breakthrough innovation and invention within a company often occurs within the research team. Incremental innovation is more typically found in development organizations. The R&D sector within a company has a long history of relying on the scientific method to aid in discovery. In the next section of this chapter we will discuss the scientific method and its relationship with design of experiments.

3 The Scientific Method and Design of Experiments

Scientists and engineers solve problems of interest to society by the efficient application of scientific principles. This is usually accomplished by either refining an existing product or process, or by designing a new product or process that meets customers' needs. The scientific (or engineering) method is the approach typically used in formulating and solving these problems. Montgomery and Runger (2014) identify the steps in the scientific (or engineering) method as follows:

1. Develop a clear and concise description of the problem.
2. Identify, at least tentatively, the important factors that affect this problem or that may play a role in its solution.
3. Propose a model for the problem, using scientific or engineering knowledge of the phenomenon being studied. This model may be a theory or hypothesis about how the phenomena of interest behaves. State any limitations or assumptions of the model.
4. Conduct appropriate experiments and collect data to test or validate the tentative model or conclusions made in steps 2 and 3.
5. Refine the model on the basis of the observed data.
6. Manipulate the model to assist in developing a solution to the problem.
7. Conduct an appropriate experiment to confirm that the proposed solution to the problem is both effective and efficient.
8. Draw conclusions or make recommendations based on the problem solution.

The steps in the scientific method are shown in Fig. 1. Many of the fields of science are employed in the scientific method: the physics and the mechanical sciences (statics, dynamics), fluid science, thermal science, electrical science, the science of materials, chemistry, biochemistry and biological sciences. Notice that the scientific method features a strong interplay between the problem, the factors that may influence its solution, a model of the phenomenon, and experimentation to verify the adequacy of the model and the proposed solution to the problem.

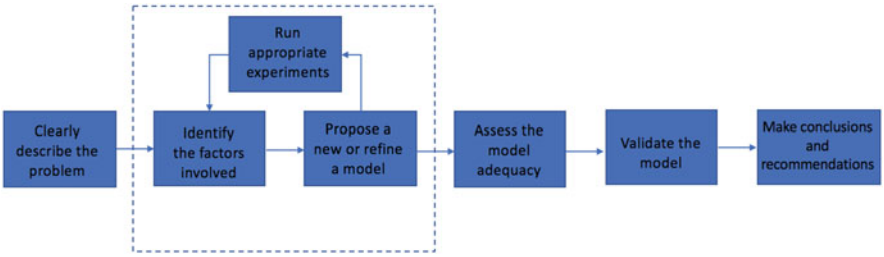


Fig. 1 The scientific (or engineering) method adapted from Montgomery and Runger (2014)

Steps 2–4 in Fig. 1 are enclosed in a box, indicating that several cycles or iterations of these steps may be required to obtain an appropriate solution. The boxed steps should not suggest that these steps can be bypassed, in fact, they are critical to the process and as we argue, necessary for innovation.

Scientists and engineers must know how to efficiently plan experiments, collect data, analyze and interpret the data, and understand how the observed data are related to the model they have proposed for the problem under study. An experiment is a test or series of tests in which purposeful changes are made to the input variables of a process or system so that we may observe and identify the reasons for changes that may be observed in the output response. We usually want to determine which input variables are responsible for the observed changes in the response, develop or refine a model relating the response to the important input variables and to use this model for process or system improvement or other decision-making. In R&D activities we are often trying to discover how some system behaves or performs, or to validate a theory about how the system should perform.

There are at least three distinct strategies of experimentation. In the best-guess approach the experimenter makes an educated guess based on his or her experience and scientific/engineering knowledge about the phenomena being studied behaves. Based on the outcome of this experiment, another experiment or series of experiments is planned and conducted. This process is continued until either (1) success is achieved, (2) no further guesses about the problem are forthcoming so testing is halted, or (3) the organization abandons the effort. Best-guess experimentation is sometimes very successful, but it can take a long time and there is no assurance that any solution found is the best one. This approach is one in which the experimenter decides appropriate levels of factors to tests at and makes adjustments based on subject matter knowledge or trial in error. In this method, sometimes factors are varied simultaneously, other times, only one factor is varied at a time. The range of experimentation may stay relatively consistent, or be wildly different. While using the best-guess approach may be successful regarding a single solution, it usually results in less knowledge about the system as a whole.

The one-factor-at-a-time or OFAT strategy is very popular in some fields. In this approach, a list of potential factors to be studied is constructed and then experiments are performed in which all factors but one are held constant at some reference or baseline level while one factor is varied over its range. This is repeated until all

factors have been varied over their range while simultaneously holding all others constant. Then a decision is made about the problem by examining the one-factor-at-a-time results. This decision is often a pick-the-winner process, where the best combination of factors is read from a series of plots. The well-known disadvantage of this approach is that any interaction between the factors will not be discovered. Interactions occur relatively often and in many cases they are the key to problem solution.

Statistically designed experiments are the recommended strategy. Usually these are experiments based on the idea of factorial designs (see Montgomery 2012). This approach varies factors together which among other things facilitates the discovery of interaction effects. The famous statistician George Box was often quoted as saying that if "... scientists and engineers only knew about the simplest factorial design (the 2^k) and only knew how to visually examine the data this would have a huge impact on innovation and competitive position in this country" (see Box 1990).

Some argue that successful invention and innovation requires creativity and original thinking and that the use of formal statistical methods like designed experiments stifles or retards the creative 'trial and error' process. We think of designed experiments as an efficient and well-organized approach to trial-and-error experiments. Perhaps a key difference is that a sound approach to designed experiments is that a pre-experimental planning activity is highly recommended. Refer to Coleman and Montgomery (1993) and Chapter 1 of Montgomery (2012), including the supplemental material for that chapter and the additional references therein. Charles Hicks, a famous professor of statistics and mathematics at Purdue University, is said to have told his design of experiments students that "... if you have 10 weeks to solve a problem, you should spend 8 weeks planning the experiment, one week running it, and one week analyzing the data." It is important to remember that all experiments are designed experiments. Ones that are poorly planned and executed will usually deliver disappointing results, while careful pre-experimental planning and execution of an experiment will usually produce results helpful and even essential in eventual problem solution.

It is often a capital mistake in invention and innovation activities to over-rely on theory. An example of this occurred early in the history of powered flight. In the early part of the twentieth century Samuel P. Langley was the most famous authority in aerodynamics of his era. He was sponsored by the US government to develop a *flying machine*. Langley built an airplane based entirely on his understanding of theoretical aerodynamic principles. At the same time, Orville and Wilbur Wright, two bicycle mechanics from Dayton, Ohio, were building an airplane based on their experimental work. They developed a working knowledge of aerodynamics from a home-built wind tunnel in which they conducted numerous experiments. Their experiments included flying kites and eventually gliders at Kitty Hawk, North Carolina. They also developed control systems for the airplane based on the wing-warping technique and developed a propulsion system experimentally. This work took place over a period of several years. Langley tested his airplane by launching it from a ramp. It fell into the Potomac River and never flew. The Wright brothers were highly successful, becoming the founding fathers of modern aviation. Good

pre-experimental planning which brings in a variety of backgrounds, viewpoints, and experiences is often effective in avoiding over-reliance on theory.

4 The Role of Design of Experiments in Innovation

As noted in the introduction, Hindo and others think that statistical methodologies can stifle innovation. In fact, many people believe that any specific framework, for example, Design Thinking, may suppress creative thinking or in general the creative thinking process. We are proponents of using appropriate toolsets when warranted. For example, control charting, and specifically a Shewhart chart, cannot be used until there is a process in place in which measurement may be taken and thus sampling and charting can be applied.

Misguided use of methodologies and a lack of understanding of toolsets can lead to failure or lack of success. It is wrong for a manager to say, "Use design of experiments to innovate me a new product." Design of experiments will not produce results; people will produce results. It is more appropriate to understand that design of experiments can provide a very effective and efficient aid that leads to innovation and invention. Hindo argues that "defenders of Six Sigma at 3M claim that a more systematic new-product introduction process allows innovations to get to market faster." Six sigma is about reducing variability in key product quality characteristics, not a tool to create a new-product. See Montgomery (1992, 1999) for a more thorough discussion of statistical process control and the role of experimental design within process control.

So, when should Design of Experiments be applied for innovation? The process can be used when an idea has been formed regarding use or development of a *thing*. Noting back to flight testing, a notion of an airplane and flight was developed. Determining the notion of flight leads way into the first step of design of experiments "statement or recognition of a problem." Prior to figuring out what this statement is, the design of experiments framework cannot begin. Once the statement is formed, or the problem is recognized, then design of experiment may be applied.

Based on the notion of 'creating a vehicle that can fly,' it was important to determine *how* to fly and what factors might influence flight. In order to determine the how and why, it is important to conduct experiments. Whether it is a small or large number of tests or trials, design of experiments can be extremely effective for determining what to test, where to test, and how much to test.

5 Barriers Hindering the Use of Design of Experiments

We believe that designed experiments should be much more widely used in invention and innovation activities. As alluded to earlier in the quote attributed to Box, even the use of simple techniques such as 2^k factorial designs, has the potential

to greatly spark innovation and research and development productivity. So, why aren't the basic design of experiments concepts and techniques more widely used? We think there are several barriers that hinder the more widespread use of design experiments and probably statistical methods in general.

Resistance to change is certainly an issue. Many scientists and engineers were educated in an environment where the OFAT approach was used in their university laboratory courses. In many cases it's not just the scientists and engineers, but often the managers and executives responsible for R&D that have this experience in their background. This can make it difficult to effectively integrate designed experiments as a standard part of R&D activities. Furthermore, many individuals may view the use of designed experiments as more time-consuming and difficult that the traditional approach such as an OFAT.

Prior negative experiences with statistical methods including designed experiments may also be a factor. Prior experiments may not have been successful because appropriate design and analysis techniques were not used. For example, one of us was engaged as a consultant by a company to provide some training on design of experiments to their R&D organization. It turned out that there had been a previous round of training by a consultant who had focused exclusively on Taguchi methods. However, most of the experiments actually conducted in this organization were mixture experiments and the scientists and engineers quickly became disillusioned with deigned experiments when they were unable to see how to use the L18 and L27 orthogonal array for the kind of problems they encountered. There was a lot of negative energy to overcome to convince them that there were appropriate techniques that would be useful to them.

Sometimes a failed experiment could be the result of poor pre-experimental planning. As noted in Coleman and Montgomery (1993) and Montgomery (2012), good experimental design is almost always a team effort. Letting one person design the experiment is almost always a mistake, especially if that person is an expert in the field. This often results in a situation where the expert already *knows the answer* and as a result designs an experiment to prove his or her conjecture. This can lead to an experiment that is too narrow in scope and that produces disappointing results.

Sometimes scientists and engineers have a weak statistical background that inhibits their understanding and use of designed experiments. Sadly, many scientific and engineering disciplines don't recognize the value of statistics and require very minimal (if any) university education in the field. Equally sadly, university courses are sometimes poorly taught. Often the statistics course for engineers and scientists is a service course and assigned to someone with little interest in how the subject matter could actually be used by the students. Sometimes the course disintegrates into a semester-long exposition of balls and urns and almost nothing that illustrates the power and beauty of using statistical methods to solve real problems is actually covered. Sometimes even a full course in design of experiments is not taught well. Many faculty members lack practical experience with designed experiments and don't have full appreciation of its use in an R&D environment. They do not present real and meaningful examples and case studies in class. Furthermore, students are not encouraged to conduct a real experiment as a course term project requirement.

Finally, many university design of experiments courses really don't focus enough on design, with too much course content devoted to analysis. Integration of computer software into the course could change that emphasis.

Over-reliance on knowledge of underlying theory is another all-to-common problem; team leadership believes that the project can be addressed by relying on *first principles*. So the product or system design is carried out using a purely theoretical modeling and analysis approach. Utilizing one's knowledge of the underlying theory is an integral part of the successful use of the scientific method but it needs to be integrated into a well-thought-out approach to research and development that also makes use of sound experimental strategy at important steps along the way. The first principles approach often leads to viewing experimentation as confirmation only, and testing comes too late in the development cycle to take advantage of the discovery and exploration aspects of good experimental strategy. The story of Samuel Langley and the Wright brothers discussed previously is an excellent example of how things can go wrong when we rely too much on first principles.

6 Recent Developments in Design of Experiments

There have been several developments in recent years in the design of experiments field that have great potential to enhance innovation and drive more efficient product and process development. Here we mention only a few of these.

The first of these is new design methodology that can reduce the amount of experimentation, reduce resources required for testing, and reduce development time. The use of non-regular fractional factorial designs can be very useful in this regard. These are designs in which many of the factorial effects are not completely aliased. Jones and Montgomery (2010) identify a class of designs for 6–8 two-level factors in 16 runs that do not alias any main effects with two-factor interactions and no two-factor interactions are completely aliased with each other (although they are correlated). These designs are good alternatives to the usual resolution IV fractions in which the two-factor interactions are completely aliased. If there are significant two-factor interactions the usual resolution IV designs would require follow-on experimentation to identify which two-factor interactions are active. Unless there are many two-factor interactions these non-regular designs allow experimenters to identify important main effects and two-factor interactions without additional experimentation. The ability to isolate both main effects and two-factor interactions from a single relatively small experiment has the potential to greatly accelerate the development cycle. Shinde et al. (2014) explore the projection properties of these designs and provide some insight on potential analysis methods. Krishnamoorthy et al. (2015) demonstrate how one modern regression technique, the Dantzig selector, can be used to analyze these designs. In a subsequent chapter Jones et al. (2015a) present 16-run designs for 9–14 two-level factors that do not completely alias any main effects with two-factor interactions and no two-

factor interactions are completely aliased with each other, although these effects are correlated. These designs can be thought of as alternative to the regular resolution III 16-run fractions.

The definitive screening designs developed by Jones and Nachtsheim (2011) are three-level designs that require only one more run than twice the number of factors. These designs are small enough to allow efficient screening of potentially many factors yet they can accommodate many second-order effects without additional runs. These designs have the following desirable properties:

1. The number of required runs is only one more than twice the number of factors. Consequently, these are very small designs.
2. Unlike resolution III designs, main effects are completely independent of two-factor interactions. As a result, estimates of main effects are not biased by the presence of active two-factor interactions, regardless of whether the interactions are included in the model.
3. Unlike resolution IV designs, two-factor interactions are not completely aliased with other two-factor interactions, although they may be correlated.
4. Unlike resolution III, IV and V designs with added center points, all quadratic effects can be estimated in models comprised of any number of linear and quadratic main effect terms.
5. Quadratic effects are orthogonal to main effects and not completely aliased (although they are correlated) with interaction effects.
6. With six or more factors, the designs are capable of estimating all possible full quadratic models involving three or fewer factors with very high levels of statistical efficiency.

These designs are an excellent compromise between Resolution III fractions for screening and small RSM designs. They also admit the possibility of moving directly from screening to optimization using the results of a single experiment. Jones and Nachtsheim found these designs using an optimization technique they had previously developed for finding minimum aliasing designs. This procedure minimizes the sum of squares of the elements of the alias matrix subject to a constraint on the *D*-efficiency of the resulting design. These designs can also be constructed directly from conference matrices.

Griffin et al. (2012) discuss the extensive use of simulation models as an aid to innovation. They state, “serial innovators typically have these types of hard data [powerful data supported by evidence] because they run detailed experiments testing their models of how things work.” Both physical and computer simulations can utilize experimental design methods. Experimental designs for deterministic computer models is another relatively new area of application that has great potential to accelerate innovation. Many engineering design activities make use of these types of models which include finite element models, computational fluid dynamics models, computational thermodynamic models, environmental models, and electrical circuit and device design software. Some of these models have many variables that must be studied and they can have very long execution times even on very fast computers. A widely used way to use these models is to deploy an

experimental design on the computer model and then fit a response surface of some type as a meta-model to the resulting output. Standard experimental design techniques such as factorial designs and response surface designs often do not work well in these applications because the low-order models that these designs support don't usually lead to an approximating meta-model that fits the response surface with the desired accuracy.

The approach that is widely used in practice is to use a space-filling design and fit the meta-model using the Gaussian process model. Jones and Johnson (2009) give an introduction and overview of these methods. Other useful references on space-filling designs and associated modeling techniques include Johnson et al. (2011), Silvestrini et al. (2013), and Jones et al. (2015b). Space-filling designs are not recommended for use in modeling response surfaces with low-order polynomials because of undesirable prediction variance properties, see Johnson et al. (2010).

7 Conclusions

It is our view that design of experiments is the most statistical powerful tool that is useful in enhancing both breakthrough and incremental innovation. Yet it is not as widely used as it could be. Based on research of 3M practice, Hindo discusses that "for a long time, 3M had allowed researchers to spend years testing products." Design of experiments could greatly improve the testing process and Six Sigma practice can be used to reduce noise when the product is formed and being produced. Making statements that a culture of quality stifles activities such as testing seems to be a misunderstanding of toolsets. Aside from this misunderstanding, we have identified four main reasons for barriers to design of experiments, but which can be thought of as barriers to any formal statistical toolset:

1. Resistance to change
2. Prior negative experiences with statistical methods
3. Lack of statistical knowledge of key personnel in the organization
4. Over-reliance on underlying theory or a *first-principles* approach

Design of experiments provides a structured methodology for experimentation and this can greatly aid in creative thinking. This structured methodology can improve creative thinking in many instances because it allows one the ability to iterate through ideas in a very efficient manner. There is always struggle with regards to innovation and invention. The struggle cannot and should not be removed. Creating the starting point, that leads the way to the use of designed experiments takes time and energy, but will be very rewarding. Applying statistical methodology is an important aid in the innovative process and should be employed for improved results.

References

Anderson-Cook, C. M., Lu, L., Clark, G., DeHart, S. P., Hoerl, R., Jones, B., et al. (2012a). Statistical engineering—Forming the foundations. *Quality Engineering, 24*(2), 110–132.

Anderson-Cook, C. M., Lu, L., Clark, G., DeHart, S. P., Hoerl, R., Jones, B., et al. (2012b). Statistical engineering—Roles for statisticians and the path forward. *Quality Engineering, 24*(2), 133–152.

Antony, J., Coleman, S., Montgomery, D. C., Anderson, M. J., & Silverstrini, R. T. (2011). Design of experiments for non-manufacturing processes: Benefits, challenges and some examples. *Journal of Engineering Manufacture, 225*(11), 2088–2095.

Bisgaard, S. (2012). The future of quality technology: From a manufacturing to a knowledge economy & from defects to innovations. *Quality Engineering, 24*(1), 30–36.

Box, G. E. P. (1990). Do interactions matter? Center for Quality and Productivity Improvement, University of Wisconsin, Madison.

Box, G. E. P., & Woodall, W. H. (2012). Innovation, quality engineering, and statistics. *Quality Engineering, 24*(1), 20–29.

Coleman, D. E., & Montgomery, D. C. (1993). A systematic approach to planning for a designed industrial experiment (with discussion). *Technometrics, 35*(1), 1–27.

Griffin, A., Price, R., & Vojak, B. (2012). *Serial innovators: How individuals create and deliver breakthrough innovations in mature firms*. Stanford, CA: Stanford University Press.

Hockman, K. K., & Jensen, W. A. (2016). Statisticians as innovation leaders. *Quality Engineering, 28*(2), 165–174.

Hoerl, R. W., & Snee, R. D. (2010a). Moving the statistics profession forward to the next level. *The American Statistician, 64*(1), 10–14.

Hoerl, R. W., & Snee, R. D. (2010b). Closing the gap: Statistical engineering can bridge statistical thinking with methods and tools. *Quality Progress, 43*(5), 52–53.

Johnson, R. T., Montgomery, D. C., & Jones, B. (2011). An empirical study of the prediction performance of space-filling designs. *International Journal of Experimental Design and Process Optimization, 2*, 1–18.

Johnson, R. T., Montgomery, D. C., Jones, B., & Parker, P. A. (2010). Comparing computer experiments for fitting high-order polynomial models. *Journal of Quality Technology, 42*(1), 86–102.

Jones, B., & Johnson, R. T. (2009). Design and analysis for the gaussian process model. *Quality and Reliability Engineering International, 25*, 515–524.

Jones, B., & Montgomery, D. C. (2010). Alternatives to resolution IV screening designs in 16 runs. *International Journal of Experimental Design and Process Optimization, 1*(4), 285–295.

Jones, B., & Nachtsheim, C. J. (2011). A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology, 43*, 1–15.

Jones, B., Shinde, S. M., & Montgomery, D. C. (2015a). Alternatives to resolution III regular fractional factorial designs for 9–14 factors in 16 runs. *Applied Stochastic Models in Business and Industry, 31*, 50–58.

Jones, B., Silvestrini, R. T., Montgomery, D. C., & Steinberg, D. M. (2015b). Bridge designs for modeling systems with low noise. *Technometrics, 57*(2), 155–163.

Krishnamoorthy, A., Montgomery, D. C., Jones, B., & Borror, C. M. (2015). Analyzing no-confounding designs using the Dantzig selector. *International Journal of Experimental Design and Process Optimization, 4*, 183–205.

Montgomery, D. C. (1992). The use of statistical process control and design of experiments in product and process development. *IIE Transactions, 24*(5), 4–17.

Montgomery, D. C. (1999). Experimental design for product and process design and development (with commentary). *Journal of the Royal Statistical Society Series D (The Statistician), 48*, Part 2, 159–177.

Montgomery, D. C. (2012). *Design and analysis of experiments* (8th ed.). Hoboken, NJ: Wiley

Montgomery, D. C., & Runger, G. C. (2014). *Applied statistics and probability for engineers* (6th ed.). New York: John Wiley & Sons

Montgomery, D. C., & Woodall, W. H. (2008). An overview of six sigma. *International Statistical Review, 76*(3), 329–346.

Shinde, S. M., Montgomery, D. C., & Jones, B. (2014). Projection properties of no-confounding designs for six, seven, and eight factors in 16 runs. *International Journal of Experimental Design and Process Optimization, 4*(1), 1–26.

Silvestrini, R. T., Montgomery, D. C., & Jones, B. (2013). Comparing computer experiments for the Gaussian process model using integrated prediction variance. *Quality Engineering, 25*(2), 164–174.

Thomke, S. H. (2003a). *Experimentation matters: Unlocking the potential of new technologies for innovation*. Boston: Harvard Business Press.

Thomke, S. H. (2003b). R&D comes to service. *Harvard Business Review, 81*(4), 70–79.

D-Optimal Three-Stage Unbalanced Nested Designs for the Determination of Measurement Precision

Seiichi Yasui and Yoshikazu Ojima

Abstract The precision of measurement results can be quantified by variance components of random effect models. The variance components are estimated from measurement results that are obtained by performing a collaborative assessment experiment. The measurement results are statistically modeled by a nested design. Although balanced nested designs are widely used, staggered nested designs, which are one type of unbalanced nested designs, have the statistical advantage that the degrees of freedom in all stages except for the top stage are equal. Thus, balanced nested designs do not necessarily have a better performance from the statistical point of view. In this study, *D*-optimal designs are identified in general nested designs that include both balanced and unbalanced designs and consider the practical feasibility of collaborative assessment experiments as well.

Keywords Random effect · Sample size · Repeatability · Reproducibility · Sensitive analysis

1 Introduction

Nested designs are used to statistically determine the precision of measurement results in ISO 5725-3 (1994). In ISO 5725-1 (1994), the precision of measurement is defined as "the closeness of agreement between independent test results obtained under stipulated conditions". This definition implies that the precision depends on the conditions under which objects are measured. Two important conditions are repeatability and reproducibility conditions. Repeatability conditions are defined as "conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time". Reproducibility conditions are defined as

S. Yasui (✉) · Y. Ojima
Tokyo University of Science, Yamazaki, Noda, Chiba, Japan
e-mail: yasui@rs.tus.ac.jp; ojima@rs.tus.ac.jp

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_16

"conditions where test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment". In other words, repeatability conditions are conditions in which the dispersion of measurement results is minimal, whereas reproducibility conditions are conditions in which the dispersion of measurement results is maximal in the range of our interest. The precisions under such conditions are called repeatability precision and reproducibility precision, respectively.

These precisions are quantitatively determined as the variances of measurement results that are obtained from nested designs. Measurement results from nested designs are modeled by hierarchical random effect models, and the precisions are usually estimated by a linear combination of the variance component estimators based on an analysis of variance.

Although balanced nested designs are widely used, staggered nested designs, which are one type of unbalanced nested designs, have the statistical advantage that the degrees of freedom in all stages except for the top stage are equal. Thus, balanced nested designs do not necessarily have a better performance from the statistical point of view, and there are favorable situations for unbalanced nested designs. In our study, we focus on three stage nested designs, and *D*-optimal designs with respect to the estimation of repeatability, intermediate, and reproducibility precisions are investigated in some situations regarding the magnitudes of the variance components and given sample sizes.

Gold and Gaylor (1970) investigated three stage nested designs for the estimation of precision by variance components with respect to *A*-, *D*-, and adjusted (scaled) *A*-optimality and found optimal designs under the quite restricted situation that the sample sizes are multiples of 12. They assumed that the three-stage nested design with 12 observations is replicated as a block. We find *D*-optimal designs for any sample size under some variance component configurations by developing an effective algorithm to search all unbalanced nested designs in which all the degrees of freedom are non-zero.

In Sect. 2, we discuss appropriate estimators of the precision of measurements. In Sect. 3, *D*-optimal designs for some sample sizes are shown under some variance component configurations, and Sect. 4 is the conclusion.

2 *D*-Optimality for the Determination of Measurement Precision

The statistical model for unbalanced nested designs with three stages is

$$
\begin{gathered}
y_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk} \\
i = 1, \ldots, a,\ j = 1, \ldots, b_i,\ k = 1, \ldots, r_{ij} \\
\alpha_i \sim i.i.d.N(0, \sigma_A^2),\ \beta_{ij} \sim i.i.d.N(0, \sigma_B^2),\ \varepsilon_{ijk} \sim i.i.d.N(0, \sigma_E^2)\ , \qquad (1)
\end{gathered}
$$

Table 1 ANOVA table

Source	Sum of squares	Degrees of freedom	Mean square	Expected mean square
A	SSA	$\phi_A = a - 1$	$MSA = SSA/\phi_A$	$\sigma_E^2 + l_{AB}\sigma_B^2 + l_{AA}\sigma_A^2$
B	SSB	$\phi_B = b - a$	$MSB = SSB/\phi_B$	$\sigma_E^2 + l_{BB}\sigma_B^2$
A	SSE	$\phi_E = n - b$	$MSE = SSE/\phi_E$	σ_E^2

where μ is a general mean (constant). The variances σ_A^2, σ_B^2, and σ_E^2 are called variance components.

The variance components are estimated by an analysis of variance (ANOVA) which is widely used in practice. Such an estimation and its estimator are called ANOVA estimation, and ANOVA estimator, respectively. An ANOVA table is shown in Table 1. The l_{AA}, l_{AB}, and l_{BB} in Table 1 are constants which are derived by Leone et al. (1968) and Ojima (1984).

The ANOVA estimator of the variance components is the solution of the equation

$$\begin{pmatrix} MSA \\ MSB \\ MSE \end{pmatrix} = \begin{pmatrix} l_{AA} & l_{AB} & 1 \\ 0 & l_{BB} & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\sigma}_A^2 \\ \hat{\sigma}_B^2 \\ \hat{\sigma}_E^2 \end{pmatrix}. \tag{2}$$

Let the coefficient matrix of the equation be L^{-1}. Then, the ANOVA estimator is $L\vec{v}$ where $\vec{v} = (MSA, MSB, MSE)'$, and L is the inverse of L^{-1} .

In experiments to determine the precision of measurement results, the variances under repeatability conditions, intermediate conditions, and reproducibility conditions are defined as

$$\begin{aligned} \sigma_A^2 + \sigma_B^2 + \sigma_E^2 \quad & \text{(reproducibility variance)}, \\ \sigma_B^2 + \sigma_E^2 \quad & \text{(intermediate variance)}, \\ \sigma_E^2 \quad & \text{(repeatability variance)}, \end{aligned}$$

respectively. Their ANOVA estimators are provided by replacing $(\sigma_A^2, \sigma_B^2, \sigma_E^2)$ by their estimators $(\hat{\sigma}_A^2, \hat{\sigma}_B^2, \hat{\sigma}_E^2)$. Thus, the estimator of these precisions can be expressed as $\vec{C}L\vec{v}$ where

$$\vec{C} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3}$$

The elements of $CL\vec{v}$ in order from the top are the estimators of the repeatability, the intermediate, and the reproducibility variance, respectively. Note that the estimator of the variance components is expressed by matrix form as $C = I$, where I is the identity matrix.

The variance-covariance matrix of $CL\vec{v}$ is $\vec{C}\vec{L}V(\vec{v})\vec{L}'\vec{C}'$. The determinant of the matrix is

$$|CLV(\vec{v})\vec{L}'\vec{C}'| = \frac{1}{l_{AA}^2 l_{BB}^2}|V(\vec{v})|, \tag{4}$$

due to

$$\vec{L} = \begin{pmatrix} 1/l_{AA} & -l_{AB}/(l_{AA}l_{BB}) & (l_{AB} - l_{BB})/(l_{AA}l_{BB}) \\ 0 & 1/l_{BB} & -1/l_{BB} \\ 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

Gold and Gaylor (1970) provided the variances and the covariances of estimators of the variance components $LV(\vec{v})L'$ in three-stage unbalanced nested designs. Ojima (1984) derived the variances and the covariances of sums of squares based on the canonical form induced by the orthogonal transformation in three-stage unbalanced nested designs. From Ojima (1984), due to the covariances $Cov(SSA, SSE) = 0$ and $Cov(SSB, SSE)$, the determinant of the variance-covariance matrix of the estimators of the variances is

$$\frac{1}{l_{AA}^2 l_{BB}^2 \phi_A^2 \phi_B^2 \phi_E^2} V(SSE)\left[V(SSA)V(SSB) - Cov(SSA, SSB)\right]. \tag{6}$$

The matrix $\vec{C}$ can be generalized under the restriction of non-singularity. Then, since the determinant of the variance-covariance matrix of the precision estimators is proportional to the Eq. (4), the D-optimal design for the general nonsingular C is the same as that for the matrix (3). However, the matrix

$$C = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

does not make sense because such estimators are not meaningful for the precision of measurements. In particular, the lower rank matrix such as a 2×3 matrix results is "without replication" in a certain stage. For example, If the matrix is

$$C = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

the replication in the third stage is not necessary, in other word, the design with $\phi_E = 0$ is available in this case, since it is only enough to estimate $\sigma_B^2 + \sigma_E^2$. Consequently, we consider the D-optimal designs obtained by minimising the determinant (6).

3 *D*-Optimal Three-Stage Unbalanced Nested Designs

3.1 Derivation of the Optimal Designs

The *D*-Optimal Designs in the general case where there is no restriction regarding a, the b_i's, and the r_{ij}'s are mathematically interesting. However, in such a situation, unfeasible or unrealistic designs might be picked up as the optimal designs. In collaborative experiments to determine the precision of measurements, a is the number of the participating laboratories, b_i is the number of the measurement operators in the laboratory, and r_{ij} is the number of replications of the measurement. For example, the design with $a = 2$, $b_1 = 5$, and $b_2 = 1$ is too unbalanced to be practical so that a cost problem could occur. Hence, we consider the restricted designs that consist of the five fundamental structures $(d_1, d_2, d_3, d_4, d_5)$ shown in Fig. 1.

The *D*-optimal design exists in all possible combinations of fundamental structures such that all the degrees of freedom are positive. For each given number of observations n, such combinations are generated, the determinant (6) is calculated for each combination and the *D*-optimal design for n observations is identified.

The unbalanced nested design constituted from fundamental structures is denoted as $\mathcal{D} = (m_1, m_2, m_3, m_4, m_5)$, where m_i is the number of the fundamental structure d_i in $\mathcal{D}$. Thus, the total number of observations n is equal to $4m_1 + 3m_2 + 2m_3 + 2m_4 + m_5$. In order that it is possible to estimate all the variance components, at least one of the designs d_1 or d_2 or (d_3, d_4) is necessary in $\mathcal{D}$. If only either d_1 or d_2 is

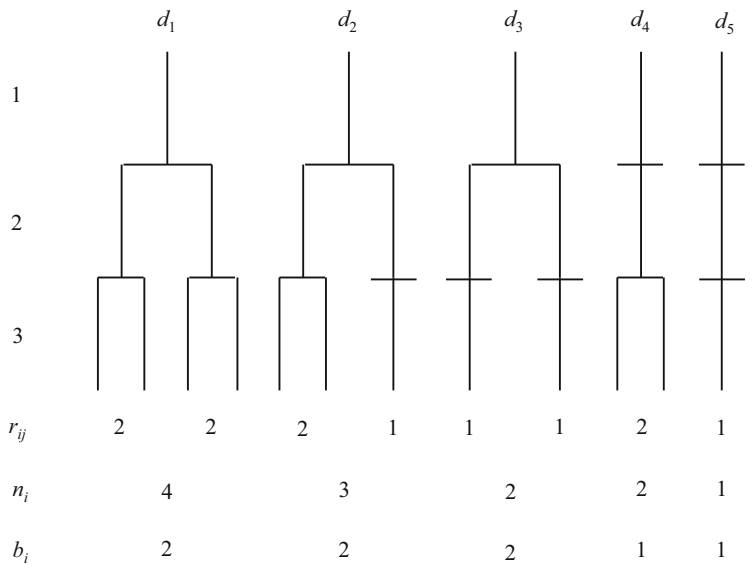


Fig. 1 Fundamental structures

included in the design, two or more structures are drawn so that $m_1 + m_2 \geq 1$ and $\sum_{i=1}^{5} m_i \geq 2$. On the other hand, if both d_3 and d_4 are included in the design, this design is always feasible, i.e. all the degrees of freedom for this design are non-zero.

Hence, $n = 4$ is the minimum number of observations, and the possible designs are $\mathcal{D}_1 = (0, 1, 0, 0, 1)$ and $\mathcal{D}_2 = (0, 0, 1, 1, 0)$. In case of $n = 5$, there are five possible designs which are $\mathcal{D}_1 = (1, 0, 0, 0, 1)$, $\mathcal{D}_2 = (0, 1, 1, 0, 0)$, $\mathcal{D}_3 = (0, 1, 0, 1, 0)$, $\mathcal{D}_4 = (0, 1, 0, 0, 2)$, and $\mathcal{D}_5 = (0, 0, 1, 1, 1)$. Let $opt_D(D_l)$ be the value of the determinant (6) for the design $\mathcal{D}_l$. We calculate $opt_D(D_1) = 55.73$, $opt_D(D_2) = 61.41$, $opt_D(D_3) = 80.64$, $opt_D(D_4) = 80.58$ and $opt_D(D_5) = 120.04$ in $\sigma_A^2 = \sigma_B^2 = \sigma_E^2 = 1^2$, and $\mathcal{D}_1$ is identified as the D-optimal design with $n = 5$ under the situation where all the variances in the stages are one.

The D-optimal designs for $n = 4(1)60$ are calculated by generating all the combinations with repetitions exhaustively. Table 2 shows the list of D-optimal designs for sample sizes $n = 5, 10, 20, 30, 60$ under several situations (ρ_A, ρ_B), where $\rho_A = \sigma_A^2/\sigma_E^2$ and $\rho_B = \sigma_B^2/\sigma_E^2$. For $n \geq 20$, the balanced design or a nearly balanced design is optimal in situations where $\rho_A \leq 2$. For $n = 60$, the balanced design is optimal in any situation except for $(\rho_A, \rho_B) = (8, 8)$.

The triplet in Table 3 denotes the degrees of freedom (ϕ_A, ϕ_B, ϕ_E) of the D-optimal design. ϕ_A is close to ϕ_B for any n, ρ_A, and ρ_B. If ρ_A and ρ_B are larger, there is less difference among the degrees of freedom ϕ_A, ϕ_B, and ϕ_E. This result means that the staggered nested designs provide more accurate estimates in situations of larger ϕ_A and ϕ_B with respect to the generalized variances of the estimators of variance components and their linear combinations.

3.2 *Sensitivity of the Generalized Variance to Sample Size n*

The precision of the estimators and the sampling cost, i.e. the choice of the sample sizes, are important aspects when determining the preferable experiment design. Often the optimality with respect to the precision of the estimators is less important than the effort regarding the number of replications in the three stages. Or in other words, in practice one would prefer a slightly less precision if it needs less sampling effort. Hence, in this section the relationship between the generalized variance expressed as Eq. (6) and the sample size n is determined.

The determinants for the optimal designs depend on three parameters: the ratios ρ_A, ρ_B and the sample size n. Since the exact Eq. (6) for the determinant $v_n = V_n(\rho_A, \rho_B)$ as a function of n is too complicated we approximate it by a linear equation.

v_n is strong nonlinearly related to n in the region of small sample sizes $n < 10$. From a practical perspective, designs with small sample sizes $n < 10$ should not be used and hence, we do not need to investigate possible cost reductions for these sample sizes. For $n \geq 10$ and any ρ_A and ρ_B an empirical relation between the

Table 2 Optimal designs

		ρ_B						
$n = 5$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	**(0,1,1,0,0)**	**(0,1,1,0,0)**	**(0,1,1,0,0)**	**(0,1,1,0,0)**
	0.25	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	**(0,1,1,0,0)**	**(0,1,1,0,0)**	**(0,1,1,0,0)**	**(0,1,1,0,0)**
	0.5	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	**(0,1,1,0,0)**	**(0,1,1,0,0)**	**(0,1,1,0,0)**	**(0,1,1,0,0)**
	1.0	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	**(0,1,1,0,0)**	**(0,1,1,0,0)**	**(0,1,1,0,0)**
	2.0	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	**(0,1,1,0,0)**	**(0,1,1,0,0)**	**(0,1,1,0,0)**
	4.0	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	**(0,1,1,0,0)**	**(0,1,1,0,0)**
	8.0	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	(1,0,0,0,1)	**(0,1,1,0,0)**	**(0,1,1,0,0)**
		ρ_B						
$n = 10$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	**(2,0,0,1,0)**	(2,0,1,0,0)	(2,0,1,0,0)	(2,0,1,0,0)	(2,0,1,0,0)	(2,0,1,0,0)	*(0,2,2,0,0)*
	0.25	**(2,0,0,1,0)**	(2,0,1,0,0)	(2,0,1,0,0)	(2,0,1,0,0)	(2,0,1,0,0)	(2,0,1,0,0)	*(0,2,2,0,0)*
	0.5	**(2,0,0,1,0)**	**(2,0,0,1,0)**	(2,0,1,0,0)	(2,0,1,0,0)	(2,0,1,0,0)	(2,0,1,0,0)	*(0,2,2,0,0)*
	1.0	**(2,0,0,1,0)**	***(2,0,0,0,2)***	(2,0,1,0,0)	(1,0,0,0,1)	(2,0,1,0,0)	(2,0,1,0,0)	*(0,2,2,0,0)*
	2.0	***(2,0,0,0,2)***	***(2,0,0,0,2)***	(2,0,1,0,0)	(1,0,0,0,1)	(2,0,1,0,0)	(2,0,1,0,0)	*(0,2,2,0,0)*
	4.0	***(2,0,0,0,2)***	***(2,0,0,0,2)***	***(2,0,0,0,2)***	***(2,0,0,0,2)***	(2,0,1,0,0)	(2,0,1,0,0)	(1,2,0,0,0)
	8.0	***(2,0,0,0,2)***	***(2,0,0,0,2)***	***(2,0,0,0,2)***	***(2,0,0,0,2)***	(2,0,1,0,0)	(2,0,1,0,0)	(1,2,0,0,0)
		ρ_B						
$n = 20$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	**(4,0,2,0,0)**	**(4,0,2,0,0)**
	0.25	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	**(4,0,2,0,0)**	**(4,0,2,0,0)**
	0.5	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	**(4,0,2,0,0)**	**(4,0,2,0,0)**
	1.0	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	**(4,0,2,0,0)**	**(4,0,2,0,0)**
	2.0	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	**(4,0,2,0,0)**	**(4,0,2,0,0)**
	4.0	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	**(4,0,2,0,0)**	*(2,4,0,0,0)*
	8.0	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	(5,0,0,0,0)	*(2,4,0,0,0)*	(0,6,1,0,0)
		ρ_B						
$n = 30$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	**(7,0,0,1,0)**	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)
	0.25	**(7,0,0,1,0)**	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)
	0.5	**(7,0,0,1,0)**	**(7,0,0,1,0)**	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)
	1.0	**(7,0,0,1,0)**	***(7,0,0,0,2)***	***(7,0,0,0,2)***	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)
	2.0	**(7,0,0,1,0)**	***(7,0,0,0,2)***	***(7,0,0,0,2)***	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)
	4.0	**(7,0,0,1,0)**	***(7,0,0,0,2)***	***(7,0,0,0,2)***	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)
	8.0	**(7,0,0,1,0)**	***(7,0,0,0,2)***	***(7,0,0,0,2)***	(7,0,1,0,0)	(7,0,1,0,0)	(7,0,1,0,0)	(0,10,0,0,0)
		ρ_B						
$n = 60$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)
	0.25	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)
	0.5	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)
	1.0	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)
	2.0	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)
	4.0	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)
	8.0	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	(15,0,0,0,0)	**(0,20,0,0,0)**

Table 3 Degrees of freedom for each optimal design

		ρ_B						
$n = 5$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**
	0.25	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**
	0.5	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**
	1.0	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**
	2.0	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**
	4.0	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**
	8.0	(11,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**
		ρ_B						
$n = 10$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	**(2,2,5)**	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)	*(3,4,2)*
	0.25	**(2,2,5)**	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)	*(3,4,2)*
	0.5	**(2,2,5)**	**(2,2,5)**	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)	*(3,4,2)*
	1.0	**(2,2,5)**	***(3,2,4)***	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)	*(3,4,2)*
	2.0	***(3,2,4)***	***(3,2,4)***	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)	*(3,4,2)*
	4.0	***(3,2,4)***	***(3,2,4)***	***(3,2,4)***	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)
	8.0	***(3,2,4)***	***(3,2,4)***	***((3,2,4)***	(2,3,4)	(2,3,4)	(2,3,4)	(2,3,4)
		ρ_B						
$n = 20$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	**(5,6,8)**
	0.25	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	**(5,6,8)**
	0.5	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	**(5,6,8)**
	1.0	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	**(5,6,8)**
	2.0	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	**(5,6,8)**
	4.0	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	***(5,6,8)***
	8.0	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	***(5,6,8)***	(6,7,6)
		ρ_B						
$n = 30$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	**(7,7,15)**	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)
	0.25	**(7,7,15)**	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)
	0.5	**(7,7,15)**	**(7,7,15)**	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)
	1.0	**(7,7,15)**	***(8,7,14)***	***(8,7,14)***	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)
	2.0	**(7,7,15)**	***(8,7,14)***	***(8,7,14)***	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)
	4.0	**(7,7,15)**	***(8,7,14)***	***(8,7,14)***	(7,8,14)	(7,8,14)	(7,8,14)	(7,8,14)
	8.0	**(7,7,15)**	***(8,7,14)***	***(8,7,14)***	(7,8,14)	(7,8,14)	(7,8,14)	(9,10,10)
		ρ_B						
$n = 60$		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	0.25	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	0.5	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	1.0	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	2.0	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	4.0	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	8.0	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	**(19,20,20)**

Table 4 Results for (7) given various variance ratios ρ_A and ρ_B: upper value—$\hat{\beta}_0$, lower—$\hat{\beta}_1$

		ρ_B						
		0.125	0.25	0.5	1.0	2.0	4.0	8.0
ρ_A	0.125	5.203	4.834	5.900	7.509	9.581	11.924	14.398
		−3.574	−3.293	−3.303	−3.332	−3.365	−3.380	−3.371
	0.25	4.606	5.200	6.166	7.665	9.657	11.958	14.415
		−3.269	−3.274	−3.280	−3.306	−3.342	−3.364	−3.362
	0.5	5.258	5.784	6.632	7.969	9.819	12.03	14.452
		−3.250	−3.253	−3.254	−3.272	−3.308	−3.338	−3.347
	1.0	6.152	6.598	7.351	8.504	10.145	12.198	14.531
		−3.234	−3.231	−3.232	−3.240	−3.240	−3.299	−3.319
	2.0	7.233	7.620	8.288	9.309	10.723	12.537	14.701
		−3.211	−3.214	−3.212	−3.216	−3.230	−3.253	−3.279
	4.0	8.433	8.789	9.407	10.347	11.578	13.139	15.025
		−3.209	−3.200	−3.197	−3.203	−3.205	−3.213	−3.225
	8.0	9.720	10.057	10.640	11.532	12.658	14.010	15.618
		−3.203	−3.193	−3.188	−3.192	−3.192	−3.185	−3.182

logarithm of the determinant v_n and the logarithm of the sample size n,

$$\ln v_n = \beta_0 + \beta_1 \ln n + \varepsilon_n, \quad n \geq 10, \tag{7}$$

is assumed, and the coefficients β_0 and β_1 are calculated as $\hat{\beta}_0$ and $\hat{\beta}_1$ by ordinary least squares. For all the situations Table 4 shows in each cell $\hat{\beta}_0$ as the upper value and $\hat{\beta}_1$ as the lower value. For all the situations the coefficients of determination are larger than 0.99, all the slopes $\hat{\beta}_1$ are negative, which means that the generalized variance decreases with increasing sample size. The intercepts $\hat{\beta}_0$ are increasing if ρ_A and ρ_B become larger.

In order to consider the possibility to choose a smaller sample size and preserve an acceptable precision, the sensitivity of the empirical relation (7) is considered. The prediction function of the empirical relation (7) is

$$v_n = \mathrm{e}^{\hat{\beta}_0} n^{\hat{\beta}_1}, \quad n \geq 10. \tag{8}$$

The first derivate of Eq. (8) is

$$\frac{dv_n}{dn} = \mathrm{e}^{\hat{\beta}_0} \hat{\beta}_1 n^{\hat{\beta}_1 - 1}, \quad n \geq 10. \tag{9}$$

If the generalized variance v_n is not considerably increasing if the sample size is reduced, the less precise design with smaller sample size is acceptable in practice. Let A_L be an acceptable limit of the maximal deterioration of v_n that can be ignored in practice. Thus, if for the optimal design with sample size n, $|dv_n/dn| \leq A_L$

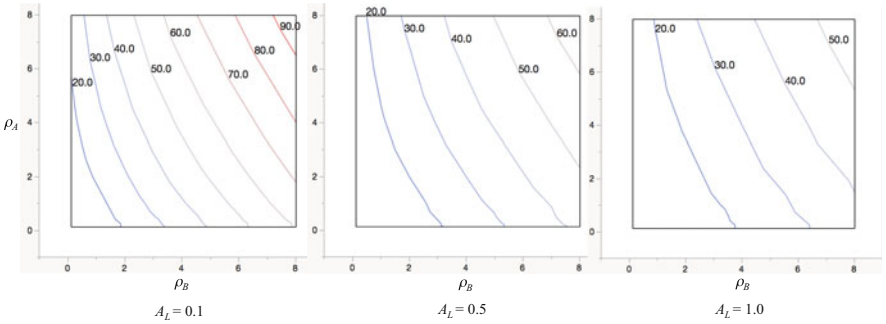


Fig. 2 The smallest sample size n^* for a given acceptable limit A_L

holds, the design is called a practically acceptable optimal design in this chapter. The smallest sample size n^* is

$$\left\lceil \left(-A_L \mathrm{e}^{-\hat{\beta}_0} / \hat{\beta}_1 \right)^{\frac{1}{\hat{\beta}_1 - 1}} \right\rceil, \tag{10}$$

where $\hat{\beta}_1$ is assumed to be negative.

Contours of the smallest sample sizes n^* that satisfy the acceptable limit A_L are shown in Fig. 2. For instance, if $\rho_A = 4.0$, $\rho_B = 4.0$ and $A_L = 1.0$, the smallest sample size is $n^* = 30$. The optimal design with $n = 30$ is given in Table 2 as $(7, 0, 1, 0, 0)$.

Hence, an acceptable design might be determined based on both a precision of the estimator (8) and a sensitivity (9). However, it should be noted that the variance ratios ρ_A and ρ_B are unknown and must be predicted based on past empirical data.

4 Conclusions

We obtain D-optimal designs for the determination of the precision of measurement under 49 variance component configurations. Balanced designs are optimal in a wide range of these configurations. If the variance components of both the first and the second upper stage are much larger than that of the third stage variance component, e.g., $\rho_A = \rho_B = 8.0$, staggered or nearly staggered nested designs are optimal.

In order to choose the optimal design, practitioners have to specify the unknown ratios of variance components ρ_A and ρ_B by a priori assumptions. Though the exact specification of the values of the variance component ratios is difficult in practice, there are situations in which they can be predicted approximately.

Since the lowest stage corresponds to measurement under repeatability conditions and the highest stage corresponds to measurement under reproducibility

conditions, the variance components have a coherent relation $\sigma_E^2 \leq \sigma_B^2 \leq \sigma_A^2$, or in other words, $1 \leq \rho_B \leq \rho_A$. In such a region, for $n \geq 30$, the optimal design is unique except for $\rho_A = \rho_B = 8.0$, e.g. $(7, 0, 1, 0)$ and $(15, 0, 0, 0)$ are optimal designs for $n = 30$ and $n = 60$, respectively. For the case of extreme variance ratios ($\rho_A = \rho_B = 8.0$ or more) and $n = 30, 60$, it is found that the staggered nested designs are preferable.

In Chap. 3, we developed a procedure to find the optimal design from a given acceptable limit for the overall precision of estimation. The acceptable limit regarding the precision of the estimators of the variance components (8) and its sensitivity (9) should be supplied by the practitioners. For instance, if $\rho_A = 4.0$ and $\rho_B = 4.0$ are assumed and an acceptable limit is determined as $A_L = 1.0$, we find that the smallest sample size is $n^* = 30$ and from Table 2 the optimal design with $n = 30$ is $(7, 0, 1, 0, 0)$. In addition, even if ρ_A and ρ_B are different from the assumed values the design $(7, 0, 1, 0, 0)$ remains suitable.

This chapter provides theoretical rather than practical features regarding the performance of precision experiments based on unbalanced nested designs. In practice additional aspects must be taken into consideration: measurement cost (the number of replications or sample size), the precision of estimation, the problems of the variance components to be unknown, and so on. Thus, the practical optimality must be formulated in a decision-making theory framework. If the utility function with arguments ρ_A and ρ_B can appropriately be defined, we are able to obtain the preferable optimal design systematically by mathematical programming. This is an issue of our future work. Furthermore, optimal designs in more general unbalanced nested designs should be found and investigated in future work. In general unbalanced cases, the number of candidate designs rapidly increases according to sample size. Combinatorial optimization should be invoked to solve the problem.

References

Goldsmith, C. H., & Gaylor, D. W. (1970). Three stage nested designs for estimating variance components. *Technometrics, 12*, 487–498.

ISO 5725-1. (1994). Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions. International Organization for Standardization, Geneva, Switzerland.

ISO 5725-3. (1994). Accuracy (trueness and precision) of measurement methods and results – Part 3: Intermediate measures of the precision of a standard measurement method. International Organization for Standardization, Geneva, Switzerland.

Leone, F. C., Nelson, L. S., Johnson, N. L., & Eisenstat, S. (1968). Sampling distributions of variance components II. Empirical studies of unbalanced nested designs. *Technometrics, 10*, 719–737.

Ojima, Y. (1984). The use of canonical forms for estimating variance components in unbalanced nested designs. *Reports of Statistical Application Research, 31*, 1–18.

Part III
Related Areas

Sampling Inspection by Variables Under Weibull Distribution and Type I Censoring

Peter-Th. Wilrich

Abstract The lifetime (time to failure) of a product is modeled as Weibull distributed (with unknown parameters); in this case the logarithms of the lifetimes are Gumbel distributed. Lots of items shall be accepted if their fraction p of nonconforming items (items the lifetime of which is smaller than a lower specification limit t_L) is not larger than a specified acceptable quality limit. The acceptance decision is based on the $r \leq n$ observed lifetimes of a sample of size n which is put under test until a defined censoring time t_C is reached (Type I censoring). A lot is accepted if $r = 0$ or if the test statistic $y = \hat{\mu} - k\hat{\sigma}$ is not smaller than the logarithm of the specification limit, $x_L = \log(t_L)$, where k is an acceptance factor and $\hat{\mu}$ and $\hat{\sigma}$ are the Maximum Likelihood estimates of the parameters of the Gumbel distribution. The parameters of the sampling plan (acceptance factor k, sample size n and censoring time t_C) are derived so that lots with $p \leq p_1$ shall be accepted with probability not smaller than $1 - \alpha$. On the other hand, lots with fractions nonconforming larger than a specified value p_2 shall be accepted with probability not larger than β. n and t_C are not obtained separately but as a function that relates the sample size n to the censoring time t_C. Of course, n decreases if the censoring time t_C is increased. For $t_C \to \infty$ the smallest sample size, i.e. that of the uncensored sample, is obtained. Unfortunately, the parameters of the sampling plan do not only depend on the two specified points of the OC, $P_1(p_1, 1 - \alpha)$ and $P_2(p_2, \beta)$, but directly on the parameters τ and δ of the underlying Weibull distribution or equivalently, on the parameters $\mu = \log(\tau)$ and $\sigma = 1/\delta$ of the corresponding Gumbel distribution. Since these parameters are unknown we assume that the hazard rate of the underlying Weibull distribution is nondecreasing ($\delta \geq 1$). For the design of the sampling plan we use the limiting case $\delta = 1$ or $\sigma = 1/\delta = 1$. A simulation study shows that the OC of the sampling plan is almost independent of σ if the censoring time t_C is not smaller than the specification limit t_L.

P.-Th. Wilrich (✉)
Institut für Statistik und Ökonometrie, Freie Universität Berlin, Berlin, Germany
e-mail: wilrich@wiwiss.fu-berlin.de

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_17

Keywords Sampling inspection · Inspection by variables · Variables sampling · Lifetime · Life test · Weibull distribution · Gumbel distribution · Censoring

1 Introduction

The lifetime (time to failure) is an important quality characteristic of many types of product. If a lower specification limit t_L for the lifetime is established an item is nonconforming if its lifetime t is smaller than t_L. In order to test whether the fraction of nonconforming items in a lot of a product, p, is small so that it can be accepted, or that p is large so that it should be rejected, a sample of size n is put on test and the lifetimes of the samples items are noted. In sampling inspection by attributes the number of lifetimes of the sample being smaller than the lower specification limit is used for the acceptance decision whereas in sampling inspection by variables the lifetimes of the sample are statistically evaluated for the acceptance decision.

Technical Report TR 3 (1961), Technical Report TR 4 (1962), Technical Report TR 6 (1963), Technical Report TR 7 (1965), based on Goode and Kao (1961, 1962, 1963) present sampling plans for inspection by attributes for the lifetime assumed to be Weibull distributed with known shape parameter δ and specification limits established for the mean life, the hazard rate or the reliable life. Since it is known that sampling by attributes requires larger sample sizes than sampling by variables in order to work with equal efficiency it seems favorable to apply sampling plans for inspection by variables. Most of the existing sampling plans for inspection by variables as, e.g. ISO 3951-1 (2005), ISO 3951-2 (2005), cannot be applied to lifetimes because they assume a normal distribution of the quality characteristic which is unrealistic for lifetimes, and they require the lifetimes of all sampled items to be measured. Instead of the normal distribution the Weibull distribution is very often an appropriate assumption for the distribution of lifetimes. And economical considerations require the life test to be finished when only a specified number r of items of the sample have failed (Type II censoring) or a specified test time t_C has elapsed (Type I censoring).

Type II censored sampling plans for inspection by variables under Weibull distribution have been presented by Fertig and Mann (1980) and Hosono et al. (1981). They used best linear unbiased estimators (BLUEs) of the parameters of the Weibull distribution for the acceptance decision which need tables of the coefficients being available only for small sample sizes. Schneider (1989) based the acceptance procedure on Maximum Likelihood estimators and their asymptotic normal distribution.

I deal with Type I censored sampling plans for inspection by variables which have the advantage of the test time t_C being fixed in advance. Section 2 presents the Weibull distribution and the Gumbel distribution as the underlying model. Section 3 describes the sampling plans and their design. Section 4 gives an example. Section 5 presents a graphical procedure that uses Weibull probability chapter. The Maximum Likelihood estimators of the parameters of the Gumbel distribution

and the asymptotic variance of the test statistic are derived in Annexes A and B, respectively.

2 The Model

The lifetime of a product is modelled as a random variable T that is Weibull distributed with the probability density function

$$f_T(t) = \frac{\delta}{\tau}\left(\frac{t}{\tau}\right)^{\delta-1} \exp\left(-\left(\frac{t}{\tau}\right)^{\delta}\right); x > 0 \tag{1}$$

where $\tau > 0$ is a scale parameter and $\delta > 0$ is a shape parameter. The cumulative distribution function of T is

$$F_T(t) = P(T \le t) = 1 - \exp\left(-\left(\frac{t}{\tau}\right)^{\delta}\right), \tag{2}$$

the survival function is

$$G_T(t) = P(T > t) = \exp\left(-\left(\frac{t}{\tau}\right)^{\delta}\right) \tag{3}$$

and the failure rate (hazard rate) is

$$h_T(t) = \frac{f_T(t)}{1 - F_T(t)} = \frac{\delta}{\tau}\left(\frac{t}{\tau}\right)^{\delta-1}; \tag{4}$$

$h_T(t)$ is monotonically increasing (decreasing) for $\delta > 1$ ($\delta < 1$). For $\delta = 1$, $h_T(t)$ is constant, $h_T(t) = 1/\tau$; in this case T follows the exponential distribution.

The transformed random variable $X = \ln T$ has the survival function

$$\begin{aligned} G_X(x) &= P(\ln T > x) = P(T > e^x) = G_T(e^x) \\ &= \exp(-(e^x/\tau)^{\delta}) = \exp(-\exp(\delta(x - \ln\tau)) = \exp(-\exp((x-\mu)/\sigma)). \end{aligned} \tag{5}$$

This location and scale parameter distribution with location parameter $\mu = \ln\tau \in \mathbb{R}$ and scale parameter $\sigma = 1/\delta > 0$ (Note: μ and σ are not expectation and standard deviation of X) is the Type I asymptotic distribution of the smallest extreme value in a sample of size $n \to \infty$, often denoted as Gumbel distribution.

The linear transformation $Z = (X - \mu)/\sigma$ transforms this distribution into the standardized Gumbel distribution with the survival function

$$G_Z(z) = \exp(-\exp(z)) \tag{6}$$

and the probability density function

$$f_Z(z) = \exp(z - \exp(z)) = \exp(z)G_Z(z); \tag{7}$$

it has no parameters. In the following we use the Gumbel distribution of $X = \ln T$ instead of the Weibull distribution of T because, as a location and scale distribution, it has many advantages in the design of sampling plans.

3 The Sampling Plan

A lower limit t_L for the lifetime T of the items of a product is specified. An item is nonconforming if its lifetime is smaller than t_L, $T < t_L$. A lot of items is acceptable if its fraction of nonconforming items, p, is not larger than a specified value p_1. The sampling plan shall accept a lot with $p \leq p_1$ with probability not smaller than $1 - \alpha$. On the other hand, lots with fractions nonconforming larger than a specified value p_2 shall be accepted with probability not larger than β. $(p_1, 1 - \alpha)$ and (p_2, β) are design specifications for the sampling plan. We put n items on a life test and note the lifetimes $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(r)}$ of all items that fail until an established test time t_C is reached, i.e. the sample is censored at the right with censoring time t_C. Note that r is a random variable. Based on the logarithms $x_i = \ln t_{(i)}$ of the lifetimes $t_{(i)}$ the Maximum Likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ of the parameters μ and σ of the Gumbel distribution are calculated; see Annex A.

The lot is accepted if the test statistic

$$y = \hat{\mu} - k\hat{\sigma} \tag{8}$$

is not smaller than $x_L = \ln t_L$,

$$y = \hat{\mu} - k\hat{\sigma} \geq x_L \tag{9}$$

where k is the acceptance factor of the sampling plan (k, n, t_C), or equivalently

$$(x_L - \hat{\mu})/\hat{\sigma} \leq -k \tag{10}$$

or

$$\hat{p} = F_Z((x_L - \hat{\mu})/\hat{\sigma}) \leq F_Z(-k) = p_{crit}, \tag{11}$$

where $\hat{p}$ is an estimate of the fraction nonconforming in the lot. k, n and t_C shall be fixed so that the probabilities of acceptance of the lot are $1 - \alpha$ and β if the fractions of nonconforming items in the lot are p_1 and p_2, respectively. Since the test statistic and the estimate of the fraction nonconforming cannot be calculated if the observed number of failures is $r = 0$ the decision rules (9) and (11) are amended by the rule to accept the lot if $r = 0$; this causes a very small increase of the probability of acceptance of a lot.

Asymptotically, the test statistic $y = \hat{\mu} - k\hat{\sigma}$ is normally distributed with expectation $E(y) = \mu - k\sigma$ and variance $V(y) = \sigma_y^2 = V(\hat{\mu}) + k^2 V(\hat{\sigma}) - 2kCov(\hat{\mu}, \hat{\sigma})$; see Annex B.

The operating characteristic function (OC), i.e. the probability of acceptance of the lot as a function of its fraction nonconforming, p, is

$$\begin{aligned} L(p) &= P(y \geq x_L|p) = P(\hat{\mu} - k\hat{\sigma} \geq x_L|p) \\ &= P\left(\frac{(\hat{\mu} - k\hat{\sigma}) - (\mu - k\sigma)}{\sigma_y} \geq \frac{x_L - \mu + k\sigma}{\sigma_y}\right) \\ &= P\left(U \geq \frac{1}{A}\left(\frac{x_L - \mu}{\sigma} + k\right)|p\right) = 1 - P\left(U \leq \frac{1}{A}(z_L + k)|p\right) \\ &= 1 - \Phi\left(\frac{1}{A}(z_p + k)\right) \end{aligned} \tag{12}$$

where U is the standardized normal variable and

$$A = \sigma_y/\sigma; \tag{13}$$

$\Phi(\cdot)$ is the cumulative distribution function of the standardized normal distribution. The standardized lower specification limit $z_L = (x_L - \mu)/\sigma$ is equal to the p-quantile $z_p = \ln(-\ln(1-p))$ of the standardized Gumbel distribution if the fraction nonconforming in the lot is p.

A and k are obtained by solving the equations

$$\begin{aligned} L(p_1) &= 1 - \Phi\left(\frac{1}{A}(z_{p_1} + k)\right) = 1 - \alpha \\ L(p_2) &= 1 - \Phi\left(\frac{1}{A}(z_{p_2} + k)\right) = \beta \end{aligned} \tag{14}$$

for A and k. From the first equation we get $\Phi(\frac{1}{A}(z_{p_1} + k)) = \alpha$ or

$$\frac{1}{A}(z_{p_1} + k) = u_\alpha, \tag{15}$$

and from the second equation

$$\frac{1}{A}(z_{p_2} + k) = u_{1-\beta}, \tag{16}$$

where u_p is the p-quantile of the standardized normal distribution. Equations (15) and (16) have the solutions

$$k = \frac{z_{p_1}u_{1-\beta} - z_{p_2}u_\alpha}{u_\alpha - u_{1-\beta}} \tag{17}$$

and

$$A = \frac{z_{p_1} - z_{p_2}}{u_\alpha - u_{1-\beta}}. \tag{18}$$

The OC of the sampling plan passes through the two points $P_1(p_1, 1-\alpha)$ and $P_2(p_2, \beta)$ if the parameters of the sampling plan are k and A according to (17) and (18). The value of A according to (18) has to be equal to $A = \sigma_y/\sigma$ according to (19):

$$A = \frac{z_{p_1} - z_{p_2}}{u_\alpha - u_{1-\beta}} = \frac{\sqrt{v_{11} + k^2 v_{22} - 2k v_{12}}}{\sqrt{n}} = \frac{f(k, z_C)}{\sqrt{n}} \tag{19}$$

or

$$n = \frac{f^2(k, z_C)}{A^2}, \tag{20}$$

where v_{11}, v_{12} and v_{22} are the elements of the asymptotic covariance matrix of the estimators $\hat{\mu}$ and $\hat{\sigma}$ according to (49).

The parameters of the sampling plan, k and A, being fixed according to (17) and (18), this equation defines a series of pairs (z_C, n) for which the design requirement is met. For $z_C \to \infty$, i.e. for the case of no censoring, n takes its smallest value,

$$n_{min} = \frac{1 + \frac{6(k+1-\gamma)^2}{\pi^2}}{A^2} \tag{21}$$

according to (54), where $\gamma = 0.57721566490\ldots$ is Euler's constant (see Erdélyi (1954), p. 148). Depending on the cost of sampled items and test time the user of the sampling plan can choose smaller test times with larger sample sizes and vice versa.

In order to calculate the right hand side of (20) we need the standardized censoring time $z_C = (x_C - \mu)/\sigma$. However, we have only the established censoring time $x_C = \ln t_C$, and we cannot convert it into the standardized censoring time z_C because μ and σ are unknown.

We solve this problem with the assumption that the failure rate of the Weibull distribution of the lifetime is nondecreasing, i.e. that the failure rate of an item does not decrease if its lifetime increases. This corresponds to the case where the shape parameter of the Weibull distribution is larger or equal to 1, $\delta \geq 1$, and the scale parameter of the Gumbel distribution is not larger than 1, $\sigma = 1/\delta \leq 1$. We fix σ at $\sigma_0 = 1$ (and discuss this choice in Sect. 4). Since the fraction nonconforming in the lot is $p = F_Z(z_L) = F_Z((x_L - \mu)/\sigma_0)$ the unknown parameter μ now only depends on the fraction nonconforming p. We then choose μ so that the corresponding $p = F_Z(x_L - \mu)/\sigma_0) = p_{50\%}$ is the

indifferent quality of the sampling plan, i.e. that the probability of acceptance according to (12) is 50%, $L(p_{50\%}) = 50\%$. For this case, $z_p = (x_L - \mu)/\sigma_0 = z_{p50\%} = -k$ and we see that, according to (11), $p_{50\%} = p_{crit}$. With $\mu = x_L + k\sigma_0$ we finally obtain $z_C = (x_C - \mu)/\sigma_0 = (x_C - x_L)/\sigma_0 - k = x_C - x_L - k$.

If we calculate the standardized censoring time as

$$z_C = x_C - x_L - k = \ln t_C - \ln t_L - k \tag{22}$$

and hence,

$$n = \frac{f^2(k, \ln t_C - \ln t_L - k)}{A^2}, \tag{23}$$

we get a sampling plan the OC's of which pass through the indifference point $(p_{crit}, 50\%)$ and for $\sigma = 1$ ($\delta = 1$) through the design points $P_1(p_1, 1 - \alpha)$ and $P_2(p_2, \beta)$.

If we choose the censoring time (test time) t_C equal to the specification limit (specified lifetime) t_L the sample size n is only slightly smaller than the sample size n_{att} of the sampling plan for inspection by attributes with the same OC curve (in the example of Sect. 4 we find $n = 103$ and $n_{att} = 109$). Hence, one might decide to use the attributes sampling plan in order to avoid the assumption of a Weibull distribution of the lifetime. However, in contrast to attributes sampling variables sampling allows to choose a test time t_C that is different from the specified lifetime t_L. If the test time is larger than the specified lifetime, e.g. $t_C = 2t_L$, the sample size n is much smaller than n_{att} ($n = 75$ in the example of Sect. 4) whereas the sample size is much larger if t_C is smaller than t_L (see Fig. 2). Larger test times require a smaller number of items to be put on test and vice versa. An appropriate choice of n and t_C should be based on cost considerations.

The parameters of the sampling plan, k and n, are derived under the assumption that the test statistic y is normally distributed. This assumption is based on the asymptotic normality of the estimators $\hat{\mu}$ and $\hat{\sigma}$. In order to check this assumption I have performed simulations for different design points $P_1(p_1, \alpha)$ and $P_2(p_2, \beta)$ resulting in different sample sizes n. In all cases the normal distribution was a good approximation of the distribution of the test statistic y and even better of the estimator $\hat{p}$ of the fraction nonconforming. Hence, the simulated OC curves were in very good agreement with the asymptotic OC's passing through $P_1(p_1, \alpha)$ and $P_2(p_2, \beta)$ (see also Fig. 1).

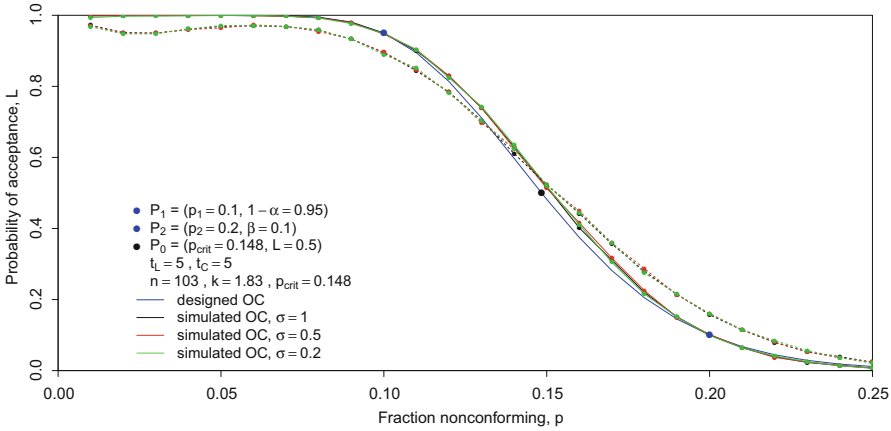


Fig. 1 The asymptotic OC curve (blue) of the sampling plan $(k, n, t_C) = (1.83, 103, 5)$ that passes through the points $P_1(p_1 = 0.1, 1 - \alpha = 0.95)$ and $P_2(p_2 = 0.2, \beta = 0.1)$. The black, red, green curves are the simulated OC curves for $\sigma = 1, 0.5, 0.2$, respectively (solid: numerical acceptance decision, dashed: graphical acceptance decision, see Sect. 5). Each point represents the average of 10^4 simulation runs

4 An Example

The lifetime T of a particular product is assumed to be Weibull distributed. An item of the product is defined as nonconforming if its lifetime t is smaller than the lower specification limit $t_L = 5$. A sampling plan for inspection by variables has to be designed so that lots with fraction nonconforming $p_1 = 0.1$ are accepted with probability $1 - \alpha = 0.95$, and lots with fraction nonconforming $p_2 = 0.2$ are accepted with probability $\beta = 0.1$.

According to (17) and (18) the parameters of the sampling plan are $k = 1.83$ and $A = 0.256$; the critical fraction nonconforming according to (11) is $p_{crit} = 0.148$. A lot is accepted if, according to (9), the test statistic y is not smaller than the lower specification limit $x_L = \ln t_L = 1.61$, or equivalently according to (11), if the estimate $\hat{p}$ is not larger than the critical fraction nonconforming, $p_{crit} = 0.148$. The blue curve of Fig. 1 shows the OC curve of this sampling plan. The two blue points on this curve are the design points $P_1(p_1 = 0.1, 1 - \alpha = 0.95)$ and $P_2(p_2 = 0.2, \beta = 0.1)$. The black point $P_0(p_{crit} = 0.148, L = 0.5)$ indicates the indifferent quality.

Figure 2 is a plot of the sample size n as a function of the censoring time t_C. If we choose the censoring time as $t_C = 2t_L, t_L, t_L/2$ we obtain the sample sizes $n = 75, 103, 296$, respectively. The smallest sample size, for the case of no censoring, is $n_{min} = 63$. The corresponding attributes sampling plan is $(n_{att} = 109, c = 16)$: if not more than 16 lifetimes are smaller than $t_L = 5$ the lot is accepted. For such an attribute sampling plan, the life test can always be finished at t_L, and hence, the censoring time is equal to the specification limit, $t_C = t_L$. It is interesting to note that

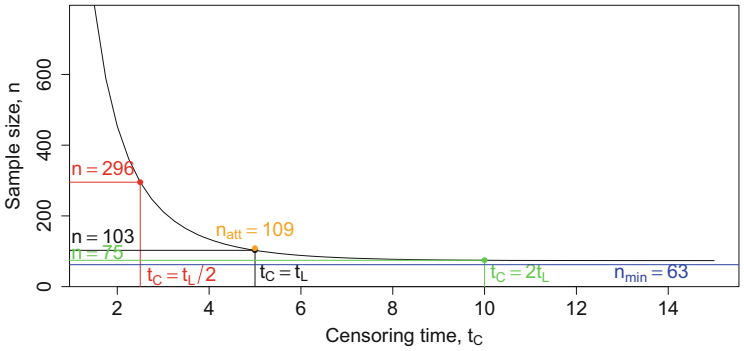


Fig. 2 The sample size n as a function of the censoring time t_C for the sampling plan the OC of which passes through the points $P_1(p_1 = 0.1, 1 - \alpha = 0.95)$ and $P_2(p_2 = 0.2, \beta = 0.1)$. For the censoring times $t_C = 2t_L$ (green), $t_C = t_L$ (black), $t_C = t_L/2$ (red) we obtain the sample sizes $n = 75, 103, 296$, respectively. The smallest sample size, for the case of no censoring, is $n_{min} = 63$ (blue). The orange point indicates the sample size of the corresponding attributes sampling plan, $n_{att} = 109$

the sample size of the attributes sampling plan, $n_{att} = 109$ (orange point in Fig. 2), is not much larger than the sample size of the variables sampling plan for $t_C = t_L$, $n = 103$.

We now start sampling with the censoring time $t_C = t_L = 5$. In a simulation experiment we choose $\sigma = 1, 0.5, 0.2$, calculate for various p the corresponding $\mu = x_L - z_p\sigma$, generate samples of size $n = 75$ with censoring time $t_C = t_L = 5$ and count the number of simulation runs in which the test statistic is larger than $x_L = \ln t_L = 1.61$. The black, red, green curves of Fig. 1 are the simulated OC curves for $\sigma = 1, 0.5, 0.2$, respectively, which are almost equal to the theoretical OC. If we now fix the censoring time at $t_C = t_L/2 = 2.5$ the sampling plan is $(k, n, t_C) = (1.83, 296, 2.5)$. Figure 3 shows that the OC's now depend very much on the standard deviation σ of the distribution of the log-lifetime, i.e. on the shape parameter $\delta = 1/\sigma$ of the distribution of the lifetime. We note that the sampling plan becomes less efficient (OC more flat) if the standard deviation is smaller than the value that had been used for the design of the sampling plan, $\sigma_0 = 1$.

Figure 4 gives an explanation of this unexpected behavior of the sampling plan. In the upper graph the censoring time is $t_C = t_L = 5$, in the lower graph it is $t_C = t_L/2 = 2.5$. The green simulated distributions of the test statistic y belong to $\sigma = 1(\delta = 1)$ of the underlying lifetime distribution, the blue distributions to $\sigma = 0.5(\delta = 2)$. The solid distributions belong to the fraction $p_1 = 0.1$ of nonconforming items in the lot, the dashed distributions to $p_2 = 0.2$. In the upper

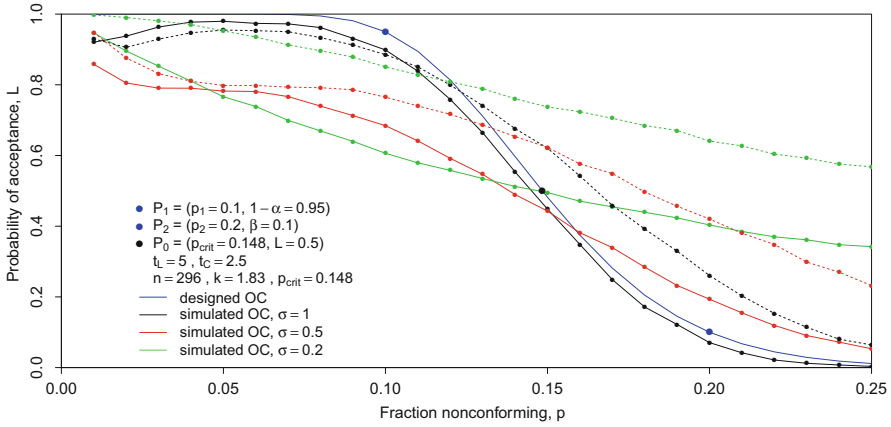


Fig. 3 The asymptotic OC curve (blue) of the sampling plan $(k, n, t_C) = (1.83, 296, 2.5)$ that passes through the points $P_1(p_1 = 0.1, 1 - \alpha = 0.95)$ and $P_2(p_2 = 0.2, \beta = 0.1)$. The black, red, green curves are the simulated OC curves for $\sigma = 1, 0.5, 0.2$, respectively (solid: numerical acceptance decision, dashed: graphical acceptance decision). Each point represents the average of 10^4 simulation runs

graph for $p_1 = 0.1$ the fraction of accepted lots (area of the distribution to the right of the specification limit $x_L = \ln 5 = 1.61$, indicated as red vertical line) is 0.948 if $\sigma = 1$ (solid green) and 0.945 if $\sigma = 0.5$ (solid blue). For $p_2 = 0.2$ it is 0.102 (dashed green) if $\sigma = 1$ and 0.103 if $\sigma = 0.5$. All these results of 10^4 simulation runs are in excellent agreement with the specified values $1 - \alpha = 0.95$ and $\beta = 0.1$, respectively. However, in the lower graph for $p_1 = 0.1$ the fraction of accepted lots is 0.896 if $\sigma = 1$ (solid green) and 0.374 if $\sigma = 0.5$ (solid blue). For $p_2 = 0.2$ it is 0.065 (dashed green) if $\sigma = 1$ and 0.319 if $\sigma = 0.5$. Whereas for $\sigma = 1$ the fractions of accepted lots are in agreement with the specified values, they are extremely different from them if $\sigma = 0.5$. A comparison of the blue distributions with the green distributions of y shows that they have a smaller standard deviation if $\sigma = 0.5(\delta = 2)$ than if $\sigma = 1(\delta = 1)$, and this would increase the efficiency of the sampling plan. On the other hand, the distributions (and the expected values of the test statistic y, indicated as points) are shifted towards the specification limit if σ decreases (δ increases), and this stronger effect decreases the efficiency of the sampling plan. Simulations show that the choice of a smaller σ_0 than $\sigma_0 = 1$ is no practical solution: it slightly turns all OC's clockwise around the point of indifferent quality, however this efficiency increasing effect is small and the price is a much larger sample size n. The best recommendation is not to use censoring times t_C smaller than the specification limit t_L. Figure 2 demonstrates another reason for this recommendation: for censoring times decreasing from the specification limit to 0 the sample size increases sharply.

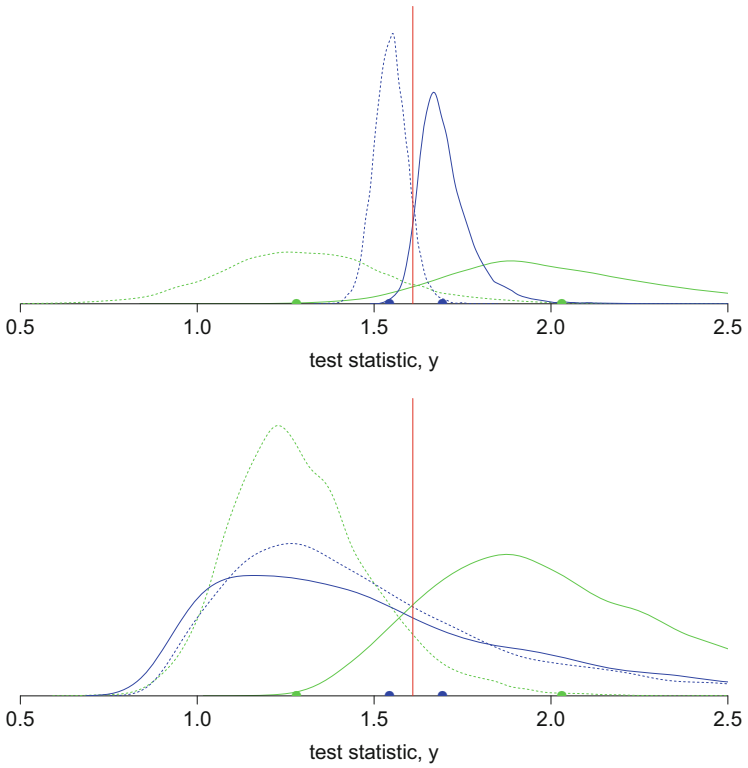


Fig. 4 The distributions of the test statistic y for censoring time $t_C = t_L = 5$ (upper graphs), $t_C = t_l/2 = 2.5$ (lower graphs), $\sigma = 1$ (green), $\sigma = 0.5$ (blue), $p_1 = 0.1$ (solid) and $p_2 = 0.2$ (dashed) obtained by 10^4 simulation runs. The expected values of the test statistic are indicated as points on the horizontal axis. The specification limit $x_L = 1.61(t_L = 5)$ is indicated as red vertical line

5 A Graphical Approach

The cumulative distribution function of the Weibull distribution is

$$F = 1 - \exp\left(-\left(\frac{t}{\tau}\right)^{\delta}\right). \tag{24}$$

By taking twice the logarithm of $1 - F$ we get

$$\ln(-\ln(1 - F)) = \delta(\ln t - \ln \tau). \tag{25}$$

This equation relates $\ln(-\ln(1 - F))$ linearly to $\ln t$. Hence, in a coordinate system with a logarithmic horizontal axis for t and a vertical axis according to $\ln(-\ln(1-F)$ for F the cumulative distribution function of any Weibull distribution is represented as a straight line. The slope of this straight line is equal to the parameter δ and

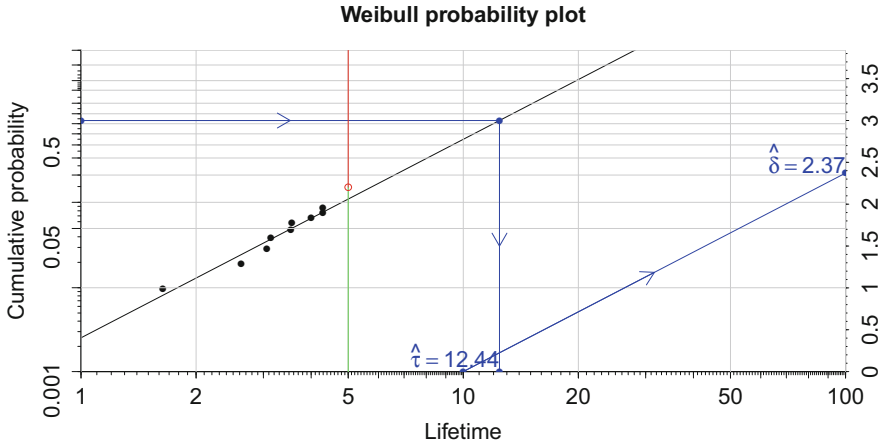


Fig. 5 In this particular example of the application of our sampling plan ($n = 103, k = 1.83, p_{crit} = 0.148, t_L = 5, t_C = 5$) 9 lifetimes $t_{(1)}, \ldots, t_{(9)}$ have been observed and are plotted against $1/(n+1), \ldots, 9/(n+1)$ (black points). The "best fit" straight line (black) intersects with the vertical line through $t_L = 5$ in the green part for which the estimate $\hat{p}$ is smaller than p_{crit} and hence, the lot is accepted (the blue lines demonstrate how the parameters of the Weibull distribution can be estimated graphically)

the parameter τ is the lifetime t for which the cumulative distribution is equal to $1-\exp(-1) = 0.632$. Graph chapter with such a coordinate system exists as Weibull probability chapter.

We can use the Weibull probability chapter for the application of the sampling plans based on the Weibull distribution (but not for its design). We plot the points $(t_{(i)}, \mathbb{E}(F_T(t_{(i)})) = i/(n+1))$ and draw a "best fit" straight line through these points. At the intersection of this straight line with the vertical line through the specification limit t_L we can read an estimate $\hat{p}$ of the fraction of nonconforming items in the lot. If $\hat{p}$ is not larger than the critical fraction p_{crit} given by the sampling plan we accept the lot. Figure 5 shows a particular example of the application of our sampling plan ($n = 103, k = 1.83, p_{crit} = 0.148, t_L = 5, t_C = 5$). 9 lifetimes $t_{(1)}, \ldots, t_{(9)}$ have been observed and are plotted against $1/(n+1), \ldots, 9/(n+1)$ (black points). The "best fit" straight line (black) intersects with the vertical line through $t_L = 5$ in the green part for which the estimate $\hat{p}$ is smaller than p_{crit} and hence, the lot is accepted. If the intersection were in the red part of the vertical line $\hat{p}$ were larger than p_{crit} and the lot would be rejected.

In our simulation experiment we have used the graphical procedure parallel to the numerical procedure of Sect. 3. The dashed curves of Fig. 1 are the simulated OC curves of the graphical procedure corresponding to the solid curves of the numerical procedure. The OC curves are a little more flat, i.e. the graphical procedure is slightly less efficient. However, the graphical procedure depends on the visually fitted straight line and this fit might cause dispute if the intersection with the vertical line is close to the critical value p_{crit}.

6 Conclusions

The lifetime (time to failure) of a product is modeled as Weibull distributed (with unknown parameters); in this case the logarithms of the lifetimes are Gumbel distributed. Lots of items shall be accepted if their fraction p of nonconforming items (items the lifetime of which is smaller than a lower specification limit t_L) is not larger than a specified acceptable quality limit. The acceptance decision is based on the $r \leq n$ observed lifetimes of a sample of size n which is put under test until a defined censoring time t_C is reached (Type I censoring). A lot is accepted if $r = 0$ or if the test statistic $y = \hat{\mu} - k\hat{\sigma}$ is not smaller than the logarithm of the specification limit, $x_L = \log(t_L)$, where k is an acceptance factor and $\hat{\mu}$ and $\hat{\sigma}$ are the Maximum Likelihood estimates of the parameters of the Gumbel distribution. The parameters of the sampling plan (acceptance factor k, sample size n and censoring time t_C) are derived so that lots with $p \leq p_1$ shall be accepted with probability not smaller than $1 - \alpha$. On the other hand, lots with fractions nonconforming larger than a specified value p_2 shall be accepted with probability not larger than β. n and t_C are not obtained separately but as a function that relates the sample size n to the censoring time t_C. Of course, n decreases if the censoring time t_C is increased. For $t_C \to \infty$ the smallest sample size, i.e. that of the uncensored sample, is obtained. Unfortunately, the parameters of the sampling plan do not only depend on the two specified points of the OC, $P_1(p_1, 1 - \alpha)$ and $P_2(p_2, \beta)$, but directly on the parameters τ and δ of the underlying Weibull distribution or equivalently, on the parameters $\mu = \log(\tau)$ and $\sigma = 1/\delta$ of the corresponding Gumbel distribution. Since these parameters are unknown we assume that the hazard rate of the underlying Weibull distribution is nondecreasing ($\delta \geq 1$). For the design of the sampling plan we use the limiting case $\delta = 1$ or $\sigma = 1/\delta = 1$. A simulation study shows that the OC of the sampling plan is almost independent of σ if the censoring time t_C is not smaller than the specification limit t_L.

If the censoring time t_C is chosen smaller than the specification limit t_L then the sample size of the sampling plan is rather large, if $t_C = t_L$ the sample size is not much smaller than the sample size of the corresponding attributes sampling plan, whereas for t_C larger than t_L the sample size is, e.g. for $t_C = 2t_L$, about 10–30% smaller than that of the corresponding attributes sampling plan.

Annex A: Maximum Likelihood Estimation of the Parameters of the Gumbel Distribution

r lifetimes $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(r)}$ (assumed to be Weibull distributed) are observed in a life test with n items put on test and the test finished at time t_C (Type I censoring to the right); all $n - r$ unobserved lifetimes $t_{(r+1)} \leq t_{(r+2)} \leq \ldots \leq t_{(n)}$ are larger than t_C; $r = 0, 1, \ldots, n$ is a random variable.

We transform the lifetimes $t_{(i)}$ to $x_i = \ln t_{(i)}$. The likelihood function of the sample is

$$L(\mu,\sigma) = \prod_{i=1}^{r} f_X(x_i) \cdot G_X^{n-r}(x_C) = \frac{1}{\sigma^r} \prod_{i=1}^{r} f_Z(z_i) \cdot G_Z^{n-r}(z_C)$$

$$= \frac{1}{\sigma^r} \prod_{i=1}^{r} \exp(z_i - \exp(z_i))(\exp(-\exp(z_C)))^{n-r} \tag{26}$$

with $z_i = (x_i - \mu)/\sigma$ and $z_C = (x_C - \mu)/\sigma$. The loglikelihood function is

$$l(\mu,\sigma) = -r \ln \sigma + \sum_{i=1}^{r} (z_i - \exp(z_i)) - (n-r)\exp(z_C). \tag{27}$$

With $\partial z_i/\partial \mu = -1/\sigma$ and $\partial z_i/\partial \sigma = -x_i/\sigma^2$ we obtain the first derivatives of the loglikelihood as

$$\frac{\partial l(\mu,\sigma)}{\partial \mu} = -\frac{1}{\sigma}\left[\sum_{i=1}^{r}(1 - \exp(z_i)) - (n-r)\exp(z_C)\right] \tag{28}$$

$$= -\frac{1}{\sigma}\left[r - \exp(-\mu/\sigma)\left(\sum_{i=1}^{r}\exp(x_i/\sigma) + (n-r)\exp(x_C/\sigma)\right)\right]$$

and

$$\frac{\partial l(\mu,\sigma)}{\partial \sigma} = -\frac{r}{\sigma} - \frac{1}{\sigma^2}\left[\sum_{i=1}^{r}(x_i - x_i\exp(z_i)) - (n-r)x_C\exp(z_C)\right] \tag{29}$$

$$= -\frac{r}{\sigma} - \frac{\sum_{i=1}^{r} x_i}{\sigma^2} - \frac{\exp(-\mu/\sigma)}{\sigma^2}\left[\sum_{i=1}^{r} x_i \exp(x_i/\sigma) + (n-r)x_C\exp(x_C/\sigma)\right].$$

The Maximum Likelihood estimates are the roots of the equations $\frac{\partial l(\mu,\sigma)}{\partial \mu} = 0$ and $\frac{\partial l(\mu,\sigma)}{\partial \sigma} = 0$. With (28) and (29) we find

$$\exp(-\hat{\mu}/\hat{\sigma}) = \frac{r}{\sum_{i=1}^{r}\exp(x_i/\hat{\sigma}) + (n-r)\exp(x_C/\hat{\sigma})} \tag{30}$$

$$\exp(-\hat{\mu}/\hat{\sigma}) = \frac{r\hat{\sigma} + \sum_{i=1}^{r} x_i}{\sum_{i=1}^{r} x_i \exp(x_i/\hat{\sigma}) + (n-r)x_C\exp(x_C/\hat{\sigma})}, \tag{31}$$

respectively, and by Eqs. (30) and (31) we obtain a nonlinear equation for the determination of $\hat{\sigma}$:

$$\hat{\sigma} + \frac{\sum_{i=1}^{r} x_i}{r} - \frac{\sum_{i=1}^{r} x_i \exp(x_i/\hat{\sigma}) + (n-r)x_C \exp(x_C/\hat{\sigma})}{\sum_{i=1}^{r} \exp(x_i/\hat{\sigma}) + (n-r)\exp(x_C/\hat{\sigma})} = 0. \tag{32}$$

From (30) we finally obtain

$$\hat{\mu} = -\hat{\sigma} \ln\left(\frac{r}{\sum_{i=1}^{r} \exp(x_i/\hat{\sigma}) + (n-r)\exp(x_C/\hat{\sigma})}\right). \tag{33}$$

It shall be noted that the estimation of the parameters is not possible if $r = 0$ (no lifetime observed).

Annex B: The Variance of the Test Statistic $y = \hat{\mu} - k\hat{\sigma}$

We write the likelihood of a single observation $z = (x-\mu)/\sigma = (\ln t - \mu)/\sigma$ as

$$L(\mu, \sigma) = f_Z^I(z) G_Z^{1-I}(z) \tag{34}$$

where

$$I = \begin{cases} 1 & : \quad z \le z_C \\ 0 & : \quad z > z_C. \end{cases} \tag{35}$$

indicates that z is observed. The loglikelihood is

$$l = l(\mu, \sigma) = I(-\ln\sigma + z - \exp(z)) - (1-I)\exp(z_C)$$

With $\partial z/\partial\mu = -1/\sigma$ and $\partial z/\partial\sigma = -z/\sigma$ the first partial derivatives of l become

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma}\left[I(1-\exp(z)) - (1-I)\exp(z_C)\right] \tag{36}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{1}{\sigma}\left[I(1+z-\exp(z)) - (1-I)\exp(z_C)\right]. \tag{37}$$

The second derivatives of l are

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{1}{\sigma^2}\left[I\exp(z) + (1-I)\exp(z_C)\right] \tag{38}$$

$$\frac{\partial^2 l}{\partial \mu \partial \sigma} = -\frac{1}{\sigma^2}\left[I(-1+\exp(z)) + (1-I)\exp(z_C)\right]$$

$$
\begin{aligned}
&-\frac{1}{\sigma^2}\left[Iz\exp(z)+(1-I)z_C\exp(z_C)\right]\\
&=-\frac{1}{\sigma^2}\left[-\sigma\frac{\partial l}{\partial\mu}+Iz\exp(z)+(1-I)z_C\exp(z_C)\right] \qquad (39)\\
\frac{\partial^2 l}{\partial\sigma^2}&=-\frac{1}{\sigma^2}\left[I(1+z-z\exp(z))-(1-I)z_C\exp(z_C)\right]\\
&\quad-\frac{1}{\sigma^2}\left[I(-1-z+z\exp(z)+1+z^2\exp(z))-(1-I)(-z_C\exp(z_C)-z_C^2\exp(z_C))\right]\\
&=-\frac{1}{\sigma^2}\left[-2\sigma\frac{\partial l}{\partial\sigma}+I(1+z^2\exp(z))+(1-I)z_C^2\exp(z_C)\right] \qquad (40)
\end{aligned}
$$

The expectations of the second derivatives are, with $\mathbb{E}(I)=P(Z\le z_C)=F_Z(z_C)$, $\mathbb{E}(\frac{\partial l}{\partial\mu})=0$ and $\mathbb{E}(\frac{\partial l}{\partial\sigma})=0$:

$$
\mathbb{E}\left(\frac{\partial^2 l}{\partial\mu^2}\right)=-\frac{1}{\sigma^2}\left[\int_{-\infty}^{z_C}\exp(z)f_Z(z)dz+(1-F_Z(z_C))\exp(z_C)\right]. \qquad (41)
$$

By partial integration with $u=\exp(z)$, $v'=f_Z(z)$, $u'=\exp(z)$, $v=F_Z(z)$ and $u'v=\exp(z)F_Z(z)=\exp(z)-\exp(z)G_Z(z)=\exp(z)-f_Z(z)$ we obtain

$$
\begin{aligned}
\int_{-\infty}^{z_C}\exp(z)f_Z(z)dz&=\exp(z_C)F_Z(z_C)-\int_{-\infty}^{z_C}(\exp(z)-f_Z(z))\,dz\\
&=-(1-F_Z(z_C)\exp(z_C)+F_Z(z_C)
\end{aligned}
$$

and hence,

$$
\mathbb{E}\left(\frac{\partial^2 l}{\partial\mu^2}\right)=-\frac{1}{\sigma^2}F_Z(z_C)=-\frac{1}{\sigma^2}f_{11}. \qquad (42)
$$

For $z_C\to\infty$ we have $f_{11}\to f_{11,\infty}=1$.

$$
\mathbb{E}\left(\frac{\partial^2 l}{\partial\mu\partial\sigma}\right)=-\frac{1}{\sigma^2}\left[\int_{-\infty}^{z_C}z\exp(z)f_Z(z)dz+(1-F_Z(z_C))z_C\exp(z_C)\right].
$$

By partial integration with $u=z\exp(z)$, $v'=f_Z(z)$, $u'=(1+z)\exp(z)$, $v=F_Z(z)$ and $u'v=(1+z)\exp(z)F_Z(z)=(1+z)\exp(z)-(1+z)\exp(z)G_Z(z)=(1+z)\exp(z)-(1+z)f_Z(z)$ we obtain

$$
\int_{-\infty}^{z_C}z\exp(z)f_Z(z)dz=z_C\exp(z_C)F_Z(z_C)-\int_{-\infty}^{z_C}(1+z)\,(\exp(z)-f_Z(z))\,dz,
$$

$$
\int_{-\infty}^{z_C}(1+z)\,(\exp(z)-f_Z(z))\,dz=\underbrace{\int_{-\infty}^{z_C}(1+z)\exp(z)dz}_{J_1}+\underbrace{\int_{-\infty}^{z_C}(1+z)f_Z(z)dz}_{J_2}
$$

$$J_1 = \exp(z_C) + z_C \exp(z_C) - \exp(z_C)$$
$$J_2 = -z_C \exp(z_C)(1 - F_Z(z_C)) + \int_{-\infty}^{z_C} (1+z) f_Z(z) dz$$
$$\Longrightarrow$$
$$\int_{-\infty}^{z_C} (1+z)(\exp(z - f_Z(z)) dz = z_C \exp(z_C) F_Z(z_C) + \int_{-\infty}^{z_C} (1+z) f_Z(z) dz$$

and hence,

$$\mathbb{E}\left(\frac{\partial^2 l}{\partial \mu \partial \sigma}\right) = -\frac{1}{\sigma^2} \int_{-\infty}^{z_C} (1+z) f_Z(z) dz = -\frac{1}{\sigma^2} f_{12}. \tag{43}$$

With the substitution $u = \exp(z)$, f_{12} becomes for $z_C \to \infty$

$$f_{12,\infty} = \int_{-\infty}^{\infty} (1+z) f_Z(z) dz = 1 + \int_0^{\infty} \ln u \exp(-u] du = 1 - \gamma \tag{44}$$

where $\gamma = 0.57721566490\ldots$ is Euler's constant (see Erdélyi (1954), p. 148).

$$\begin{aligned}\mathbb{E}\left(\frac{\partial^2 l}{\partial \sigma^2}\right) &= -\frac{1}{\sigma^2}\left[\int_{-\infty}^{z_C} (1 + z^2 \exp(z)) f_Z(z) dz + (1 - F_Z(z_C)) z_C^2 \exp(z_C)\right] \\ &= -\frac{1}{\sigma^2}\left[\int_{-\infty}^{z_C} f_Z(z) dz + \int_{-\infty}^{z_C} z^2 \exp(z) f_Z(z) dz + (1 - F_Z(z_C)) z_C^2 \exp(z_C)\right]\end{aligned}$$

By partial integration with $u = z^2 \exp(z)$, $v' = f_Z(z)$, $u' = (2z + z^2)\exp(z)$, $v = F_Z(z)$ and $u'v = (2z+z^2)\exp(z)F_Z(z) = (2z+z^2)\exp(z) - (2z+z^2)\exp(z)G_Z(z) = (2z + z^2)\exp(z) - (2z + z^2) f_Z(z)$ we obtain

$$\begin{aligned}\int_{-\infty}^{z_C} z^2 \exp(z) f_Z(z) dz &= z_C^2 \exp(z_C) F_Z(z_C) - \int_{-\infty}^{z_C} (2z + z^2)\exp(z) F_Z(z) dz, \\ &= z_C^2 \exp(z_C) F_Z(z_C) - \underbrace{\int_{-\infty}^{z_C} (2z + z^2)\exp(z) dz}_{J_3} + \int_{-\infty}^{z_C} (2z + z^2) f_Z(z) dz,\end{aligned}$$
$$J_3 = 2\int_{-\infty}^{z_C} z \exp(z) dz + z_C^2 \exp(z_C) - 2\int_{-\infty}^{z_C} z \exp(z) dz = z_C^2 \exp(z_C)$$
$$\Longrightarrow$$
$$\int_{-\infty}^{z_C} z^2 \exp(z_C) f_Z(z) dz = -(1 - F_Z(z_C)) z_C^2 \exp(z_C) + \int_{-\infty}^{z_C} (2z + z^2) f_Z(z) dz$$

and hence,

$$\mathbb{E}\left(\frac{\partial^2 l}{\partial \sigma^2}\right) = -\frac{1}{\sigma^2} \int_{-\infty}^{z_C} (1+z)^2 f_Z(z) dz = -\frac{1}{\sigma^2} f_{22}. \tag{45}$$

With the substitution $u = \exp(z)$, f_{22} becomes for $z_C \to \infty$

$$f_{22,\infty} = 1+2\int_{-\infty}^{\infty} z f_Z(z)dz + \int_{-\infty}^{\infty} z^2 f_Z(z)dz = 1+2\underbrace{\int_0^{\infty} \ln u \exp(-u)du}_{J_4} + \underbrace{\int_0^{\infty} \ln^2 u \exp(-u)du}_{J_5}.$$

With $J_4 = -\gamma$ according to (43) and $J_5 = \gamma^2 + \frac{\pi^2}{6}$ (see Erdélyi (1954), p. 149) we get $f_{22,\infty} = (1-\gamma)^2 + \frac{\pi^2}{6}$.

The equations for f_{11}, f_{12} and f_{22} in (42), (43) and (45) are equivalent to equations derived in Harter and Moore (1968). The integrals in (43) and (45) cannot be solved directly. Escobar and Meeker (1986) present series expansions

$$\begin{aligned} f_{12} &= F_Z(z_C) + \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j!} \left(z_C - \frac{1}{j}\right) (\exp(z_C))^j \\ f_{22} &= 2f_{12} - F_Z(z_C) + \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j!} \left(\left(z_C - \frac{1}{j}\right)^2 + \frac{1}{j^2}\right) (\exp(z_C))^j \end{aligned} \quad (46)$$

and recommend to use these series expansions if $z_C < 0$ and to split the integrals into

$$\begin{aligned} \int_{-\infty}^{z_C} (1+z) f_Z(z)dz &= \int_{-\infty}^{1} (1+z) f_Z(z)dz + \int_1^{z_C} (1+z) f_Z(z)dz \\ &= 0.2720757938345342 + \int_1^{z_C} (1+z) f_Z(z)dz \\ \int_{-\infty}^{z_C} (1+z)^2 f_Z(z)dz\, n &= \int_{-\infty}^{1} (1+z)^2 f_Z(z)dz + \int_1^{z_C} (1+z)^2 f_Z(z)dz \\ &= 1.475933122158450 + \int_1^{z_C} (1+z)^2 f_Z(z)dz \end{aligned} \quad (47)$$

for $z_C \geq 1$ and to calculate the integrals on the right hand side by numerical integration.

The Fisher information matrix of a sample of size n is

$$\mathbf{F} = -n \begin{pmatrix} \mathbb{E}\left(\frac{\partial^2 l}{\partial \mu^2}\right) & \mathbb{E}\left(\frac{\partial^2 l}{\partial \mu \partial \sigma}\right) \\ \mathbb{E}\left(\frac{\partial^2 l}{\partial \mu \partial \sigma}\right) & \mathbb{E}\left(\frac{\partial^2 l}{\partial \sigma^2}\right) \end{pmatrix} = \frac{n}{\sigma^2} \begin{pmatrix} f_{11} & f_{12} \\ f_{12} & f_{22} \end{pmatrix} \quad (48)$$

with f_{11}, f_{12}, f_{22} according to (42), (43), (45), respectively.

The asymptotic covariance matrix of the estimators $\hat{\mu}$ and $\hat{\sigma}$ is the inverse of the Fisher information matrix,

$$\mathbf{V} = \begin{pmatrix} \sigma_{\hat{\mu}}^2 & \sigma_{\hat{\mu}\hat{\sigma}} \\ \sigma_{\hat{\mu}\hat{\sigma}} & \sigma_{\hat{\sigma}}^2 \end{pmatrix} = \mathbf{F}^{-1} = \frac{\sigma^2}{n} \begin{pmatrix} f_{11}\, f_{12} \\ f_{12}\, f_{22} \end{pmatrix}^{-1} = \frac{\sigma^2}{n} \begin{pmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{pmatrix}. \tag{49}$$

We note that the inverse $(f_{ij})^{-1} = (v_{ij})$ only depends on the standardized censoring time $z_C = (x_C - \mu)/\sigma$.

For $z_C \to \infty$ the Fisher information matrix is

$$\mathbf{F}_\infty = \frac{n}{\sigma^2} \begin{pmatrix} f_{11,\infty}\, f_{12,\infty} \\ f_{12,\infty}\, f_{22,\infty} \end{pmatrix} = \frac{n}{\sigma^2} \begin{pmatrix} 1 & 1-\gamma \\ 1-\gamma & (1-\gamma)^2 + \frac{\pi^2}{6} \end{pmatrix}, \tag{50}$$

and the asymptotic covariance matrix is

$$\mathbf{V}_\infty = {\mathbf{F}_\infty}^{-1} = \frac{\sigma^2}{n} \cdot \frac{6}{\pi^2} \begin{pmatrix} (1-\gamma)^2 + \frac{\pi^2}{6} & \gamma - 1 \\ \gamma - 1 & 1 \end{pmatrix}. \tag{51}$$

The asymptotic variance of the test statistic $y = \hat{\mu} - k\hat{\sigma}$ becomes

$$\sigma_y^2 = \sigma_{\hat{\mu}}^2 + k^2\sigma_{\hat{\sigma}}^2 - 2k\sigma_{\hat{\mu}\hat{\sigma}} = \frac{\sigma^2}{n}\left(v_{11} + k^2 v_{22} - 2k v_{12}\right) = \sigma^2 A^2 \tag{52}$$

with

$$A = \frac{\sigma_y}{\sigma} = \frac{\sqrt{v_{11} + k^2 v_{22} - 2k v_{12}}}{\sqrt{n}} = \frac{f(k, z_C)}{\sqrt{n}} \tag{53}$$

where the numerator $f(k, z_C)$ only depends on the acceptance factor k and the standardized censoring time z_C, and the denominator only on the sample size n. For $z_C \to \infty$ we get

$$A = \frac{\sigma_y}{\sigma} = \frac{\sqrt{\frac{6}{\pi^2}\left((1-\gamma)^2 + \frac{\pi^2}{6} + k^2 + 2k(1-\gamma)\right)}}{\sqrt{n_{min}}} = \frac{f(k)}{\sqrt{n_{min}}} \tag{54}$$

References

Erdélyi, A. (Ed.). (1954). *Tables of Integral Transforms* (Vol. I). London/New York: McGraw-Hill.

Escobar, L. A., & Meeker, W. Q. (1986). Elements of the fisher information matrix for the smallest extreme value distribution and censored data. *Journal of the Royal Statistical Society Series C (Applied Statistics), 35*, 80–86.

Fertig, K. W., & Mann, N. R. (1980). Life-test sampling plans for two-parameter Weibull populations. *Technometrics, 22*, 165–177.

Goode, H. P., & Kao, J. H. K. (1961). Sampling plans based on the Weibull distribution. In *Proceedings of the Seventh National Symposium on Reliability and Quality Control* (pp. 24–40).

Goode, H. P., & Kao, J. H. K. (1962). Sampling procedures and tables for life and reliability testing based on the Weibull distribution (Hazard Rate Criterion). In *Proceedings of the Eighth National Symposium on Reliability and Quality Control* (pp. 37–58).

Goode, H. P., & Kao, J. H. K. (1963). Weibull tables for bio-assaying and fatigue testing. In *Proceedings of the Ninth National Symposium on Reliability and Quality Control* (pp. 270–286).

Harter, H. L., & Moore, A. H. (1968). Maximum-Likelihood estimation, from doubly censored samples, of the parameters of the first asymptotic distribution of extreme values. *Journal of the American Statistical Association, 63*, 889–901.

Hosono, Y., Ohta, H., & Kase, S. (1981). Design of single sampling plans for doubly exponential characteristics. In H. J. Lenz, G. B. Wetherill, & P. T. Wilrich (Eds.), *Frontiers in Statistical Quality Control*. Würzburg: Physica-Verlag.

ISO 3951-1. (2005). *Sampling procedures for inspection by variables – Part 1: Specification for Single Sampling Plans Indexed by Acceptance Quality Limit (AQL) for Lot-by-Lot Inspection – Single Quality Characteristic and Single AQL.* Geneva: International Standardization Organization.

ISO 3951-2. (2005). *Sampling Procedures for Inspection by Variables – Part 2: General Specification for Single Sampling Plans Indexed by Acceptance Quality Limit (AQL) for Lot-by-Lot Inspection of Independent Quality Characteristics.* Geneva: International Standardization Organization.

Quality Control and Reliability Technical Report TR 3. (1961). *Sampling Procedures and Tables for Life and Reliability Testing Based on the Weibull distribution (Mean Life Criterion).* Office of the Assistant Secretary of Defense (Installations and Logistics), U.S. Government Printing Office, USA.

Quality Control and Reliability Technical Report TR 4. (1962). *Sampling Procedures and Tables for Life and Reliability Testing Based on the Weibull distribution (Hazard Rate Criterion).* Office of the Assistant Secretary of Defense (Installations and Logistics), U.S. Government Printing Office, USA.

Quality Control and Reliability Technical Report TR 6. (1963). *Sampling Procedures and Tables for Life and Reliability Testing Based on the Weibull distribution (Reliable Life Criterion).* Office of the Assistant Secretary of Defense (Installations and Logistics), U.S. Government Printing Office, USA.

Quality Control and Reliability Technical Report TR 7. (1965). *Factors and Procedures for Applying MIL STD-105D Sampling Plans to Life and Reliability Testing.* Office of the Assistant Secretary of Defense (Installations and Logistics), U.S. Government Printing Office, USA.

Schneider, H. (1989). Failure-censored variables-sampling plans for lognormal and Weibull distributions. *Technometrics, 31*, 199–206.

Approximate Log-Linear Cumulative Exposure Time Scale Model by Joint Moment Generating Function of Covariates

Watalu Yamamoto and Lu Jin

Abstract Online monitoring data contain various measurements of system activity. The amount of work resulting from system activity is also measured in various ways. When we model the reliability of a system, i.e., the intensity or risk of failure events, we need to choose a time scale. Though there should be genuine time scales for each failure phenomenon, the field data, including online monitoring data, may not provide evidence for them. There are many uncontrollable factors in the field. Many variables increase monotonically and are highly correlated with each other within a system. Yet they also represent the differences among systems. This article attempts to build a bridge between two useful approaches, alternative time scale and cumulative exposure model, by assuming the stationarity of the increments in these measurements within a system.

Keywords Cumulative exposure model · Accelerated failure-time model · Approximation · Moment generating function

1 Time Scale Models

When it is natural to model the failure time distribution of a system with the age on the chronological time, say T_0, we call it as the time scale for the failure time. Following the tradition of reliability engineering, we call the target subjects observed failure times and covariates as systems. The calendar time is a typical time scale. But it is not the only one. The total operating time, T_1, may be the time scale, if the total hours of operation vary among systems. The total usage amount, T_2,

W. Yamamoto (✉) · L. Jin
University of Electro-Communications, Tokyo, Japan
e-mail: watalu@inf.uec.ac.jp; jinlu@inf.uec.ac.jp

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_18

on a work amount scale may be the time scale, if the contents of operation vary among systems. The examples of total usage amounts include the total mileage of an automobile, and the number of sheets printed by a copier. The total operating time is also an example of the total usage amount.

Let us assume that we observe a set of variables, T_0, T_1, ..., T_p, each of which measures the failure time on a candidate time scale. T_0 is the chronological failure time. We conduct statistical inference of either the choice among them or the synthesis of them, based on the observed data. If the engineering knowledge suggests that T_k is suitable for modeling the failure time with a probability distribution $F(t)$, then a series of goodness-of-fit tests or a statistical comparison of the estimates of Kullback-Leibler divergences between the data on individual time scales and the failure time distribution allows us to check that suggestion. If more than one time scale could be bases of the time scale, then the synthesis is investigated using the time scale models.

Farewell and Cox (1975) were possibly the first authors to investigate combining multiple time scales to obtain a more suitable time scale in the context of life testing. The problem of time scales was also investigated by Kordonsky and Gertsbakh (1993, 1995a,b, 1997). They considered the so called linear time scale model,

$$U_L = \beta_0 T_0 + \beta_1 T_1 + \cdots + \beta_p T_p, \tag{1}$$

and investigated properties by estimating parameters with a minimum coefficient of variation. The choice of the estimating criterion was made because it is scale invariant. In their studies, the parameter space was

$$\Theta_L = \left\{ \boldsymbol{\beta};\ \beta_k \geq 0,\ k = 0, \ldots, p \text{ and } \sum_k \beta_k = 1 \right\}. \tag{2}$$

There is also another time scale model,

$$U_M = {T_0}^{\beta_0} {T_1}^{\beta_1} \cdots {T_p}^{\beta_p},$$

called the multiplicative time scale model.

Duchesne and Lawless (2000) called these models as alternative time scales. Both models coincide with T_k if $\beta_k = 1$ and $\beta_{k'} = 0$ for $k' \neq k$ hold.

In the study of time scales, it is assumed that the random quantity synthesized with a time scale model, U_L or U_M, is distributed with a common failure time distribution. This assumption holds under the condition of collapsibility, proposed by Oakes (1995). A time scale model is collapsible if the probability to survive at a point $\left(T_0, T_1, \ldots, T_p\right)$ in the time space depends only on that point and does not depend on the life path $\mathcal{H}_{T_0} = \left\{\left(t, T_1(t), \ldots, T_p(t)\right); 0 \leq t \leq T_0\right\}$ to that point. Under this condition, the endpoints $\left(T_0, T_1, \ldots, T_p\right)$ of individual paths are sufficient for estimating parameters of time scale models. Duchesne and Lawless (2002) propose a semiparametric estimator of the parameters under the collapsibility

condition using the theory of estimating functions, which doesn't need to assume a specific distribution of the time scale.

As for failure time data, the endpoint of the life path, $\left(T_0, T_1, \ldots, T_p\right)$, is the only available record. Recently, there have been studies on assessing the reliability of products under continuous on-line surveillance, which allows us to observe the path $\mathcal{H}_{T_0}$ up to the failure. Hong and Meeker (2013) proposed to use Nelson's cumulative damage model to model the effect of use rate variation onto the failure-time of a product and predict the failure-time distribution by estimating the use-rate process. Use rates are the derivatives $dT_k(t)/dt$, $k = 1, \ldots, p$, of individual time variable $T_k(t)$, $k = 1, \ldots, p$, with respect to the chronological time t. Hong et al. (2015) modeled a physical degradation process using the dynamic measurements of the environmental conditions. They applied a smoothing regression technique to estimate the trends in degradation paths. We believe that the cumulative damage model is also useful for the problem of time scales.

2 Cumulative Exposure Time Scale Model

When the path to the failure time is observed, it is convenient to choose the chronological time as the reference time t. To construct the cumulative exposure time scale, the derivative process $X_k(t) = dT_k(t)/dt$, $k = 1, \ldots, p$ of the lifetime progress serves as covariates of the cumulative exposure model. Furthermore we denote the observed version of $X_k(t)$'s as $x_k(t)$, $k = 1, \ldots, p$.

A general cumulative damage model is specified by a pair of formulas, cumulative damage,

$$u(T) \,|\mathcal{H}_\infty = \int_0^T \mathcal{D}(s; \mathcal{H}_s)\, ds, \tag{3}$$

and distribution on the cumulative damage time scale

$$T \,|\mathcal{H}_\infty \sim F(u(t)), \tag{4}$$

where $\mathcal{H}_t$ is the history of the covariate process up to time t. This model is a dynamic scale-accelerated failure-time model with covariate process. $\mathcal{D}(t; \mathcal{H}_t)$ is the speed of the progress of time at t. This speed can be affected by the covariate process. If the sample path is linear in t, $\mathcal{D}(t; \mathcal{H}_t)$ is constant through the time. In that case, this model results in the scale accelerated failure time model (Escobar and Meeker 2006).

$F(u)$ is usually chosen from log the location scale family which has a density function of the form

$$f(u; \mu, \sigma) = \frac{1}{u\sigma} f_0\left(\frac{\log u - \mu}{\sigma}\right) = \frac{1}{u\sigma} f_0\left(\log\left(\frac{u}{\exp \mu}\right)^{1/\sigma}\right). \tag{5}$$

Bagdonavičius and Nikulin (2001) called scale-shape family. $\mathcal{D}(t; \mathcal{H}_t)$ is defined to be $\mathcal{D}(t; \mathcal{H}_t) \equiv 1$ under the standard constant condition $\boldsymbol{x}(t) \equiv \boldsymbol{x}_0 = (x_{10}, x_{20}, \ldots, x_{p0})$. Then μ and σ are the parameters of the failure-time distribution under the constant condition $\boldsymbol{x}_0$.

Generally the speed of the progress of time $\mathcal{D}(t; \mathcal{H}_t)$ may depend on the history up to t, $\mathcal{H}_t$. However, it is difficult to model in such a flexible manner. Therefore, we restrict ourselves to model failure time data with continuous monitoring as

$$\mathcal{D}(\mathcal{H}_t) \approx \mathcal{D}(\boldsymbol{x}(t)). \tag{6}$$

This model can be used for modeling the time scale under continuous monitoring. $\mathcal{D}(s; \boldsymbol{x}(s))$ can assess how the variations in $x_k(t)$'s affects the failure time.

Since our interest lies in the modeling of time scales, we restrict our attention to cases with dynamic use-rates as covariates, as in Hong and Meeker (2010). A use rate is defined as the increment in some usage variable or a time scale variable per unit reference time. Note that $\int_0^{T_0} ds$ is the failure time on the chronological time scale.

There are two primary choices of parametric models. One is the linear model,

$$\mathcal{D}_L(\boldsymbol{x}(t)) = \beta_0 + \beta_1 x_1(t) + \cdots + \beta_p x_p(t). \tag{7}$$

This model is derived from a general model by approximating with Taylor expansion around the standard condition $\boldsymbol{x}_0$;

$$\mathcal{D}_L(\boldsymbol{x}(t)) \approx \mathcal{D}_L(\boldsymbol{x}_0) + \sum_k (x_k(t) - x_{k0}) \left.\frac{\partial \mathcal{D}}{\partial x_k}\right|_{x_k = x_{k0}}. \tag{8}$$

This model is same as the linear time scale model, U_L. However the parameter space does not need to be positive.

Another typical class of time scales is the *log-linear cumulative exposure model*;

$$\mathcal{D}_M(\boldsymbol{x}(t)) = {x_1}^{\beta_1}(t) \cdots {x_p}^{\beta_p}(t) = \exp\left(\beta_1 \tilde{x}_1(t) + \cdots + \beta_p \tilde{x}_p(t)\right), \tag{9}$$

where $\tilde{x}_k(t) = \log x_k(t)$. The log-linear model is derived from a general model by approximating the logarithm with Taylor expansion around $\boldsymbol{x}_0$;

$$\log \mathcal{D}_M(\boldsymbol{x}(t)) \approx \log \mathcal{D}_L(\boldsymbol{x}_0) + \sum_k (\tilde{x}_k - \tilde{x}_{k0}) \left.\frac{\partial \log \mathcal{D}_M}{\partial \tilde{x}_k}\right|_{\tilde{x}_k = \tilde{x}_{k0}}. \tag{10}$$

Unlike the linear model, the log-linear model is not the same as the multiplicative time scale model. We developed a useful relationship between this model and an accelerated failure time model.

Note that the addition of a constant term to Eq. (9) might cause aliasing of the scale parameter for scale-shape family.

3 Formulas for Maximum Likelihood Estimation

Before going into detail of our proposition, we present the formulas for maximum likelihood estimation of the parameters of cumulative exposure models. Hereafter we denote $\tilde{x}$ as x.

We assume that an online monitoring scheme collects the sample path of its covariate process $\mathcal{X}_{i,\infty}$, time of event t_i, and type of event δ_i, from each system to be monitored. The term $\delta_i = 1$ indicates that the system failed and $\delta_i = 0$ indicates that is was censored. Let $\boldsymbol{x}_i(t)$ be the vector of all variables $x_{i1}(t), \ldots, x_{ip}(t)$ which correspond to the observation of the covariate process on i-th sample. The contribution of each system to the log-likelihood is

$$\begin{aligned}\log L_i = \delta_i \left\{ \boldsymbol{\beta}' \boldsymbol{x}_i(t_i) + \log f\left(u\left(t_i; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right); \boldsymbol{\theta}\right)\right\} \\ + (1-\delta_i) \log\left\{1 - F\left(u\left(t_i; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right); \boldsymbol{\theta}\right)\right\}\end{aligned}$$

where $\boldsymbol{\beta}' \boldsymbol{x}_i(t_i)$ is $\log \partial u\left(t; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right)/\partial t$ evaluated at $t = t_i$. We abbreviate $u\left(t_i; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right)$ as u_i and $u\left(t_i; \hat{\boldsymbol{\beta}} \left| \mathcal{X}_{i,\infty}\right.\right)$ as $\hat{u}_i$.

The score vector consists of

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log L_i = \delta_i \boldsymbol{x}_i(t_i) + \delta_i \frac{\partial u\left(t_i; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right)}{\partial \boldsymbol{\beta}} \left.\frac{\partial}{\partial u} \log f(u; \boldsymbol{\theta})\right|_{u=u_i} \\ + (1-\delta_i) \frac{\partial u\left(t_i; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right)}{\partial \boldsymbol{\beta}} \left.\frac{\partial}{\partial u} \log\{1 - F(u; \boldsymbol{\theta})\}\right|_{u=u_i}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \log L_i = \delta_i \frac{\partial}{\partial \boldsymbol{\theta}} \log f(u_i; \boldsymbol{\theta}) \\ + (1-\delta_i) \frac{\partial}{\partial \boldsymbol{\theta}} \log\{1 - F(u_i; \boldsymbol{\theta})\}.\end{aligned}$$

The observed Fisher information matrix consists of

$$\begin{aligned}\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log L_i = \delta_i \frac{\partial^2 u\left(t_i; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \left.\frac{\partial}{\partial u} \log f(u; \boldsymbol{\theta})\right|_{u=u_i} \\ + (1-\delta_i) \frac{\partial^2 u\left(t_i; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \left.\frac{\partial}{\partial u} \log\{1 - F(u; \boldsymbol{\theta})\}\right|_{u=u_i} \\ + \delta_i \frac{\partial u\left(t_i; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty}\right.\right)}{\partial \boldsymbol{\beta}} \frac{\partial}{\partial \boldsymbol{\beta}^t} \left[\left.\frac{\partial}{\partial u} \log f(u; \boldsymbol{\theta})\right|_{u=u_i} \right]\end{aligned}$$

$$+\delta_i \frac{\partial u\left(t_i;\boldsymbol{\beta}\left|\mathcal{X}_{i,\infty}\right.\right)}{\partial \boldsymbol{\beta}} \frac{\partial}{\partial \boldsymbol{\beta}^t}\left[\frac{\partial}{\partial u}\frac{\partial}{\partial \boldsymbol{\theta}}\log\{1-F(u;\boldsymbol{\theta})\}\Big|_{u=u_i}\right]$$

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log L_i = \delta_i \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(u_i;\boldsymbol{\theta}) + (1-\delta_i)\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log\{1-F(u_i;\boldsymbol{\theta})\}$$

and off diagonal components

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\beta}'}\log L_i = \delta_i \frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{\partial}{\partial u}\log f(u;\boldsymbol{\theta})\Big|_{u=u_i}\right]\frac{\partial u\left(t_i;\boldsymbol{\beta}\left|\mathcal{X}_{i,\infty}\right.\right)}{\partial\boldsymbol{\beta}'} + (1-\delta_i)\frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{\partial}{\partial u}\log\{1-F(u;\boldsymbol{\theta})\}\Big|_{u=u_i}\right]\frac{\partial u\left(t_i;\boldsymbol{\beta}\left|\mathcal{X}_{i,\infty}\right.\right)}{\partial\boldsymbol{\beta}'}. \tag{11}$$

The first and second derivatives with respect to $\boldsymbol{\theta}$ given $\boldsymbol{\beta}$ are readily available on many packages or software programs for log-location scale family, which help us in fitting parametric failure time distributions to the failure data that include censoring. So it is rather straightforward to solve the set of equations

$$\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log L_i\left(\hat{u}_i\right)=\mathbf{0}, \tag{12}$$

where $L_i\left(\hat{u}_i\right)$ is the likelihood contribution with $u=\hat{u}_i$.

However, the derivatives with respect to the components of parameter $\boldsymbol{\beta}$ require numerical integration for each system every time we need to evaluate.

4 Log-Linear Cumulative Exposure Model as Approximate Accelerated Failure Time Model

We discuss further the cumulative exposure model by Hong and Meeker (2013) as a time scale model. We focused on the integral processes of work amounts (or use rates) among many types of covariates. If the covariate process $\mathcal{X}_{i,\infty}$ is stationary,

$$u\left(t;\boldsymbol{\beta}\left|\mathcal{X}_{i,\infty}\right.\right)/t=\frac{1}{t}\int_0^t \exp\left(\boldsymbol{\beta}'\boldsymbol{x}(s)\right)ds \tag{13}$$

is a nonparametric estimate of the joint moment generating function

$$M_{\boldsymbol{X}}(\boldsymbol{\beta}) = E\left[\exp\left(\boldsymbol{\beta}'\boldsymbol{X}(t)\right)\right] \tag{14}$$

with respect to the joint distributions of $\boldsymbol{X}(t)$. Under certain regularity conditions for the existence of the moment generating function, this estimate, also called *the empirical moment generating function*, is consistent.

Csörgő (1980) and Feuerverger (1989) proved that for all β for which $M_X(\beta)$ exists,

$$\sup\left|\hat{M}_X(\beta) - M_X(\beta)\right| = \sup|e_X(\beta)| \to 0 \tag{15}$$

as $t \to \infty$, and that

$$t^{1/2}\left\{\hat{M}_X(\beta) - M_X(\beta)\right\} = t^{1/2} e_X(\beta) \tag{16}$$

converges to Gaussian as $t \to \infty$. From these results,

$$\int_{t_0}^{t_1} \exp\left(\beta_0 X_0(s) + \beta_1 X_1(s) + \cdots + \beta_p X_p(s)\right) ds \to (t_1 - t_0) M_X(\beta) \tag{17}$$

as $t_1 \to \infty$ and $t_1 - t_0 \to \infty$. Thus, the log-linear cumulative exposure is approximated as

$$\int_0^t \exp\left(\boldsymbol{\beta}'\boldsymbol{x}(s)\right) ds \simeq M_X(\boldsymbol{\beta})\, t. \tag{18}$$

Hence, $M_X(\boldsymbol{\beta})$ plays a role of the acceleration factor for the accelerated failure time model as

$$T\left|\mathcal{X}_{i,\infty} \sim F\left(M_X(\boldsymbol{\beta})\, t\right).\right. \tag{19}$$

Once the empirical moment generating function of the covariate process is estimated as $\hat{M}_{\boldsymbol{X}}(\boldsymbol{\beta})$, we can approximate the cumulative exposure as

$$u\left(t; \boldsymbol{\beta}\left|\mathcal{X}_{i,\infty}\right.\right) \simeq \hat{M}_{bmX}(\beta)\, t. \tag{20}$$

This approximation also establishes the relationship between the cumulative exposure model and accelerated failure time model. The empirical moment generating function $\hat{M}_{\boldsymbol{X}}(\boldsymbol{\beta})$ serves as an acceleration factor for the latter.

The marginal distribution is much easier to identify than the joint distribution. For example, if the covariate processes are stationary and are distributed marginally with multivariate normal distribution, we can reduce the amount of calculation for U by substituting the estimates of mean vector $\boldsymbol{\mu}$ and covariance matrix Σ to calculate

the moment generating function. By plugging the estimates of the first two moments $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ into the joint moment generating function, we have the Gaussian type moment generating function as

$$\hat{M}_{\boldsymbol{X}}(\beta) = \exp\left(\hat{\boldsymbol{\mu}}'\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}'\hat{\Sigma}\boldsymbol{\beta}\right). \tag{21}$$

Note that the model by Hong and Meeker (2013) allows us to assess the effects of covariates of a wider class than the class we assume. The covariates and integral processes do not need to be positive or monotone for their purposes.

5 Further Approximations of Empirical Moment Generating Function

The amount of computation required for the evaluation of $\hat{M}_{\boldsymbol{X}}(\boldsymbol{\beta})$ for a given $\boldsymbol{\beta}$ is the same as that for the evaluation of $U\left(t; \boldsymbol{\beta} \left| \mathcal{X}_{i,\infty} \right.\right)$. The estimation of the cumulative exposure model requires the evaluation of this function for each individual product within the online monitoring data. If we want to regularly monitor the changes in fitting of the model, the total amount of computation for this model increases at every moment we receive a new record. Therefore, it is useful to decrease the amount of computation.

The simplest way is Taylor series approximation of the moment generating function. If this function exists, it has the Taylor-series expansion

$$M_X(\boldsymbol{\beta}) = 1 + \boldsymbol{\beta}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\beta}'\left(\boldsymbol{\mu}\boldsymbol{\mu}' + \Sigma\right)\boldsymbol{\beta} + \cdots \tag{22}$$

around the origin of the space of $\boldsymbol{\beta}$. The first order approximation of the moment generating function is

$$\tilde{M}_1(\boldsymbol{\beta}) = 1 + \boldsymbol{\beta}'\boldsymbol{\mu}. \tag{23}$$

A moment estimator of $\boldsymbol{\mu}$ is the vector of the sample means of $x_{ik}(t)$. This approximation holds under the first order stationarity where the expected values of covariates do not depend on time, i.e., $E[\boldsymbol{X}_i(t)] = \boldsymbol{\mu}_i$. Furthermore if the covariate process is a line, then this model coincides with the linear time scale in Eq. (1).

The second order approximation provides another formula

$$\tilde{M}_2(\boldsymbol{\beta}) = 1 + \boldsymbol{\beta}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\beta}'\left(\boldsymbol{\mu}\boldsymbol{\mu}' + \Sigma\right)\boldsymbol{\beta}. \tag{24}$$

This approximation holds under the second order stationarity where the covariance functions as well as autocorrelation functions do not depend on time. Further expansions are also possible.

If the marginal distribution is unimodal and symmetric, an approximation by the normal distribution

$$\tilde{M}_F(\boldsymbol{\beta}) = \exp\left(\boldsymbol{\beta}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}\right) \tag{25}$$

can be considered.

If the covariates are conditionally independent of each other within a system, the joint moment generating function is a product of the moment generating functions of the marginal distributions of each covariate. Therefore, we have another identification by

$$\tilde{M}_{\boldsymbol{X}}(\boldsymbol{\beta}) = \prod_j \tilde{M}_{X_j}(\beta_j). \tag{26}$$

6 Simulation Study

A set of simulations was conducted to compare the approximate cumulative exposure models introduced in the previous section. Two covariates $X_1(t)$ and $X_2(t)$ were continuously observed on each system. The $\log X_1(t)$ and $\log X_1(t)$ were assumed to be distributed with normal distributions with means μ_1 and μ_2, variances 0.8^2, and to be independent with each other. The means μ_1 and μ_2 were also assumed to vary among systems and distributed with normal distributions with means 0, variances 0.5^2, and to be independent with each other. The sample sizes were set as 100 and 200. U is assumed to be distributed with Weibull distribution with shape parameter 4. The simulated failure data were then censored as type II.

We applied four estimation procedures to the same simulated data.

1. Estimation using the original cumulative exposure model.
2. Estimation using the approximate cumulative exposure model with Eq. (23).
3. Estimation using the approximate cumulative exposure model with Eq. (24).
4. Estimation using the approximate cumulative exposure model with Eq. (25).

The resulting sampling distributions are shown in Figs. 1, 2, 3, and 4.

Figures 2 and 3 suggest that both models with Eqs. (23) and (24) cannot reproduce the sampling distributions of the cumulative exposure model, as shown in Fig. 1. The effects of approximation include large variances that have not been affected by censoring. The approximated estimator also has a bias. These results indicate that even though the cumulative exposure time scale model with the delta method described as Eq. (23) is related to the linear time scale in Eq. (1), Taylor approximation may not be a good choice for formulating the acceleration factor in (20).

Figure 4 shows that Eq. (25) leads to smaller variances of parameter estimates than the previous two, though these variances seem less sensitive in the presence

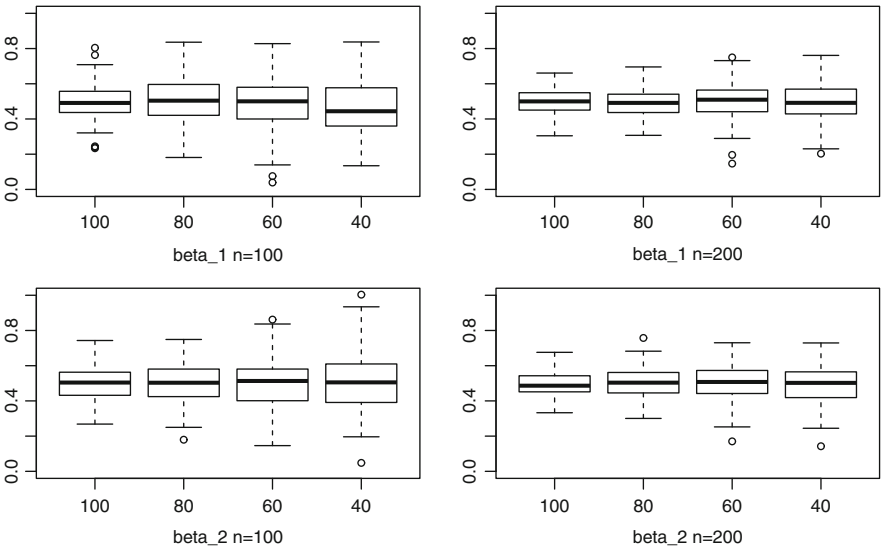


Fig. 1 Boxplots of sampling distributions of parameter estimates using original cumulative exposure model under various censoring rates (horizontal axes indicate censoring ratios and vertical axes indicate estimated values)

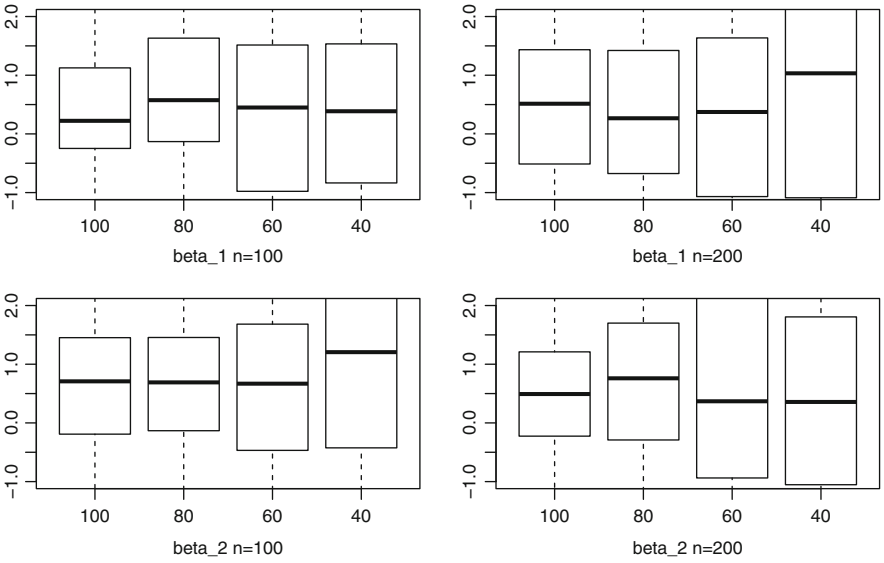


Fig. 2 Boxplots of sampling distributions of parameter estimates using approximate cumulative exposure model (23) under various censoring rates (horizontal axes indicate censoring ratios and vertical axes indicate estimated values)

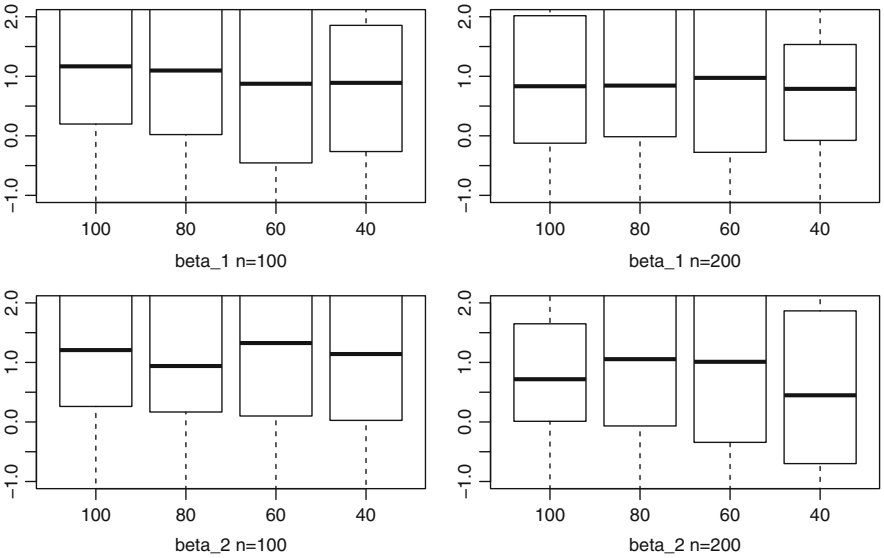


Fig. 3 Boxplots of sampling distributions of parameter estimates using approximate cumulative exposure model (24) under various censoring rates (horizontal axes indicate censoring ratios and vertical axes indicate estimated values)

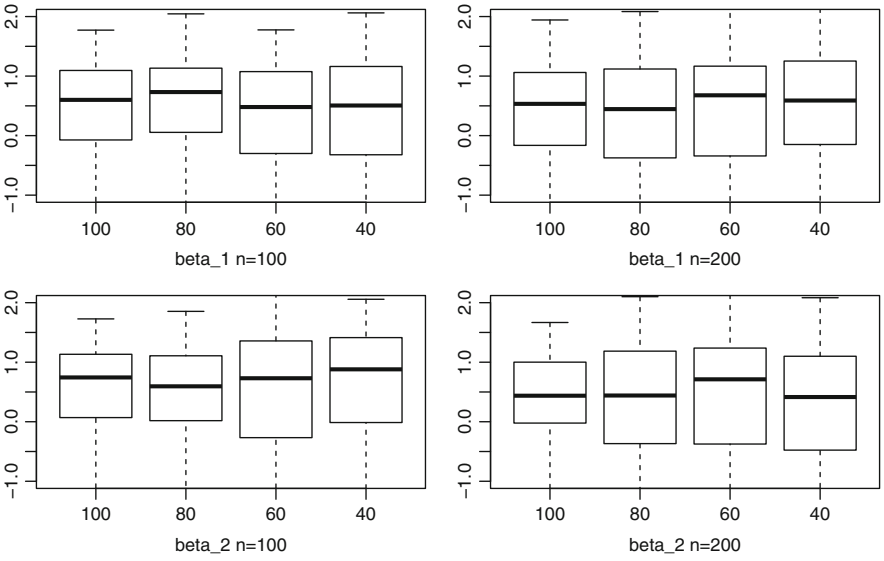


Fig. 4 Boxplots of sampling distributions of parameter estimates using approximate cumulative exposure model (25) under various censoring rates (horizontal axes indicate censoring ratios and vertical axes indicate estimated values)

of censoring again. We would like to investigate the characteristics and properties of the approximation of the cumulative exposure time scale with an accelerated life time model with a parametric moment generating function of the covariate process as the acceleration factor.

7 Remarks

There are other ways of approximation, and we now discuss two of them. One is the combination of a rough grid and multilinear interpolation. By preparing the values of $\hat{M}_{\boldsymbol{X}}(\boldsymbol{\beta})$ for the set of specified points $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$, the multilinear interpolation is obtained as

$$\tilde{M}_L(\boldsymbol{\beta}) = \sum_k N_k \hat{M}_{\boldsymbol{X}}(\boldsymbol{\beta}_k), \tag{27}$$

where N_k is the normalizing constant, which depends on both $\boldsymbol{\beta}$ and the set of points $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p\}$.

The other way is to have a random set of points $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p\}$ and construct multi-dimensional spline interpolation by multiple adaptive regression splines (Friedman 1991) or generalized additive models (Hastie and Tibshirani 2004). Note that though there are many flexible and useful interpolation techniques, they tend to increase the amount of computation.

Acknowledgements This work is partly supported by Grant-in-Aid for Scientific Research (C) No. 15K00042 and No. 25750121 from the Japanese Society for the Promotion of Science. The authors would like to thank the reviewers for their fruitful comments on the early version of this manuscript.

References

Bagdonavičius, V., & Nikulin, M. (2001). *Accelerated Life Models: Modeling and Statistical Analysis*. Boca Raton: Chapman and Hall/CRC.

Csörgő, S. (1980). The empirical moment generating function. In I. Vincze, (Ed.) *Nonparametric Statistical Inference, Budapest: Colloquia Mathematica Societaitis Janos Bolyai* (Vol. 32, pp. 139–150). Amsterdam: North-Holland.

Duchesne, T., & Lawless, J. F. (2000). Alternative time scales and failure time models. *Lifetime Data Analysis, 6*, 157–179.

Duchesne, T., & Lawless, J. F. (2002). Semiparametric inference method for general time scale models. *Lifetime Data Analysis, 8*, 263–276.

Escobar, L. A., & Meeker, W. Q. (2006) A review of accelerated test models. *Statistical Science, 21*, 552–577.

Farewell, V. T., & Cox, D. R. (1975). A note on multiple time scales in life testing. *Applied Statistics, 28*, 115–124.

Feuerverger, A. (1989). On the empirical saddlepoint approximation. *Biometrika, 76*, 457–464.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*, 1–67.
Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*, 297–310.
Hong, Y., Duan, Y., Meeker, W. Q., Stanley, D. L., & Gu, X. (2015). Statistical methods for degradation data with dynamic covariates information and an application to outdoor weathering data. *Technometrics, 57*, 180–193.
Hong, Y., & Meeker, W. Q. (2010). Field-failure and warranty prediction based on auxiliary use-rate information. *Technometrics, 52*, 148–159.
Hong, Y., & Meeker, W. Q. (2013). Field-failure predictions based on failure-time data with dynamic covariate information. *Technometrics, 55*, 135–149.
Kordonsky, K. B., & Gertsbakh, I. (1993). Choice of the best time scale for system reliability analysis. *European Journal of Operational Research, 65*, 235–246.
Kordonsky, K. B., & Gertsbakh, I. (1995a). System state monitoring and lifetime scales – I. *Reliability Engineering and System Safety, 47*, 1–14.
Kordonsky, K. B., & Gertsbakh, I. (1995b). System state monitoring and lifetime scales – II. *Reliability Engineering and System Safety, 49*, 145–154.
Kordonsky, K. B., & Gertsbakh, I. (1997). Multiple time scales and the lifetime coefficient of variation: Engineering applications. *Lifetime Data Analysis, 2*, 139–156.
Oakes, D. (1995). Multiple time scales in survival analysis. *Lifetime Data Analysis, 1*, 7–18.

A Critique of Bayesian Approaches within Quality Improvement

G. Geoffrey Vining

Abstract Bayesian approaches are increasingly popular within the statistics community. However, they currently do not seem to find wide application within the industrial statistics/quality improvement community. This chapter examines some of the basic reasons why. It begins by reviewing Box's perspective on the scientific method and discovery. It then examines Deming's concepts of analytic versus enumerative studies. Together, these concepts provide a framework for evaluating when Bayesian approaches make good sense, where they make little sense, and where they fall somewhere in between. This chapter touches on statistical sampling plans, statistical process monitoring, and the design and analysis of experiments.

Keywords Design of experiments · Scientific method · Statistical process monitoring · Analytic studies · Prior distributions

1 Introduction: Scientific Method—Box and Deming

For centuries now, the scientific method has been the fundamental approach for developing solutions to scientific and engineering problems. The proper use of the scientific method has been the major reason for much of modern progress in science and engineering.

Box (1999) provides an excellent overview of the role of the scientific method, which is an iterative inductive/deductive process that involves constant interplay between the concrete and abstract universes. The actual problem and its context form the concrete universe. Historically, first principle mathematical models form the abstract universe used to explain the behavior of the concrete. More recently, people use complex mathematical algorithms. These mathematical models provide useful approximations to the true behavior within the concrete universe. Scientists

G. G. Vining (✉)
Department of Statistics, Virginia Tech, Blacksburg, VA, USA
e-mail: vining@vt.edu

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_19

and engineers develop solutions based on the insights gained from these approximations. Ultimately, however, people must interact with the concrete universe to confirm the adequacy of these proposed solutions. This interaction with the concrete universe requires the collection and the interpretation of data.

A succinct summary of the scientific method:

1. Define the problem (inductive)
2. Propose an educated theory, idea or model (inductive)
3. Collect data to test the theory (deductive)
4. Analyze the results (deductive)
5. Interpret the data and draw conclusions (deductive)

This process continues until a reasonable solution emerges. Ultimately, the scientific method is a sequential learning strategy, which is the basic point to Box! The proper application of the scientific method requires:

- Model building
- Data collection
- Data analysis
- Data interpretation

These methods provide the opportunity to test the adequacy of the abstract formulation of the problem for modeling the actual concrete problem. It is for this reason that Marquardt (1987) called statistics the "handmaiden of the scientific method." Vining (2011) and Freeman et al. (2013) discuss the importance of the scientific method for the proper design and analysis of experiments in more detail.

Clearly, data are essential for the scientific method. A fundamental principle is that the data must stand purely upon themselves. Researcher bias, both in terms of the data themselves and in terms of the analysis, must be treated with great caution. Obviously, there is a major difference between data cleaning, which is fundamental in any real scientific/engineering study, and eliminating "inconvenient" data, inconvenient in the sense that they are not consistent with the researcher's hypothesis/model. However, in both cases the researcher may claim simply that he/she simply removed "outliers." Bias in the analysis is much more subtle. Frankly, bias in either area must raise serious concerns about any conclusions that result from the analysis, especially if data cleaning and data analytic procedures are not clearly stated in the final report.

2 Box and Deming

Box began his career as a chemist. He learned experimental design during World War II when he served as a sergeant dealing with toxic agents (see Box for more details). Box never really stopped being a scientist over his entire career. His instincts as a scientist strongly shaped his approach to statistics.

For Box, statistics is essential for scientists and engineers in their discovery processes. The focus on discovery is fundamental. Discovery is not mathematically coherent; rather, it is a journey involving a series of phases or steps. Each new phase builds upon what is discovered (learned) from the previous. Each phase has a different purpose, and each specific purpose guides how the scientist should approach that specific problem. For Box, discovery is a sequential learning adventure based on the scientific method. Discovery always is an investigation!

The early phases are pure exploration, trying to see how first principles and previous experimentation apply to the investigation at hand. In the early phases, no one truly knows what factors are of real interest. People may not even know what responses to measure. There is no single model to be estimated/tested. Rather, the goal is to begin to develop what appear to be the truly important factors and how they relate to the critical responses that reflect the problem at hand. Over time, the important factors and an approximate model emerge. In the final phases, the researchers seek to confirm the model and to provide very good estimates of the important parameters.

The discovery process is extremely dynamic, changing, often dramatically, from phase to phase. Experimentation must support model robustness as the researchers seek to develop reasonable models to explain the concrete behavior. The models proposed are never correct, but they are useful. Especially in the early phases, the models proposed can be simultaneously under and over-specified. These models may not reflect all of the important factors. The proposed ranges for the factors being studied may not be close to their "optimal" values.

During the 1970s, Box and Kiefer, the father of optimal experimental design theory, had a fierce debate. Especially interesting is an issue of Biometrika in 1975. Kiefer (1975) appeared just a few pages before Box and Draper (1975). Kiefer discussed robustness to the choice of variance based criterion for selecting an optimal design. Box and Draper discussed robustness to the model and other assumptions, in particular outliers.

It is clear from this discussion that Kiefer's focus was on confirmation. He assumes that the model is correct and the real purpose of the experiment is the precise estimation (think final estimation) of the model parameters. Model robustness is of no concern to him. Confirmation is an important phase in the discovery process; however, it is only one phase, the final phase. The confirmation phase is much closer to a static situation than the entire discovery process.

The issue of discovery versus confirmation, dynamic versus static is similar to Deming (1986) concepts of analytic versus enumerative studies. Deming's real contributions to statistics are in sampling. A census is a classic example of an enumerative study. The goal is to describe a static population at a specific point in time. Analytic studies, on the other hand, deal with dynamic processes. For Deming, control charts are a classic example of an analytic study on a dynamic process, and hypothesis tests are classic examples of enumerative studies on a static population. A process being monitored by a control chart is not static but subject to change at any time. As a result, to view a control chart as a series of hypothesis tests completely violates Deming's world view. It lacks profound knowledge. Of course, Deming's

world view does not include the differences between Phase I (very dynamic) and Phase II (much more static, especially under Deming's basic assumptions about assignable causes) control charts. Nonetheless, the point is valuable: Static processes lend themselves to different analytical techniques than dynamic, just as discovery involves a great deal more than confirmation.

3 Basic Issues with Bayesian Methods

The key to Bayesian analysis is the posterior distribution of the data, which has the form (following Casella and Berger (2002), p. 324)

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\mathbf{y})}$$

where

- $\mathbf{y}$ is the vector of the observed data
- $\pi(\boldsymbol{\theta})$ is the prior distribution on the parameter vector $\boldsymbol{\theta}$
- $f(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function
- $m(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta$ is the marginal distribution of $\mathbf{Y}$.

Bayesian analysis requires strong distributional assumptions, unlike ordinary least squares that only makes second moment assumptions. First, the Bayesian analysis assumes a strong distributional form for the fundamental likelihood function. It then adds another strong distributional assumption for the prior distribution. Often Bayesian analysts soften the strength of their assumptions by assuming diffuse or non-informative priors. We discuss this issue in more detail later in this section.

Bayesian inference uses the posterior distribution to calculate the risk function (see Casella and Berger, p. 349). Let $\delta(\mathbf{y})$ be an estimator of θ. The risk function, $R(\theta, \delta)$, is given by

$$R(\theta, \delta) = \mathrm{E}_\theta \left[\ell(\theta, \delta(\mathbf{Y})\right] \ ,$$

where $\ell(\theta, \delta(\mathbf{Y}))$ is an appropriate loss function. Analysts often use squared error loss of the form $\ell(\theta, a) = (a - \theta)^2$. The "best" parameter estimate minimizes the risk function. A Bayesian optimal experimental design optimizes the appropriate risk function both over the parameter space and over the experimental region.

Our discussion requires us to focus on the dependence of the analysis on the prior distribution. The key point is that formal Bayesian approaches impart bias. This bias is extremely useful if it reflects the truth. The problem is that the prior distribution never completely represents the truth. However, if the prior information closely reflects reality, then the Bayesian analysis can speed the investigation precisely because we allow the prior distribution to bias the data in the "correct" direction. On the other hand, the prior distribution also can impede the speed of the investigation

because the prior distribution can dominate the data, especially for small data sets. However, even for moderate to large sample sizes, the prior can be so strong that it continues to dominate. Of especial danger are prior distributions that are much stronger than the analyst understands.

Consider the situation where one can model the data by a normal distribution with a known variance σ^2 and unknown mean μ and with a normal prior distribution with mean μ_0 and variance τ^2. Please note that τ^2 controls the strength of the prior distribution. Let μ_p be the posterior mean, and let σ_p^2 be the posterior variance. Under the assumption of squared error loss function, μ_p minimizes the Bayes risk function. Let $\overline{y}_n$ be the sample mean for a random sample of n observations. With quite a bit of algebra, one can show that

$$\mu_p = \frac{\sigma^2\mu_0 + n\tau^2\overline{y}_n}{\sigma^2 + n\tau^2}. \tag{1}$$

We note that as $\tau^2 \to 0$, $\mu_p \to \theta$ without regard to the data! The message is clear: The stronger the prior, the less important are the data. We also can show that

$$\sigma_p^2 = \frac{\tau^2\sigma^2}{\sigma^2 + n\tau^2}.$$

We now note that as $\tau^2 \to \infty$, $\mu_p \to \overline{y}_n$ and $\sigma_p^2 \to \frac{\sigma^2}{n}$. As a result, the posterior mean is simply the standard frequentist estimate, which does not even require the assumption of normality by the Central Limit theorem. Casella and Berger, p. 326 make similar points. The key observation for this chapter is that the analyst gains nothing by the use of the diffuse prior at the expense of much stronger assumptions.

The assumption of the point prior ($\tau = 0$) is obviously extreme; however, it does make a basic point. At what point do the data overwhelm the prior information? In the next section, we illustrate that much less extreme and actually plausible priors have a huge influence on the posterior mean. However, for the moment, we need to consider the naive practitioner, who does not understand how the prior biases the posterior mean.

For example, I have worked with a Bayesian statistician at NASA who has fallen in love with WINBUGS, a popular software for performing Bayesian analyses based on Markov chain Monte Carlo (website: winbugs-development.mrc-bsu.cam.ac.uk). I remember how proud he was of an analysis he did on some of our preliminary data. He had assumed a prior distribution and then analyzed some updated data. He was extremely proud that WINBUGS could even plot the posterior distribution, which he claimed maximum likelihood could not. Ironically, his plot was bimodal. The actual data were not consistent with his prior distribution. The sample size was too small to allow the data to overwhelm his prior. Of course, the final irony was that the maximum likelihood estimate asymptotically followed a normal distribution. As a result, the maximum likelihood analysis did provide a plot for the resulting estimate, and that plot was much more intuitive (single peaked).

Good Bayesian analysts understand the basic issues. They recognize that the quality of the resulting inference requires that the prior distribution is essentially correct. The common use of diffuse or non-informative priors occurs when the analyst has very little information about the possible values for the parameters. Diffuse priors allow all of the possible values for the parameter to be considered, unlike some strong priors which give very little or no possibility for portions of the parameter space. However, in general diffuse priors provide almost no benefit over standard frequentist analysis at the expense of stronger assumptions. This is especially true within the experimental design community where estimation is based on least squares and inference assumes Central Limit theory. The real benefit as well as the real risk comes from the use of stronger prior distributions, generally based on recent historical data. We then run the risk of falling into the trap of my NASA colleague.

In the enumerative study or confirmation experiment, the researchers may have valid, strong prior information about the system because it is much more static and stable. In such a case, Bayesian analysis based on a strong prior seems very reasonable. However, in the analytic study or the discovery process, the system is extremely dynamic. The researchers have some idea about the nature of the system, but the quality of that information is questionable because it may not be relevant, in which case it has strong potential to bias the analysis. In a dynamic situation, the available prior information often does not provide a good basis for the use of a strong prior.

4 Applications of Bayesian Approaches to Process Monitoring

Consider a simple illustrative example involving a batch production process. The monitoring scheme focuses on the sample mean of a continuous characteristic. The organization uses a modification of a control chart; thus, it rejects the batch if the sample mean exceeds a threshold value. A common Bayesian approach uses a normal distribution as the prior distribution. The resulting posterior distribution is also normal, and Eq. (1) gives the resulting posterior mean.

The first question is how to choose the prior distribution. Many practitioners would borrow the concepts of Phase I and Phase II from statistical process monitoring. Phase I in this situation uses a base period of m batches *to define* the prior distribution. A common practitioner view of the prior distribution is the historic behavior of the process. Treating the Phase I estimates as the basis for defining the prior distribution follows naturally, especially in the absence of any other information. This situation uses the Type I period slightly differently from standard control charts, where the focus is on purely "in-control" rational subgroups. In this application, the Type I period uses all batches, both accepted and rejected, in order to define the true variability in the batch means.

In the previous section, we established that using a non-informative, diffuse prior is little different than using the frequentist estimate of μ. Suppose that the organization has m batches as the Phase I period. Let y_{ij} be the j^{th} observation sampled from the i^{th} batch for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. Let $\overline{y}_{..}$ be the sample mean for the quality characteristic over the base period. In the tradition of treating the Phase I "historic" estimate of the parameter as the true value, then $\mu_0 = \overline{y}_{..}$. Let σ_b^2 be the historic batch-to-batch variability, and let σ^2 be the historic within batch variability. It can be shown that the variance of $\overline{y}_{..}$ is

$$\frac{1}{m}\left[\sigma_b^2 + \frac{\sigma^2}{n}\right],$$

which a practitioner has valid reasons to use as an appropriate variance for the prior distribution for μ in Phase II. It is the variance for the empirical Bayes estimate of μ for an extremely diffuse prior in Phase I.

Assume that $\sigma_b^2 = k\sigma^2$, where $k\sigma^2$ represents the historic batch-to-batch variability. Typically, a reasonable guess is that $1 \leq k \leq 5$. For simplicity, assume that $m > 50$ and that n is of moderate size. A reasonable approximation for τ^2, the variance of the prior distribution, is

$$\tau^2 = \frac{1}{m}\left[\sigma_b^2 + \frac{\sigma^2}{n}\right] = \frac{k + 1/n}{m} \approx \frac{k\sigma^2}{m}. \tag{2}$$

Let $\overline{y}_{i.}$ be the observed sample mean for the i^{th} batch. The posterior batch mean then becomes

$$\mu_p = \frac{\mu_0 + \frac{k}{m} n \overline{y_{i.}}}{\frac{k}{m} n + 1}.$$

Phase II control charts recognize that we can treat the sequential monitoring of a process as a sequence of hypothesis tests. For simplicity assume that the organization rejects a batch only if it believes that the batch's true mean is too large. Casella and Berger, pp. 379 and 380, outline a Bayesian test for such a situation. Let $\theta_0 > \mu_0$ be a suitably chosen constant. Their test is defined in terms of the observed sample mean for the i^{th} batch, $\overline{y}_{i.}$, and rejects the batch if

$$\overline{y}_{i.} > \theta_0 + \frac{\sigma^2(\theta_0 - \overline{y}_{..})}{n\tau^2}.$$

which is equivalent to rejecting the batch if its posterior mean for the i^{th} batch is greater than θ_0. By applying (2), the decision becomes to reject the batch if

$$\overline{y}_{i.} > \theta_0 + \frac{m(\theta_0 - \overline{y}_{..})}{kn}.$$

An important question is at what point does the sample mean begin to overwhelm the prior distribution. Let ν be the relative weight given to the sample mean. The sample size required to give at least ν weight to the sample mean is

$$n \geq \frac{m\nu}{k(1-\nu)}.$$

Consider the case where $m = 50$, $k = 5$, and we wish to have a sample size that gives exactly the same weight to the sample mean and the mean of the prior ($\nu = 0.5$). The resulting sample size is $n = 10$. Giving the sample mean only equal weight is not really dominating. Consider the same scenario with $\nu = 0.9$. The resulting sample size is $n = 90$. Even larger sample sizes are required as m increases because we have more precision for the prior distribution.

Given what we know about Phase I control charts, requiring a minimum number of batches of 50 for the base period is quite reasonable. Yet, it is clear that the resulting prior distribution is quite strong and almost surely much stronger than the naive analyst assumes. Once again, if the data are consistent with the prior, everything is OK. However, the point of the monitoring scheme is to protect the organization from shipping a bad batch.

The reality is that all production processes are truly dynamic except for relatively short periods of time. The primary purpose for k in our argument is to account for the typical batch-to-batch variability. However, the typical batch-to-batch variability probably does not reflect the more serious quality issues that this process faces over longer periods of time. Treating the process as static (performing an enumerative study) creates serious problems with overconfidence about the monitoring scheme's ability to detect serious problems.

The reviewer rightfully points out that "... sampling and process monitoring cannot start with the assumption of a constant general mean because they are intended to find out a change in the general mean." I cannot agree more with this statement. In this specific example the naive practitioner defines the general mean as the mean for the entire production process, not the mean for the acceptable production from this process. Recall, the practitioner used both the accepted and the rejected batches to define the prior distribution. His/her approach is quite legitimate for that approach. It essentially is an empirical Bayes estimate of this overall mean, assuming that the base period accurately accounts for the total batch-to-batch variability.

One may argue, quite legitimately, that the problem here is a very naive use of Bayesian analysis. Once again, I cannot agree more. However, there is little guidance for practitioners of how to construct reasonable prior distributions other than non-informative, diffuse ones. Converting available historical information into valuable priors is essential for the success of Bayesian approaches in the practitioner world! It is too easy for the naive user to create an extremely strong prior distribution without realizing it. As a result, this practitioner commits an error of the third kind: an elegant solution to the wrong problem. Essentially, the practitioner is applying an enumerative/confirmation approach to a analytic/discovery problem. Unfortunately,

practitioners make such mistakes on a too regular basis, particularly in applying Bayesian inference.

As a final note, the beginning of Section 2 of Box (1980) uses the Bayesian predictive distribution to assess the reasonableness of the posterior mean. Essentially, Box is assessing the reasonableness of the prior distribution given the data. Box notes that the prior distribution is part of the model for the data, and, of course, for Box all models are wrong. Model robustness to the choice of prior deserves much more discussion in the practical literature.

Originally, I had hoped to discover a much richer literature on Bayesian statistical monitoring procedures. I was disappointed to find very little. The purpose of this section was to highlight a possible reason. I should have paid more attention to Woodall and Montgomery (2014), who note "These (Bayesian) methods do not seem widely used." I had expected to see issues with inertia, and I was curious to see what approaches authors used to combat that problem.

5 Experimental Design and Analysis

Freeman et al. (2013) outline the basic stages in planning experiments as:

1. Define the Problem and the Specific Objective for the Experiment
2. Select the Responses
3. Determine Appropriate Factors
4. Define the Region of Operability (the set of all possible values for the factors)
5. Define the Specific Experimental Region (the set of values for the specific experiment)
6. Identify Nuisance Factors
7. Define Tentative Model
8. Understand What Are Alternative Models
9. Choose the Design
10. Check for an Error of the Third Kind (an elegant solution to the wrong problem!)
11. Train People to Conduct the Experiment
12. Collect the Data
13. Analyze the Data in Light of the Actual Experiment Conducted

It is vital to note that all experiments are sequential! They build, either formally or informally, upon previous experimentation. Subject matter expertise and insight, both based on discipline specific first principles and on experience, are essential for success, especially for determining the factors, the experimental regions, and the initial tentative model.

A valid question is why do so many people view experiments as "one-shot"? A very basic answer is that most textbooks illustrate experiments in that manner. The focus is on the analysis more than the actual planning phases. The planning reflects the true sequential nature of experimentation. For example, a classic textbook

example is an agricultural field trial. In most parts of the world, a researcher has only one growing season per year to conduct experimentation. As a result, she/he plans an experiment to obtain as much information as possible. The resulting experiment appears to be stand alone. The reality, however, is that each year's experiment builds upon what was learned from the previous years experience. A "one-shot" agricultural field trial may reflect the contribution of a masters' level student's thesis. The Ph.D. dissertation, however, reflects the full sequential nature of the experiment, including the full sequential learning.

Box clearly shows the sequential nature of industrial experimentation. He notes two primary reasons: immediacy and sequentiality. Even in the 1950s, a researcher could conduct an experiment, especially in a pilot plant 1 week, analyze the results the next week, and then conduct a follow up experiment the next. The ability to get results almost immediately (unlike the agricultural field trial) allows the researcher to conduct a true experimental campaign consisting of series of experiments within a sequential learning strategy.

Classical approaches to planning experiments clearly embrace the need for prior information; however, it also understands the limitations on that prior information, particularly its relevance or potential lack thereof, especially in the early phases of an experimental campaign. Is the purpose of the specific phase discovery or confirmation? Do we seek to build a useful model or do we seek to provide very good estimates of the parameters for a "final" model? There is a fundamental difference between subject matter expertise and insight and the formal prior belief summarized by a prior distribution.

Until now, our focus on Bayesian approaches is on the analysis of data already collected. However, the experimental design community is now embracing the use of Bayesian approaches for constructing the experimental design, especially within the optimal design community. The issue of bias in the analysis carries over very strongly into bias in the location of the design runs.

The choice of experimental design always depends upon the approach to the analysis. Traditional optimal designs for regression models use criteria such as the maximum determinant of the information matrix or the integrated prediction variance over the region of interest. However, traditional optimal design criteria for regression models do not depend upon the parameters being estimated. A legitimate entre point for Bayesian approaches in choosing the experimental design occurs when the information matrix depends on the parameters to be estimated. The information matrix determines the points of support for the model to be estimated. These points of support are the primary candidate runs for the "optimal" design. Examples where the information matrix depends on the parameters to be estimated include:

- Non-Linear Regression Models
- Generalized Linear Models
- Reliability Experiments, especially with the Weibull Distribution
- Robust Parameter Design—Mean/Variance.

The crucial point becomes what is the fundamental purpose of the experiment: discovery or confirmation? If the purpose is discovery, then the Bayesian approach requires non-informational priors that offer little benefit at the expense of much stronger assumptions. On the other hand, if the purpose is confirmation, then the researcher may have sufficient prior information that can be converted into meaningful and insightful prior distributions.

Bates and Watts (1988) briefly discuss Bayesian optimal designs for nonlinear regression models. They make the basic point that starting with a good classical linear regression model based design is a very good option because it corresponds to the Bayesian choice for an extremely noninformative prior. Their recommendation is perfectly consistent with experiments primarily for discovery.

One of the most promising applications for the use of Bayesian optimal experimental design is nonlinear regression in areas like pharmacokinetics. Typically, the model involves only one factor, time, and the researchers justify the basic nonlinear model form using first principles based on the solutions to differential equations. Particularly in a pharmacokinetics study, there are previous studies which should be quite relevant to the new experiment. As a result, the researchers should have the background information to create a meaningful, relatively strong prior distribution. Finally, the purpose of most pharmacokinetics studies is confirmation. The key difference here from the Bates and Watts recommendation is confirmation after a great deal of formal information rather than discovery.

It is important to note that the pharmacokinetics context is quite unique. There are many situations involving nonlinear regression models where the first principles strongly suggest a specific nonlinear model form; however, there is too little prior information available/relevant to create a meaningful prior distribution. The use of apparently very diffuse prior distributions can lead to some interesting results, especially very inconvenient factor settings. I personally question the use of such priors.

Another popular area for Bayesian optimal designs is generalized linear models, especially logistic regression. Maximum likelihood is the most widely accepted method for estimating a logistic model. Maximum likelihood estimation requires data in the factor space representing the transition from all success to all failures, i.e. the probability of success is truly between 0 and 1. This requirement can present serious challenges, especially if the purpose of the study is discovery.

Once again, the issue is the quality of the information available prior to running the experiment. Using a strong prior with a logistic regression during discovery can lead to the worst case scenario of all success or all failures. Such a consequence should be very rare. On the other hand, not having any runs in the transition region of the factor space is a very serious issue and occurs more frequently than desirable.

Other issues with generalized linear models, especially for discovery, is the lack of a first principles justification for the model form. Nonlinear regression often has a solid first principles basis. Ultimately, most generalized linear models are nothing more than low-order Taylor series approximations in the linear predictor. Model robustness issues arise as a result.

6 Final Comments

The scientific method is an important sequential problem solving approach that has proved very useful over the centuries. The successful application of the scientific method requires that the data stand on their own. Issues of bias, even the potential of bias, have serious consequences for the integrity of the investigation.

The early phases of the scientific method tend to focus on discovery. The scientific method depends heavily on subject matter expertise and insight. However, in the early phases of the investigation, it is highly questionable that this expertise and insight translate well into formal mathematical prior distributions. Too much is unknown in the early phases.

Formal Bayesian approaches can have great success as the prior information becomes better defined and thus more amenable to translation as formal mathematical priors. As the investigation closes onto a solution, the more likely the prior distributions provide an accurate basis for inclusion in the analysis. Some areas where there is strong potential are

- Experiments Involving Systems of Systems
  - Subsystems Are Well Understood
  - The System of Subsystems Is Not
- Situations with Well Understood Fundamental Mechanisms with Good Insights from Other Experiments
- Final Stage Confirmation of the Model Produced from the Discovery Process.

Ultimately, we need to use the right tool for the right job. In some cases, the appropriate tools are Bayesian, if they are used with care. When determining the proper tools, it is vital to understand that discovery is a different world than confirmation. In most experimental situations, success depends on the proper understanding of the experimental context and the experimental goals.

References

Bates, D. M., & Watss, D. G. (1988). *Nonlinear Regression Analysis and Its Interpretation*. New York: John Wiley and Sons.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A, 143*, 383–430.

Box, G. E. P (1999). Statistics as a catalyst to learning by scientific method part II – discussion. *Journal of Quality Technology, 31*(1), 16–29.

Box, G. E. P., & Draper, N. R. (1975). Robust designs. *Biometrika, 62*, 347–352.

Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Deming, W. E. (1986). *Out of the Crisis*. Cambridge, MA: MIT Center for Advanced Engineering Study.

Freeman, L. J., Ryan, A. G., Kensler, J. J. K., Dickinson, R. M., & Vining, G. G. (2013). A tutorial on the planning of experiments. *Quality Engineering, 25*, 315–332.

Kiefer, J. (1975). Optimum design: Variation in structure and performance under change in criterion. *Biometrika, 62*, 277–288.

Marquardt, D. W. (1987). The importance of statisticians. *Journal of the American Statistical Association, 82*, 1–7.

Vining, G. (2011). Technical advice: Design of experiments, response surface methodology, and sequential experimentation. *Quality Engineering, 23*, 217–220.

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

A Note on the Quality of Biomedical Statistics

Elart von Collani

Abstract During the last decades numerous articles were published dealing with the bad quality of biomedical statistics. However, most of the relevant chapters confine themselves to describe misunderstandings, misinterpretations and misuses of statistical methods. In contrast, in this chapter it is argued that the bad quality of biomedical statistics is also due to the statistical methodology and statistical methods themselves. This claim is illustrated by several examples. Special emphasize is laid on significance testing the most often applied statistical method in biostatistics. This chapter aims at raising the awareness of the statistical community for what is going on in medicine and hoping that this will lead to some fundamentals improvements in statistics.

Keywords Laboratory medicine · Evidence-based medicine · Significance test · Probability · Jakob Bernoulli

1 Introduction

During the last 5 years I came in very close contact with medicine and especially the use of statistical methods in medicine. I remember one of the first disturbing moments occurred when my oncologist told me that I should not compare my blood values determined by different laboratories because even the examination results of the same blood sample may differ greatly. This could lead to different therapeutic measures and thus endanger the success of a treatment.

E. von Collani (✉)
University Würzburg, Würzburg, Germany
e-mail: elart.collani@uni-wuerzburg.de

S. Knoth, W. Schmid (eds.), *Frontiers in Statistical Quality Control 12*,
Frontiers in Statistical Quality Control,
https://doi.org/10.1007/978-3-319-75295-2_20

When discussing with physicians my concerns with respect to statistical methods in medicine, I generally meet complete agreement. However, many of them told me the following:

- Physicians not only feel left alone by statistics, but that statisticians propose the methods and interpretations which they apply and which are afterwards criticized by other statisticians.
- The education of physicians does not qualify them to be able to judge statistical methods and once working as physicians they have no time and opportunity to catch up on statistics.
- Many physicians feel that the critique of their use of statistical methods is unjustified because statistics is not their field of expertise.

During the discussions some of the physicians indicated that medicine had already reacted on the existing weaknesses by developing the so-called evidence-based medicine (EbM). They told me that EbM would provide serious evidence with respect to diagnosis and treatment. As a matter of fact EbM was new to me and their words intrigued me. I will come back later to it. To begin with lets have a closer look to laboratory medicine.

2 Laboratory Medicine

From Wikipedia we learn: "A medical laboratory or clinical laboratory is a laboratory where tests are usually done on clinical specimens in order to obtain information about the health of a patient as pertaining to the diagnosis, treatment, and prevention of disease." And "Credibility of medical laboratories is paramount to the health and safety of the patients relying on the testing services provided by these labs. The international standard in use today for the accreditation of medical laboratories is *ISO 15189 - Medical laboratories - Requirements for quality and competence*" (ISO 15189 2012). Thus, if something goes wrong with laboratory medicine then it is due to an ISO standard. Actually, ISO 15189 appears to be one of the fastest growing international quality standards in the world. By 2013 the standard was adopted by medical laboratories in over 60 countries. The quality of medical laboratories is controlled via round robin tests aiming at improving comparability of the results of different laboratories. The overall goal is that for any parameter which is determined by different laboratories the results must be comparable.

In view of this goal, my personal experiences with several commercial and hospital laboratories which were all accredited according to ISO 15189 show that it has not been reached so far. I noticed the following:

- Different laboratories use different units. This may cause errors of inexperienced personal and represents a potential danger for patients. In fact, it is a miracle to me that the units in laboratory reports may change from laboratory to laboratory.

The units should be fixed and any deviation should necessarily lead to the loss of accreditation.

- The laboratory results are given by single numbers and these numbers can differ greatly from laboratory to laboratory. In fact, according to my experience a differences of 20% is not uncommon and even 30% from the same blood sample may occur. This is, of course, due to measurement uncertainty. The relevant part of the standard reads as follows:

  ISO 15189: 5.6.2. The laboratory shall determine the uncertainty of results, where relevant and possible. Uncertainty components, which are of importance, shall be taken into account. Sources that contribute to uncertainty may include sampling, sample preparation, sample portion election, calibrators, reference materials, input quantities, equipment used, environmental conditions, condition of the sample and changes of operator.

  However, when checking the laboratory reports, I have never seen anything which could be interpreted as uncertainty of the given measurement result.
- When trying to figure out the reasons for not revealing the underlying uncertainty of measurement, I learnt that the uncertainties were hidden in the reference ranges. Actually, each laboratory has specific reference ranges. Accordingly, for many values the following citation is valid: "The reference values and the values obtained may differ significantly from laboratory to laboratory".

  The following example shall illustrate the above:

  The CRP-value (C-reactive protein) is used as a marker of inflammation and belongs to the most often determined parameters in laboratory medicine. From the reports of two laboratories we find the following statements:

Laboratory A	Range of reference 0.00–0.50	Unit mg/l
Laboratory B	Range of reference 0.00–8.00	Unit mg/dl

  Since the values of many of the examined parameters may range by several orders of magnitude, mistakes may easily be made, whenever an unexpected unit is used or if the range of reference deviates from the familiar one.

The consequence of not stating the uncertainty of the values in the laboratory reports is that the results may be misinterpreted and thus endanger health or even life of patients. Therefore, if the laboratory is not known to the physician in charge, the measurement are not trusted and, therefore, repeated.

Thus, to obtain comparable laboratory data, it is necessary that measurement uncertainty is clearly stated in the reports. Hiding measurement uncertainty by laboratory specific reference ranges does not help much and may, in some circumstances, even promote misunderstandings. Therefore, simple and straightforward methods are needed to determine the uncertainty of measurement. The reference ranges, on the other hand, should be determined by the relevant health organizations and be identical for all the accredited laboratories.

Unfortunately, statistics neglects measurement uncertainty and has left the field to metrology. More than 20 years ago the "Guide to the Expression of Uncertainty in Measurement (GUM)" (Joint Committee for Guides in Metrology 2008) was published and is still in use. However, from the very beginning the proposed methods were criticized, because they are questionable and at the same time too complicated. Since measurements are the most important means for quality control, I appeal to the statistical community to turn to this eminently important field and make available simple and easy to understand methods for determining measurement uncertainty. Actually, such methods are already available—see Collani and Dräger (2001) at least for some special cases.

Next let me turn to "evidence-based medicine" which is often looked upon as a means to avoid wrong recommendations in making diagnosis and determining on therapies in all areas of medicine.

3 Evidence-Based Medicine (EbM)

The evidence-based medicine (EbM) has developed since the 1990s (see Sack et al. 1996). EbM is defined as the medical care and treatment of patients on the basis of the best available sources of knowledge and information. Therefore, it aims at defining requirements that only those medical procedures are recommended and should be incorporated into guidelines and principles, whose positive effects have been proven. For EbM, two types of studies (called "gold standards") are primarily considered as giving evidence, namely "randomized controlled clinical trials" and "meta-studies".

- Randomized controlled clinical trials:
  A clinical study is called "controlled" if there is both an experimental group and a control group. "Randomized" means that the assignment of subjects to experimental or control group is random, that is, each subject is assigned with equal probability to the experimental group or to the control group. In addition, randomized controlled trials are usually double-blind that is, both the subject itself and the experimenter do not know whether the subject is part of the experimental or the control group.
- Meta-Studies:
  The second basis of EbM are meta-studies. Often the same treatment is investigated by several clinical trials, although contradictory results are published. A meta-study attempts to combine the results of several randomized controlled clinical trials. The results of the various published studies are compared with each other and then evaluated together. It is thereby hoped to get an overall larger sample size and thus to better sound results.

For each clinical trial, the study design and the evaluation method must be distinguished. The study design determines which indicators are to be observed when, how often, and for which of the study subjects. This depends on the specific

medical procedures to be applied and especially on the aim of the study, the type of treatment to be tested and of the study indication. Depending on the study objective there are different study designs, such as the single case study, the cohort study, the case-control study, etc. Once the observations are available, they must be analyzed statistically. This is done using the evaluation method which covers all the requirements, models and statistical methods that are to be used. In contrast to the study design, the evaluation method is less determined by the medical purpose of the study. This is primarily due to statistics that offers many models and methods for evaluation of one and the same situation. The user therefore faces the problem to select an adequate method among the various competing statistical tools. The steadily growing number of statistical analysis methods that are available in a given case lead on to errors and misinterpretation. This is one of the many reasons for the large number of articles in medical literature that report on the big rate of medical chapters with erroneous statistical analysis. Already 35 years ago, Stanton Glantz (Glantz 1980) wrote in an article entitled "Biostatistics: how to detect, correct and prevent errors in the medical literature":

> Critical reviewers of the biomedical literature have consistently found that about half the articles that used statistical methods did so incorrectly.

This state has not changed until today as the following quote from a work by Lang and Altman (2013) shows which was published in 2013:

> The first major study of the quality of statistical reporting in the biomedical literature was published in 1966. Since then, dozens of similar studies have been published, every one of which has found that large proportions of articles contain errors in the application, analysis, interpretation, or reporting of statistics or in the design or conduct of research. Further, large proportions of these errors are serious enough to call the authors' conclusions into question. The problem is made worse by the fact that most of these studies are of the world's leading peer-reviewed general medical and specialty journals.

Before the EbM approach shall be evaluated with respect to quality, we must first answer the question which claims are to be placed on a trial so that the study results may be judged as evidence or proof. In this context it is necessary to distinguish between "assertion" and "assumption". The goal of a proof is to show that the assertion follows necessarily from the assumption. If this goal is met, the assertion can be considered as true, if the made assumptions are recognized as being correct. The central criterion is the consistency of the model assumptions with reality. In order to check the consistency, the trial must meet certain requirements which shall make manipulations difficult and the results verifiable by the statistical community.

The statistical community is responsible for the validation of new findings whenever the results are obtained by applying statistical methods. Note that the requirements are not intended to regulate clinical trials, because that would be an unjustified restriction of academic freedom and would only hinder scientific progress.

- Requirements to prevent manipulations:

  1. The aim of the study must be stated clearly and unambiguously. The assertion to be derived must be consistent with the target in line. If one of these requirements is not satisfied, it remains unclear if the objective has been really achieved. If the aim of the study is not clear and unambiguous, then the trial is like a shooter who shoots on a large barn and then paints the target around the bullet hole.
  2. The study design must define clearly, when the data recording is finished and the data analysis may start. If the end of data collection is not fixed, the procedure is similar to a horse race in which it remains open when the race is over and the race ends when your own horse is ahead.
  3. All assumptions and statistical methods by means of which the assertion shall be deduced, must be stated right at the start. If this requirement is not met, assumptions and methods could be selected later on the basis of the observed data. Or in other words, one could try all possible statistical methods, until a procedure is found that leads to a "significance". This result would then be published.

- Requirement to make the result verifiable:

  4. Immediately after completion of the data collection, all raw data that have been collected during the study (including those later eliminated as outliers) must be made available to the public. If this requirement is not met, then the study results cannot be verified and should therefore not be taken as evidence. Actually, clinical trials are often conducted by companies which refuse to publish the raw data, because they represent "business secrets". If the data are business secrets then the results are also business secrets and must not be looked upon as evidence but rather as marketing tools.

These four requirements are prerequisites for a clinical trial so that the results may be considered as evidence. Whether actually evidence is given, must be examined by a review of the evaluation method and by reproducing the results. Of course, one would have to develop criteria for this review, because statistics contain many questionable methods and concepts. These include the significance test which is almost always used in clinical trials and which is briefly examined later.

In view of these requirements it must be noted that the two "gold standards" of EbM do not fulfill them. Instead, EbM stipulates a study design which makes only sense if a comparison between at least two different methods of treatment should be made. If this is not the case, the implementation of a controlled trial makes little sense. But even in the case where a comparison by means of a controlled clinical trial should be done, this can lead to evidence at best if the above requirements would be met which however is not demanded by the EbM approach. The establishment of a control group implies two additional problems. First of all the ethical issue has to be considered which emerges when ill persons are given a non-effective treatment. Moreover, the overall sample size is cut in half by the control group. This makes a study unnecessarily expensive.

Instead of demanding the above specified requirements the gold standard includes randomization. Randomization means that the available subjects are allocated randomly to the given groups. The aim of the allocation is to form as homogeneous groups as possible in view of the comparison's objective. Homogeneity refers to all the characteristics of the subjects which could play a role in the comparison. In such a situation the allocation of subjects should not be left to chance, but the subjects should be specifically selected so that the groups are as equivalent as possible with respect to the planned intervention. If the groups, as is the case for randomized trials, are randomly occupied, then it cannot be ruled out that the study is conducted with groups that are not at all homogeneous. Maybe randomization in medical studies is so common, because it makes a targeted manipulation of the grouping at least difficult.

Scientific claims must be verifiable by the corresponding scientific community otherwise they should not be accepted as evidence. This applies in particular in medicine, where it comes to the health and lives of people. By the assessment of randomized controlled trials as "gold standard" they take a position which they do not deserve. The mere fact of a randomized controlled trial makes many physicians believe in the evidence of the results. This is particularly serious because randomized controlled trials are generally used in the drug development process and the results are the basis for the regulatory decisions of the authorities. The statistical community is therefore called upon to clarify the corresponding misunderstandings and to show the way to achieve real evidence.

4 Test of Significance

The significance test is the most widely used statistical method in applications. At the same time it is also one of the most questionable one. For many decades articles are published dealing with shortcomings and false interpretations of the results of the significance tests in medicine. Nonetheless articles based on significance testing are still published in scientific journals. In many cases published chapters contain contradictory results that have led and lead to wrong decisions. Moreover reports of fraud and forgery in the application of significance tests are almost daily occurrence.

Verifiability i.e. reproducibility of results is a necessary condition for science. To allow verifiability of a scientific method it must yield with high probability a correct and sufficiently accurate result. If this condition is not met by a method, as for example by methods applied in astrology, then the method cannot be looked upon as part of science. In numerous publications it is shown that the significance test does not usually fulfill its promises. The article "A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research" by Levine et al. (2008) not only lists the main shortcomings of the significance tests, but also contains a bibliography of the many works that deal with this issue.

The significance test of today, hereinafter referred to as modern significance test, was developed from two sources: the significance test of Fisher, which is described in the work "Statistical Methods for Research Workers" (Fisher 1934) and the hypothesis test of Neyman-Pearson, which was published in 1933 in the chapter "On the issue of the Most Efficient tests of Statistical Hypotheses" (Neyman and Peason 1933). Unfortunately, neither Fisher nor Neyman and Pearson succeeded in displaying the meaning and purpose of their respective procedures sufficiently clear. So it is no wonder that their work has been misunderstood and it has come to the present day confusion.

4.1 Fisher's Significance Test

The word "significant" appeared already at the end of the nineteenth century in the statistical literature, but the "significance test" was only introduced by R.A. Fisher (1934) in his famous book "Statistical Methods for Research Workers", whose first edition was published in 1925. Fisher's approach had from the beginning two fundamental weaknesses: Fisher does not explain the meaning and purpose of a "test" nor did he clarify the meaning of the word "significant". The goal of a significance test is solely to obtain a significant result. A significant result is achieved if the so-called p-value is smaller than one of some predetermined levels of significance which Fisher set as 0.10, 0.05, 0.02 and 0.01. A significant result is interpreted as an objective indication that the treatment has the desired effect. Accordingly, the significance test of Fisher may have one of only two results. Either the target (significance) is reached or not. The latter case means that the significance test was a failure, and therefore a decision about the desired effect is impossible. It follows that a wrong decision can be made only, if a significance is falsely achieved. A failure means no wrong decision, as no decision is made. It simply means that the test does not allow a decision.

Fisher's significance test is characterized by the following issues:

- The test admits only one simple hypothesis which may be selected rather freely making manipulation of the test result possible.
- It is designed for small sample sizes, i.e. only when the difference between hypothesis and reality is considerable, the target will be reached.
- No significance level is set before the start of the experiment. Whether a significance test is successful or not, is determined only after the p-value has been calculated and compared with the proposed four levels of significance. This creates a certain arbitrariness, which should actually be avoided in scientific procedures.
- The goal is to provide a first (preliminary) indication that a particular course of action (therapy) has a desired effect. Only when such an indication exists, a larger experiment is performed and the decision about the possible effect is made.

4.2 Neyman-Pearson Hypotheses Test

In 1933 Jerzy Neyman and Egon Pearson published in the "Philosophical Transactions" of the Royal Society of London a chapter entitled "On the Problem of the Most Efficient Tests of Statistical Hypotheses". Unlike Fisher's book, which is intended for non-mathematicians Neyman and Pearson's chapter is a very mathematical work. In contrast to the significance test of Fisher it is not easy to find a meaningful example for the hypothesis test of Neyman-Pearson, because of the rather unrealistic assumptions about the situation to be examined.

The Neyman-Pearson hypotheses test is characterized by the following issues:

- The test refers to two hypotheses H_0 and H_1 and has two possible results, namely acceptance of H_0 or acceptance of H_1.
- The hypothesis H_0 represents that situation where an error (Type 1 error) has serious consequences, while H_1 represents that situation where an error (Type 2 error) is less severe.
- The probability of a Type 1 error is limited by the significance level which is defined prior to testing.
- At the specified significance level, the critical region (rejection region) for H_0 is determined so that two conditions are met: The probability of a Type 1 error is equal to the predetermined level of significance, while the probability of a Type 2 error is minimized.
- In contrast to Fisher's significance test, the goal of the hypotheses test of Neyman-Pearson is the final evaluation of a situation.
- Simple and composite hypothesis H_0 are admitted by Neyman and Pearson. However, the latter case is mathematically rather difficult and therefore applications are restricted generally to simple hypothesis H_0.

To make the hypothesis test of Neyman-Pearson meaningful, it would be necessary to admit significance levels for each of the two hypotheses. Only in this way it can be avoided, that the probability of a Type 2 error may be uncontrolled large.

4.3 Significance Test Versus Hypotheses Test

Obviously, both methods have different objectives and are based on different assumptions implying that they are not comparable. Nevertheless, Fisher and Neyman argued about which method is the better one. This dispute is hardly understandable, because Fisher's test aims at excluding one single given hypothesis, while the hypotheses test aims at detecting which of two hypotheses is the right one.

This strange controversy might also be a reason for the misunderstandings of the two methods which finally led to the "modern significance test".

4.4 Modern Significance Test

The modern significance test is a blend of the significance test of Fisher and the hypotheses test of Neyman-Pearson. Its development began in the 1940s in the social sciences. From there, the modern significance test has spread to all other areas of science and is now by far the most commonly used statistical method. It is characterized by no generally agreed rules for interpretations of the numerical results and the admissible decisions. This is certainly one of the reasons for the many reports of misuse and misinterpretation when applying a significance test.

The modern significance test and the significance test of Fisher have in common the name and the p-value. By analogy with the hypotheses test of Neyman-Pearson, there are two hypotheses namely the null hypothesis H_0 and the alternative hypothesis H_1. The alternative hypothesis represents that what one expects as a result of the test. The null hypothesis is then the complement to the alternative hypothesis. Similar to the significance test of Fisher, a significance level is often not set in the outset of the experiment. A significant result is obtained by calculating the p-value. If the value obtained is less than 0.01, the result is called "highly significant", if the result is between 0.01 and 0.05 it is called "significant" and if it between 0.05 and 0.10 "low-significant". The null hypothesis is simple or can be attributed to a simple one, which makes it possible to calculate a p-value. The probability of the Type 2 error is not minimized. A significant result is achieved if the null hypothesis is rejected, which is tantamount to the acceptance of the alternative hypothesis. There are also cases in which the result is specified as acceptance of the null hypothesis or the alternative hypothesis. It is interesting to note that the words "rejection" or "acceptance" do not occur in Fisher's original work, just as the term "null hypothesis". Only later, the term null hypothesis is introduced, possibly inspired by the symbol H_0 introduced by Neyman and Pearson. The modern significance test combines two different methods and borrows not only the weaknesses of the two method, but also adds new deficits.

Fisher intended his significance test for small samples in order to obtain a first, cost-effective and objective indication. The modern significance test demands large sample sizes making the weakness caused by the simple hypothesis a fortiori virulent. This is especially the case in so-called meta-studies in which the results of different studies are combined to increase the sample size and allegedly the reliability of results. The goal of modern significance tests, is similar to the significance test of Fisher, the rejection of the null hypothesis. If this is not possible, the procedure is a failure, i.e., it has not brought new insights. Nevertheless, in such cases the result is often stated as acceptance of the null hypothesis or rejection of the alternative hypothesis. The initial goal of the significance test of Fisher was to be an indication of the existence of an effect. In contrast, the modern significance test aims at a final judgment.

4.5 The Emergence of the Modern Significance Tests

How it could happen that such a questionable procedure as the modern significance test was developed and was able to win such a market-dominating position in science? The most important reason is probably the fact that the basic concept of statistics, the probability, is not explained clearly and each user may choose an own interpretation. By this, statistics goes against a fundamental principle of science and this fact is reflected in the statistical methods.

The modern significance test was developed with the presumably most important goal to get a "significance" and thus a publication. To achieve this goal, even questionable interpretations of the numerical results were considered. Unfortunately, there is no institution in statistics, which could exert a control function to stigmatize questionable methods and interpretations because the basis of statistics itself is questionable. The problems with the significance test is by no means a purely statistical problem but affects the whole science because the significance test is applied in all branches of science. For example the spectacular detection of new elementary particles in physics was made by means of significance tests. The may test lead in virtually all branches of science to wrong decisions. However, in medicine that deals with the health and lives of people it is especially misplaced.

5 Conclusions

Besides the above there are many more problematic issues in biomedical statistics like, for example, the widespread use of relative terms which generally assumes controlled clinical studies. Actually, many of these weaknesses may be traced back to the ambiguity of the fundamental term probability in statistics.

Two years ago I performed a survey among statisticians about the meaning of the concept "probability". The answers revealed that only very few statisticians are concerned with this question, although most of them judge it as being essential. A majority of surveyed statisticians seems to espouse the frequentist interpretation, while another big part of them are fans of the Bayes interpretation. Another surprisingly large part deems right both, the frequentist and the Bayes interpretation.

The concept probability aims at quantifying what is known as "randomness". Having this in mind it is easy to see that none of these opinions makes sense. The first one assumes a series of experiments, but randomness is independent of any series of experiments. The second one denies the existence of randomness and thereby moves statistics close to religion, while the third is simply out of question. The survey also revealed that the oldest attempt to quantify randomness is almost unknown to statisticians. Already more than 300 years ago, when Newton tried to introduce something he called mass, Jakob Bernoulli defined the concept probability of a future event as the degree of certitude of its occurrence (Bernoulli 1713). This definition reflects the fact that a future event may occur or may not occur depending on the event and the given circumstances. It is an objective quantity that exists

and is independent of any experiments and of any belief and it is in particular unambiguous. Unfortunately Jakob Bernoulli's efforts to quantify uncertainty and introduce it to science were not understood by his contemporary scientists and as a result science was developed based on certainty while uncertainty was simply discarded.

If statistics should become an acknowledged branch of science then Jakob Bernoulli's interpretation must be accepted by the entire statistical community. Moreover, results obtained by statistical methods must become verifiable, i.e. reproducible. This means that the results must occur with a known and sufficiently high probability. Any method which does not yield results meeting this requirement should be abandoned. Finally, models should be developed not following mathematical or philosophical principles, but should be guided by reality, i.e. for one situation should be only one model.

All these changes seem to be straightforward and attainable without big difficulties. The only problem is that they challenge tradition and necessitate entrenched habits. But if statistics should get rid of its bad image which let people say: "Never trust statistics you didn't fake yourself," and if the quality of biomedical statistics should be improved then these changes must come true. If statistics is not able to change then it will share at one time or another the fate of astrology or alchemy which are not anymore considered as part of science.

References

Bernoulli, J., & Sylla, E. D. (1713/2006). *The art of conjecturing together with letter to a friend on sets in court tennis*. Translated with introduction and notes by Edith Dudley Sylla. Baltimore: The Johns Hopkins University Press.

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Edinburgh: Oliver and Boyd.

Glantz, S. A. (1980). Biostatistics: How to detect, correct and prevent errors in the medical literature. *Circulation, 61*, 1–7.

ISO 15189:2012. Medical laboratories – Requirements for quality and competence. International Organization for Standardization ISO, Geneva, Switzerland.

Joint Committee for Guides in Metrology (JCGM). (2008). Evaluation of measurement data – Guide to the expression of uncertainty in measurement. BIPM.

Lang, T. A., & Altman, D. G. (2013). Basic statistical reporting for articles published in clinical medical journals: The SAMPL guidelines. In P. Smart, H. Maisonneuve, A. Polderman (Eds.), Science Editors' Handbook. Paris, France: European Association of Science Editors.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Massi Lindsey, L. L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34*, 171–187.

Neyman, J., & Peason, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, 231*, 289–337.

Sackett, D. L., Rosenberg, W. M. C., Muir Gray, J. A., Haynes, R. B., Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal, 312*, 71–72.

von Collani, E., & Dräger, K. (2001). *Binomial distribution handbook for scientists and enegineers*. Boston: Birkhäuser.