XIIth INTERNATIONAL WORKSHOP
on
Intelligent Statistical Quality Control 2016
Hamburg, Germany, August 16 – 19, 2016

organized by

Sven Knoth
Department of Mathematics and Statistics
Helmut Schmidt University
Postfach 700822
22008 Hamburg, Germany

Wolfgang Schmid
Department of Statistics
European University
Große Scharrnstraße 59
15230 Frankfurt (Oder), Germany

PROCEEDINGS

August 11, 2016

Helmut Schmidt University, Hamburg

Contents

Control Charts for Time-Dependent Categorical Processes 1
Christian H. Weiß

A Median Loss Control Chart 21
Su-Fen Yang and Shan-Wen Lu

Bayesian Reliability Analysis of Accelerated Gamma Degradation Processes with Random Effects and Time-scale Transformation 35
Tsai-Hung Fan and Ya-Ling Huang

Sampling inspection by variables under Weibull distribution and Type I censoring 45
Peter-Th. Wilrich

Design of Experiments: A Key to Successful Innovation 65
Douglas C. Montgomery and Rachel T. Silvestrini

Risk-Adjusted Exponentially Weighted Moving Average Charting Procedure Based on Multi-Responses 77
Xu Tang and Fah Fatt Gan

A Note on the Quality of Biomedical Statistics 97
Elart von Collani

Monitoring and diagnosis of causal relationships among variables 111
Ken Nishina, Hironobu Kawamura, Kosuke Okamoto, and Tatsuya Takahashi

Distribution Free Bivariate Monitoring of Dispersion 123
Ross Sparks and Subha Chakraborti

Monitoring of short series of dependent observations using a control chart approach and data mining techniques 143
Olgierd Hryniewicz and Katarzyna Kaczmarek

A Primer on SPC and Web Data 163
Erwin Saniga, Darwin Davis, James Lucas

On the Phase I Shewhart Control Chart Limits for Minimizing Mean Squared Error When the Data are Contaminated 173
Murat Caner Testik, Christian H. Weiß, Yesim Koca, and Ozlem Muge Testik

Optimal Design of the Shiryaev–Roberts Chart: Give Your Shiryaev–Roberts a Headstart 189
Aleksey S. Polunchenko

Optimal Designs of Unbalanced Nested Designs for Determination of Measurement Precision 211
Seiichi Yasui and Yoshikazu Ojima

On ARL-unbiased charts to monitor the traffic intensity of a single server queue 219
Manuel Cabral Morais and Sven Knoth

Statistical process monitoring of multivariate time-between-events data: Problems and possible solutions 245
Chenglong Li, Amitava Mukherjee, Qin Su, and Min Xie

Integrating Statistical and Machine Learning Approaches in Improving Inspection Process 251
Tomomichi Suzuki, Tatsuya Iwasawa, Kenta Yoshida, Natsuki Sano, Mirai Tanaka

A MGF Based Approximation to Cumulative Exposure Models 261
Watalu Yamamoto and Lu Jin

New results for two-sided CUSUM-Shewhart control charts 269
Sven Knoth

An Empirical Bayes Approach for Detecting Changes in the Basal Body Temperature 289
Giovanna Capizzi and Guido Masarotto

A Generalized Likelihood Ratio Test for Monitoring Profile Data 309
Yang Liu, JunJia Zhu and Dennis K. J. Lin

Challenges in Monitoring Non-Stationary Time Series 327
Taras Lazariv and Wolfgang Schmid

Phase I Distribution-Free Analysis with the R Package `dfphase1` 347
Giovanna Capizzi and Guido Masarotto

Big Data Analytics and System Monitoring & Management 369
Kwok Leung Tsui and Yang Zhao

The Variable-Dimension Approach in Multivariate SPC 385
Eugenio K. Epprecht, Francisco Aparisi and Omar Ruiz

Statistical Monitoring of Multi-Stage Processes 399
Emmanuel Yashchin

A Critique of Bayesian Approaches within Quality Improvement 419
G. Geoffrey Vining

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.

Key words: 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, certain 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

Christian H. Weiß
Helmut Schmidt University, Department of Mathematics and Statistics, Postfach 700822, 22008 Hamburg, Germany, e-mail: weissc@hsu-hh.de

measurement arrives, this is used to compute a statistic (possibly also incorporating 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 papers by Woodall (2000), Woodall & 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 about 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 Section 3 below, but with additional inspection error), or Ryan et al. (2011), Weiß (2012) (charts for continuous process monitoring, see Section 4 below); further references can be found in Woodall (1997), Topalidou & 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), 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 & 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 Section 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 Section 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 Section 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 Section 4.

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 enumerable 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 $X_1, \ldots, X_T$.

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 & 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 & 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

$$\kappa(k) \;=\; \frac{\sum_{j=0}^{m}(p_{jj}(k)-\pi_j^2)}{1-\sum_{j=0}^{m}\pi_j^2} \quad \text{and} \quad \hat{\kappa}(k) \;:=\; \frac{1}{T} \;+\; \frac{\sum_{j=0}^{m}(\hat{p}_{jj}(k)-\hat{\pi}_j^2)}{1-\sum_{j=0}^{m}\hat{\pi}_j^2}. \tag{2}$$

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 & 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 $T\,\sigma^2 = 1-(1+2\,\sum_{j=0}^{m}\pi_j^3 - 3\,\sum_{j=0}^{m}\pi_j^2)/(1-\sum_{j=0}^{m}\pi_j^2)^2$, 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 (1-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 & 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 & Lewis (1983)[1], which are motivated by the standard ARMA models for real-valued processes. As shown in Weiß & 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 $(\mathrm{p},0)$ 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 & Göb, 2008):

$$\kappa(k) \;=\; \textstyle\sum_{j=1}^{\mathrm{p}} \phi_j \cdot \kappa(|k-j|) \;+\; \sum_{i=0}^{\mathrm{q}-k} \varphi_{i+k} \cdot r(i) \qquad \text{for } k \geq 1, \tag{6}$$

where the $r(i)$ are determined by $r(i) = 0$ for $i < 0$, $r(0) = \varphi_0$, and

$$r(i) \;=\; \textstyle\sum_{j=\max\{0,i-\mathrm{p}\}}^{i-1} \phi_{i-j} \cdot r(j) \;+\; \sum_{j=0}^{\mathrm{q}} \delta_{i,j} \cdot \varphi_j \qquad \text{for } i > 0.$$

[1] The ARMA model discussed by Biswas & Song (2009) is equivalent to the NDARMA model.

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 & Biswas (2015) 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 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.

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 & Mergen (1987), Bhat & 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 $\rho \in (\max\{\frac{-\pi}{1-\pi}, -\frac{1-\pi}{\pi}\};\ 1)$ 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 $\rho > 0$):

$$E[N_k^{(n)}] \;=\; n\pi, \qquad V[N_k^{(n)}] \;=\; n\pi(1-\pi)\,\frac{1+\rho}{1-\rho}\,\underbrace{\left(1-\frac{2\rho(1-\rho^n)}{n(1-\rho^2)}\right)}_{\approx 1 \text{ for large } n};$$

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 Section 2 before. 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 & 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 even serially independent (so altogether i.i.d.), then the distribution of each $\boldsymbol{N}_k{}^{(n)}$ is a multinomial one.

Remark 1 (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 $\{\boldsymbol{n} \in \{0,\ldots,n\}^{m+1} \mid n_0 + \ldots + n_m = n\}$, and its probability mass function (PMF) is given by

$$P(\boldsymbol{N} = \boldsymbol{n}) \;=\; \text{Bin}omnn_0,\ldots,n_m \cdot \pi_0^{n_0} \cdots \pi_m^{n_m}.$$

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 1, 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), 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 $\frac{4}{n} \left(\sum_{j=0}^{m} \pi_{0;j}^3 - (\sum_{j=0}^{m} \pi_{0;j}^2)^2 \right) / (1 - \sum_{j=0}^{m} \pi_{0;j}^2)^2$, see Weiß (2011), which can be used to determine an appropiate upper limit u_G.

Remark 2 (Ordinal Data). As already briefly pointed out in Section 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 3 (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,

$$\mathbb{S}^{m+1} \;:=\; \{\boldsymbol{x} \in (0;1)^{m+1} \mid x_0 + \ldots + x_m = 1\}.$$

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 & Buccianti (2011). Approaches for monitoring i.i.d. compositional data have been investigated by Boyles (1997), 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 ${N_k}^{(n)}$ 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))

$$c \;:=\; 1 + 2 \cdot \textstyle\sum_{i=1}^{\infty} \kappa(i) \;<\; \infty \qquad (c = 1 \text{ in the i.i.d. case}),$$

then ${C_k}^{(n)}/c$ is approximately χ^2_m-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 $\rho := \phi_1$, i.e., the transition matrix of $(X_t)_{\mathbb{N}}$ is given by

$$\mathbf{P} \;=\; (p_{i|j})_{i,j} \overset{(5)}{=} \begin{pmatrix} \pi_0(1-\rho)+\rho & \pi_0(1-\rho) & \cdots & \pi_0(1-\rho) \\ \pi_1(1-\rho) & \pi_1(1-\rho)+\rho & & \vdots \\ \vdots & & \ddots & \\ \pi_m(1-\rho) & \pi_m(1-\rho) & \cdots & \pi_m(1-\rho)+\rho \end{pmatrix}, \tag{10}$$

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 & 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 & 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 $c\cdot\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is given as in Remark 1. 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 4 (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), 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 1), 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 Section 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 $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.022, 0.023, 0.022, 0.025)^\top$ (Mukhopadhyay, 2008), with dispersion $\nu_{\mathrm{G}} \approx 0.831$, 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 5 (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, 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 asympotic limits are always too large, and now worse with increasing dispersion in $\boldsymbol{\pi}_0$. As an example,

$\boldsymbol{\pi}_0 = (0.83, 0.104, 0.04, 0.026)^\top,\ n = 150,\ \mathrm{ARL}_0 = 370,$

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

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. The above table already indicates 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 Figure 1, which shows the ARL against ρ in the situation $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.022, 0.023, 0.022, 0.025)^\top$ (Mukhopadhyay, 2008) with $n = 150$. On the other hand, this implies that especially the Pearson chart might be used for uncovering increases in ρ.

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 defect is shifted downwards by a certain relative amount, i.e., $\pi_{1;0} = (1-\mathrm{shift})\,\pi_{0;0}$, and all other probabilities are increased in equal measure, $\pi_{1;k} = \frac{1-\mathrm{shift}\cdot\pi_{0;0}}{1-\pi_{0;0}}\,\pi_{0;k}$. 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, Figure 2 shows some ARL graphs for the marginal distribution $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.022, 0.023, 0.022, 0.025)^\top$, where

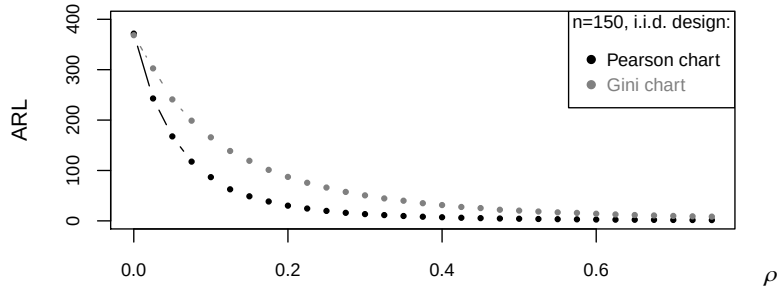


Fig. 1: ARL performance of Pearson ($u_P = 22.41094$) and Gini chart ($u_G = 1.327252$), $n = 150$, $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.022, 0.023, 0.022, 0.025)^\top$.

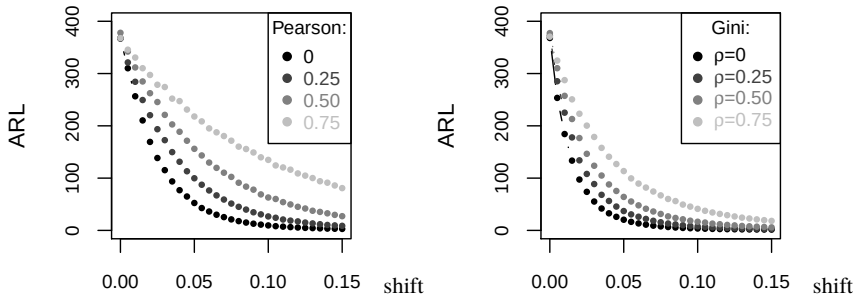


Fig. 2: ARL performance of Pearson and Gini chart ($\text{ARL}_0 \approx 370$) concerning $\pi_{1;0} = (1 - \text{shift})\,\pi_{0;0}$, $n = 150$, $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.022, 0.023, 0.022, 0.025)^\top$.

all charts are designed to give roughly the same in-control ARL. For this particular out-of-control scenario, the Gini chart is preferable, which is reasonable since the dispersion strongly increases with increasing shift size. 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 of a continuous monitoring of the categorical process $(X_t)_{\mathbb{N}}$, i.e., 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 & 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), Weiß & 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 & Stoumbos (1999), and it was extendend to the case of a binary Markov chain as in (7) by Mousavi & Reynolds (2009). Here, the idea is as sketched in Remark 4: 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 4 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 $L_t = \ln(\pi_{1;X_t}/\pi_{0;X_t})$. An alarm is triggered once S_t violates the upper control limit $h > 0$ for the first time.

In analogy to Mousavi & 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 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 1, 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 1), 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 1. 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).

Table 1: 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

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 1. 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 4, 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 2). In view of Remark 3, 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 & 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.

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*. Chapman and Hall Ltd, New York.

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, 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, 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 pages.

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 edition, 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, 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*, Physica Verlag, Heidelberg, 377–397.

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, 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. (2015) Time series analysis of categorical data using auto-odds ratio function. *Statistics*, to appear.

Marcucci, M. (1985) Monitoring multinomial processes. *Journal of Quality Technology* 17(2), 86–91.

Montgomery, D.C. (2009) *Introduction to statistical quality control*. 6th edition, John Wiley & Sons, Inc., New York.

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*, John Wiley & Sons, Ltd., Chichester.

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.

A Median Loss Control Chart

Su-Fen Yang and Shan-Wen Lu

Abstract The quality and loss of products are crucial factors separating competitive companies in many industries. Firms widely employ a loss function to measure the loss caused by a deviation of the quality variable from the target value. Monitoring this deviation from the process target value is important from the view of Taguchi's philosophy. In reality, the distribution of the quality variable may be skewed and not normal, and the in-control process mean may not be the target. We propose a median loss control chart to detect the changes in the process loss center or equivalently the shifts in the process deviation from the mean and target and/or variance for the quality variable with a skewed distribution. We also derive the median loss control chart with variable sampling intervals to detect small shifts in the process loss center. The out-of-control detection performance of the proposed median loss control chart and the median loss chart with variable sampling intervals are illustrated and compared for the process variable with a left-skewed, symmetric or right-skewed distribution. Numerical results show that the median loss chart with variable sampling intervals performs better than the median loss chart in detecting small to moderate shifts in the process loss center or in the difference of mean and target and/or variance of a process variable. The median loss chart and the median loss chart with variable sampling intervals also illustrate the best performance in detection out-of-control process for a process quality variable with a left-skewed distribution.

Su-Fen Yang
Department of Statistics, National ChengChi University, Taipei, Taiwan,
e-mail: yang@mail2.nccu.tw

Shan-Wen Lu
Department of Statistics, National ChengChi University, Taipei, Taiwan,
e-mail: waterolly415@gmail.com

1 Introduction

Control charts are commonly used tools in process signal detection to improve the quality of manufacturing processes and service processes. In the past few years, more and more statistical process control techniques have been applied to the service industry, and control charts are also becoming an effective tool in improving service quality. Tsung et al. (2008), Ning et al. (2009) and Yang and Yang (2013) are some of the few studies covering this area in the literature. Much of the service process data come from processes with variables having non-normal or unknown distributions, and thus the commonly used Shewhart variables control charts, which depend on a normality assumption, are not suitable. Some research has dealt with such a situation; see, for example, Amin et al. (1995); Chakraborti et al. (2001); Altukife (2003); Bakir (2004)(2006); Li, Tang and Ng (2010); Zou and Tsung (2010); Graham et al. (2011); Yang et al. (2011); Yang (2015) and Yang and Arnold (2015).

Product and service qualities and productivity loss are all crucial competitive factors for companies in numerous industries, and the loss function is a popular method for measuring the loss caused by variations in product or service quality. Taguchi (1986) proposed that target values are vital during process specification. Sullivan (1984) emphasized the importance of monitoring deviations from the target value. Because increases in the difference between the mean and the target or variability are the sources of out-of-control loss, it is crucial to monitor the loss variation of a manufacturing or service process. Little research has looked into monitoring a process loss center. Existing loss-function-based control charts assume that the in-control mean of the process quality variable equals the target value - see, for example, Zhang and Wu (2006) and Wu et al. (2009). However, in practice, the in-control process mean may not actually be the process target, and diagnosing the source of an out-of-control signal is crucial for correcting an out-of-control process loss center. Yang (2013a,b) and Yang and Lin (2014) proposed loss-based control charts in order to monitor the loss center caused when quality variables deviate from target values.

A major drawback of the above loss-based control charts is that they all assume the quality variable exhibits a normal distribution. In reality the distribution of the quality variable may be skewed and not symmetric, and hence the sample median is better than the sample average to measure the population center due to its robustness to the outliers (Graham, et al. (2010)), and it can be easily implemented by practitioners. Therefore it is reasonable to use median-type loss control charts to deal with process loss center monitoring.

In this paper we propose using median loss (ML) control chart for variables data to monitor the process loss center, assuming that the underlying distribution of the quality variable is skew-normal. The out-of-control detection performance of the ML chart is measured by the average run length (ARL). Furthermore, we consider the variable sampling intervals (VSI) control scheme for the proposed ML control chart in order to effectively detect small shifts in the process loss center, and investigate its out-of-control detection performance by comparing with some existing control charts.

The paper is organized as follows. Section 2 derives the sampling distribution of

the median loss for a quality variable, X, with a skew-normal distribution. Section 3 designs the ML chart and illustrates its control limits for various sample sizes and the out-of-control detection performance for small to moderate shifts in the difference of mean and target and/or variance. Section 4 constructs a VSI ML chart and measures its out-of-control detection performance for small to moderate shifts in the difference of mean and target and/or variance. Section 5 compares their performances with those of some existing control charts. Section 6 summarizes the findings and provides a recommendation.

2 Median Loss Control Chart

2.1 Skew-normal distribution

We denote random variable X has a skew-normal distribution with location parameter $\xi_0 \in (-\infty, \infty)$, scale parameter $a_0 \in (0, \infty)$, and shape parameter $b \in (-\infty, \infty)$. In other words, $X \sim SN(\xi_0, a_0, b)$. From Azzalini (1985), the probability density function (pdf) of X is

$$f_X(x) = \frac{2}{a_0}\phi(\frac{x-\xi_0}{a_0})\Phi(b\frac{x-\xi_0}{a_0}), \qquad x \in (-\infty, \infty), \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the pdf and cumulated distribution function (cdf) of the standard normal distribution.

In (1) we know that if $b = 0$, then the skew-normal distribution reduces to normal with mean ξ_0 and standard deviation a_0. The distribution is right-skewed if $b > 0$ and left-skewed if $b < 0$.

The cdf of the skew-normal random variable X is

$$F_X(x) = \Phi(\frac{x-\xi_0}{a_0}) - \frac{1}{\pi}\int_0^b \frac{\exp\{\frac{-1}{2}(\frac{x-\xi_0}{a_0})^2(1+y^2)\}}{1+y^2}\,dy, \qquad x \in (-\infty, \infty), \tag{2}$$

The expectation (μ_0) and variance (σ_0^2) of X are

$$\mu_0 = \xi_0 + a_0\frac{b}{\sqrt{1+b^2}}\sqrt{\frac{2}{\pi}}, \quad \sigma_0^2 = a_0^2[1 - \frac{2b^2}{\pi(1+b^2)}].$$

Hence, if we knowμ_0, σ_0^2, and shape b, then we can obtain

$$\xi_0 = \mu_0 - \frac{\sqrt{2}b\sigma_0}{\sqrt{(1+b^2)\pi - 2b^2}}, \quad a_0 = \frac{\sigma_0}{\sqrt{1 - \frac{2b^2}{\pi(1+b^2)}}}.$$

2.2 Taguchi Loss Function

The Taguchi loss function is defined as $L = k(X-T)^2$. Without loss of generality, we set $k = 1$. For $X \sim SN(\xi_0, a_0, b)$, the cdf of loss, $(X-T)^2$, is expressed in eq. (3).

$$\begin{aligned} F_{(X-T)^2}(t) &= P((X-T)^2 \le t) \\ &= \Phi\left(\frac{\sqrt{t}+T-\xi_0}{a_0}\right) - \Phi\left(\frac{T-\sqrt{t}-\xi_0}{a_0}\right) \\ &\quad + \frac{1}{\pi}\int_0^b \frac{e^{-\frac{1}{2}\left(\frac{T-\sqrt{t}-\xi_0}{a_0}\right)^2(1+y^2)} - e^{-\frac{1}{2}\left(\frac{\sqrt{t}+T-\xi_0}{a_0}\right)^2(1+y^2)}}{1+y^2}\,dy, \quad t>0 \end{aligned} \tag{3}$$

2.3 Derivation of the Distribution of Median Loss

Let X_i , $i = 1, 2, ..., n$, be a random sample from the in-control distribution of $SN(\xi, a, b)$. The statistic of sample median loss apparently depends on the sample size being odd or even. Without loss of generality, we only consider the case where the sample size is an odd value to make it easier and faster to compute the sample median loss.

Denote the sample median loss as $ML = (X-T)^2_{\left(\frac{n+1}{2}\right)}$.

The cdf of ML is derived as

$$\begin{aligned} F_{ML}(t) &= \int_0^t f_M(u)\,du \\ &= \frac{n!}{[(\frac{n-1}{2})!]^2}\int_0^t F_{(X-T)^2}(u)^{\frac{n-1}{2}}[1-F_{(X-T)^2}(u)]^{\frac{n-1}{2}} f_{(X-T)^2}(u)\,du \\ &= \frac{n!}{[(\frac{n-1}{2})!]^2} B\left(\frac{1}{a_0\sqrt{t}}[\phi(\tfrac{\sqrt{t}+T-\xi_0}{a_0})\Phi(b\tfrac{\sqrt{t}+T-\xi_0}{a_0}) + \phi(\tfrac{-\sqrt{t}+T-\xi_0}{a_0})\Phi(b\tfrac{-\sqrt{t}+T-\xi_0}{a_0})], \frac{n+1}{2}, \frac{n+1}{2}\right), \\ &\quad t>0, \end{aligned} \tag{4}$$

where $B(x,a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,dt$ is an incomplete beta function.

Since the process parameters $\xi_0 = \mu_0 - \frac{\sqrt{2}b\sigma_0}{\sqrt{(1+b^2)\pi - 2b^2}}$ and $a_0 = \frac{\sigma_0}{\sqrt{1-\frac{2b^2}{\pi(1+b^2)}}}$ are related with in-control population mean and standard deviation, the shifts in the

in-control population mean and/or standard deviation lead to a change in the cdf of ML. Thus, we may construct the ML control chart based on the cdf of ML to monitor changes in the loss center, or equivalently to monitor shifts in the in-control population mean (or the deviation of $\mu_0 - T$) and/or standard deviation.

2.4 Construction of the Median Loss Control Chart

Using eq. (4), we may construct the ML control chart with a specified false alarm rate . The upper control limit (UCL) and the lower control limit (LCL) of the ML chart are determined by taking the inverse cdf of ML - that is:

$$UCL = F_{ML}^{-1}(1 - \frac{\alpha}{2})$$
$$LCL = F_{ML}^{-1}(\frac{\alpha}{2})$$

If the monitoring statistic ML is smaller than LCL or larger than UCL, then the process is deemed to be out-of-control; otherwise the process is in-control. Let δ_3 denote the dispersion parameter that satisfies $\mu_0 - T = \delta_3\sigma_0$. Without loss of generality we set $\delta_3 > 0$. Table 1 gives the control limits of the ML chart for various combinations of $n = 5, 11$, $\delta_3 = 0, 1, 2$, and $b = -500, -2, 0, 2, 500$ when the in-control average run length (ARL_0) is 370.4, $\mu_0 = 0$, and $\sigma_0 = 1$. Note that when $b = -500$ or 500 the skew normal distribution converges to the left half normal distribution and right normal distribution, respectively.

From Table 1 we can see that the width of the control limits becomes narrower when n increases and b and δ_3 are fixed, and the width of the control limits becomes wider when δ_3 increases and n and b are fixed. When $\delta_3 = 0$, the width of the control limits is the widest for a symmetric ($b = 0$) distributed quality variable. When $\delta_3 \neq 0$, the width of the control limits becomes wider for an increasing b or for the distribution of the quality variable changing from left half normal, left-skewed, normal or right-skewed to right half normal.

Table 1 Control Limits of the ML Chart

n	b	δ_3		
		0	1	2
		(LCL , UCL)	(LCL , UCL)	(LCL , UCL)
5	−500	(0.006 , 3.573)	(0.021 , 4.958)	(0.157 , 10.331)
	−2	(0.004 , 3.707)	(0.016 , 6.190)	(0.176 , 12.086)
	0	(0.004 , 3.754)	(0.012 , 6.868)	(0.198 , 13.099)
	2	(0.004 , 3.707)	(0.009 , 7.546)	(0.283 , 14.040)
	500	(0.006 , 3.573)	(0.003 , 8.354)	(0.618 , 15.135)
11	−500	(0.036 , 1.661)	(0.132 , 4.264)	(0.800 , 9.290)
	−2	(0.027 , 2.192)	(0.102 , 4.374)	(0.814 , 9.463)
	0	(0.028 , 2.268)	(0.075 , 4.498)	(0.796 , 9.713)
	2	(0.027 , 2.192)	(0.054 , 4.542)	(0.855 , 9.802)
	500	(0.036 , 1.661)	(0.020 , 4.729)	(0.907 , 10.078)

3 Performance Measurement of the Median Loss Chart

We next use Average Run Length (ARL) to measure the performance of the ML chart. ARL is the average number of samples before the control chart produces a signal, which is the most popular performance measure for a control chart. ARL_0 is fixed at a request level, for example 370.4, while the out-of-control process ARL (ARL_1) is as small as can be. For the ML chart, ARL_0 is

$$ARL_0 = \frac{1}{\alpha} \tag{5}$$

where $\alpha = 1 - P(LCL < ML < UCL | in-control\ ML)$.

We derive the out-of-control distribution of the sample median loss to calculate ARL_1 values of the ML chart. Suppose that X^* is the quality characteristic for the out-of-control process, and $X^* \sim SN(\xi^*, a^*, b)$ with mean $\mu_1 = \mu_0 + \delta_1\sigma_0$, $\delta_1 \neq 0$, and standard deviation $\sigma_1 = \delta_2\sigma_0$, $\delta_2 \neq 1$. We thus now have:

$$\xi^* = \mu_0 + \delta_1\sigma_0 - \frac{\sqrt{2}b\delta_2\sigma_0}{\sqrt{(1+b^2)\pi - 2b^2}},\ a^* = \frac{\delta_2\sigma_0}{\sqrt{1 - \frac{2b^2}{\pi(1+b^2)}}}.$$

Denote the out-of-control median loss as $ML^* = (X^* - T)^2_{(\frac{n+1}{2})}$. We derive the cdf of ML^* as

$$F_{ML^*}(t) = \frac{n!}{[(\frac{n-1}{2})!]^2} B(F_{(X^*-T)^2}(t), \frac{n+1}{2}, \frac{n+1}{2}) \tag{6}$$

where $F_{(X^*-T)^2}(t) = F_{X^*}(\sqrt{t}+\mu_0+(\delta_1-\delta_3)\sigma_0) - F_{X^*}(-\sqrt{t}+\mu_0+(\delta_1-\delta_3)\sigma_0)$ and $t > 0$.

The power that the ML^* is larger than UCL or smaller than LCL is $1-\beta$. In other words, $1-\beta = 1-P(LCL < ML^* < LCL) = F_{ML^*}(LCL)+1-F_{ML^*}(UCL)$. Hence, ARL_1 is:

$$ARL_1 = \frac{1}{1-\beta} = \frac{1}{F_{ML^*}(LCL)+1-F_{ML^*}(UCL)} \tag{7}$$

To investigate the out-of-control detection performance of the proposed chart, we consider the combinations of small to moderate shifts in mean and standard deviation, $\delta_1 = 1.0, 2.0, \delta_2 = 1.0, 1.5, 2.0$, and the dispersion parameter, $\delta_3 = 0, 1, 2$, under the ARL_0 as $370.4, n = 5, \mu_0 = 0$, and $\sigma_0 = 1$, and where the quality variable has left half normal ($b = -500$), left-skewed ($b = -2$), symmetric ($b = 0$), right-skewed ($b = 2$), and right half normal ($b-500$) distributions, respectively. Table 2 gives the ARL_1s of the ML chart for all combinations of $\delta_1 = 1.0, 2.0,\ \delta_2 = 1.0, 1.5, 2.0,\ and\ \delta_3 = 0, 1, 2$.

In Table 2 we see that, no matter for $b = -500, -2, 0, 2, 500$, ARL_1 decreases when δ_1 and/or δ_2 increase under a specified δ_3; ARL1 of the ML chart decreases when δ_3 increases for a specified combination of (δ_1, δ_2, b); the ARL_1s of the ML chart with the left-skewed distributed ($b < 0$) quality variable are all smaller than those of the quality variable with symmetric ($b = 0$) and right-skewed ($b > 0$) distributions; the ARL1s of the ML chart are the smallest for the quality variable with the left half normal distribution ($b = -500$). It suggests that the ML chart should be preferred when the distribution of the process variable is left-skewed, especially for the left half normal distribution.

4 Optimal Variable Sampling Interval Median Loss Chart

Several studies on the performance of adaptive control charts have suggested the use of adaptive control schemes instead of a fixed control scheme. To show better ability in detecting a small or moderate shift of the process loss center, we let the sampling interval be variable for the median loss control chart. In the process control, we adopt two variable sampling intervals: one is long, t_1 , another is short, t_2 . The variable sampling intervals media loss chart (VSI-ML) chart is composed of UCL, warning control limits (WL), and LCL, which are in the form of:

$$\begin{aligned} UCL &= \mu_{ML} + k\sigma_{ML}, \\ WL &= \mu_{ML} + w\sigma_{ML}, \\ LCL &= 0, \end{aligned}$$

where k and w respectively denote the coefficients of UCL and WL with $0 \le w < k$.

Table 2 ARL_1 of the ML Chart

δ_1	δ_2	b	δ_3		
			0	1	2
1	1	−500	13.869	2.059	2.027
		−2	22.527	4.764	4.615
		0	24.152	8.146	8.113
		2	25.040	14.131	14.131
		500	22.726	22.207	22.027
2	1	−500	1.415	1.067	1.065
		−2	1.618	1.164	1.158
		0	1.829	1.313	1.312
		2	2.027	1.588	1.588
		500	2.308	2.308	2.308
1	1.5	−500	3.322	1.780	1.832
		−2	5.242	3.068	3.092
		0	6.591	4.528	4.545
		2	8.421	6.835	6.830
		500	9.799	9.878	9.880
2	1.5	−500	1.401	1.135	1.136
		−2	1.615	1.275	1.271
		0	1.854	1.465	1.464
		2	2.101	1.768	1.768
		500	2.447	2.447	2.447
1	2	−500	2.010	1.577	1.715
		−2	2.757	2.393	2.579
		0	3.317	3.355	3.528
		2	4.176	4.882	4.889
		500	6.860	6.867	6.452
2	2	−500	1.326	1.168	1.190
		−2	1.523	1.335	1.351
		0	1.757	1.554	1.563
		2	2.052	1.874	1.875
		500	2.519	2.522	2.522

In the VSI ML chart, the region between LCL and WL is the "central region" (CR), the region between WL and LCL is the "warning region"(WR), and the region above UCL is the "action region"(AR). We adopt the long sampling interval (t_1) when the statistic falls into CR; the short sampling interval (t_2) when the statistic falls into WR. The VSI chart parameters, k and w , are chosen to satisfy the in-control average time to signal (ATS_0) requirements. Since the time interval between samples is variable and not fixed for the proposed VSI ML chart, a more appropriate out-of-control detection performance index could be the average time to signal (ATS) and not ARL again. A smaller out-of-control ATS (ATS_1) leads to better out-of-control detection performance under a specified in-control ATS (ATS_0).

Under the specified ATS_0 , the procedure to calculate UCL and WL is described

as follows:

Step 1 Specify the values of t_0 , t_1 , t_2 , n , μ_0 , σ_0 , δ_1 , δ_2 , δ_3 , k_U (upper bound of k), and ATS_0 .

Step 2 Divide the interval $(0, UCL)$ into N subintervals, each with an equal width of $2d$, where $d = \dfrac{UCL}{2N}$.

Step 3 Denote the i^{th} interval (State) as $(m_i - d, m_i + d)$ with midpoint m_i, where $m_i = \dfrac{(2i-1)UCL}{2N}$, $i = 1, 2, ..., N$ and the $(N+1)^{th}$ interval (State) as $[UCL, \infty)$.

Step 4
$$ATS_0 = b'(I - Q)^{-1} t_0, \tag{8}$$
where $t_0 = (t_1, t_1, ..., t_2)'$ is an N-vector with element $t_i = t_1$ for $m_1 \in CR$ and $t_i = t_2$ for $m_i \in WR, i = 1, 2$.

Step 5 With ATS_0, t_1 , t_2 , n , μ_0 , σ_0 , and δ_3 , $ATS_0 = b'(I - Q)^{-1} t_0$ is an equation, including the unknown factor w and k of the chart. Use the routine "zreal" in IMSL to find w and k values satisfying the constraint, $0 \le w < k < k_U$, and determine WL and UCL.

When $t_1 = t_2 = t_0$ and $w = 0$, the VSI-ML chart is reduced to a one-sided ML chart.

In reality, engineers may not easily determine the appropriate sampling intervals t_1 and t_2 . Thus, the optimal VSI-ML chart with minimal out-of-control ATS and optimal values of t_1 and t_2 is thus recommended.

The procedure for determining the optimal VSI-ML chart is as follows.

Step 1 Specify the values of t_0 , n , μ_0 , σ_0 , δ_1 , δ_2 , δ_3 , and the requested ATS_0 .

Step 2 Determine the appropriate ranges of the variable sampling intervals and w and k of the VSI-ML chart, $t_L \le t_2 \le t_0 \le t_1 \le t_U$ and $0 \le w < k < k_U$, where the subscript L signifies the lower bound and U is the upper bound.

Step 3 Give the initial values of (t_1, t_2, w, k) . Use an optimization technique, "optim" subroutine in R package, to determine the optimal t_1^* , t_2^* , w^* and k^* that minimize ATS_1 with the constraints described in Step 2 under the requested ATS_0.

Step 4 The optimal VSI-ML chart with optimal t_1^* , t_2^* , w^* and k^* is thus constructed.

To construct the optimal VSI-ML chart, all the combinations of $\delta_1 = 1.0, 2.0$ and $\delta_2 = 1.5, 2.0$ are considered under $ATS_0 = 370, n = 5, \delta_3 = 1, \mu_0 = 0, \sigma_0 = 1, 0 \le t_2 \le t_0 = 1 \le t_1 \le 2$, and $b = -500$ and 0, respectively. Table 3 illustrates the ATS_1s of the optimal VSI ML chart with optimal (t_1, t_2) under the considered combinations of δ_1 and δ_2 when $b = -500, 0$. The last two columns of Table 3 list the ATS_1s of the VSI ML chart with specified $(t_1, t_2) = (2, 0.1)$ and the FSI ML chart. We find that the out-of-control detection performance of the optimal VSI ML chart is a little bit better than those of the specified VSI-ML chart, and that the specified VSI-ML chart is

much better than the FSI ML chart. No matter for the optimal VSI-ML, VSI-ML, or FSI ML chart, the ATS_1s of the process variable with left half normal distribution are smaller than those of the process variable with symmetric distribution. This reveals that the VSI-ML or ML charts are able to effectively monitor the out-of-control process with the left-skewed distributed quality variable, especially for the left half normal distributed quality variable.

5 Performance Comparison

We now compare the out-of-control detection performance of the proposed optimal VSI ML, the specified VSI ML, and ML charts under $b = 0$ with some existing control charts, like MEW chart proposed by Chen et al. (2001), NCS chart addressed by Costa and Rahim (2004), WLC chart proposed by Zhang and Wu (2006), and ERL chart developed by Zhang et al. (2010) with a normal distributed quality variable for $n = 5$, $ATS_0 = 370.4$, and $ARL_0 = 370.4$. For the considered combinations of small to moderate shifts in mean and/or standard deviation, we find that the ATS_1s of the optimal VSI ML and the specified VSI ML charts are all smaller than those of the existing control charts. However, the out-of-control detection performance of the FSI ML chart is always worse compared to the existing control charts for small shifts in mean.

Table 3 ATS_1 of the Optimal VSI-ML, VSI-ML, and FSI ML Charts

b	δ_1	δ_2	Optimal VSI-ML			VSI-ML	FSI-ML
			t_1^*	t_2^*	ATS_1	ATS_1	ATS_1
−500	1	1	1.848	0.000	0.282	0.435	2.059
0			1.993	0.000	0.417	0.925	7.974
−500	2	1	1.989	0.000	0.063	0.173	1.067
0			1.945	0.000	0.068	0.204	1.465
−500	1	1.5	1.848	0.000	0.340	0.474	1.781
0			1.945	0.000	0.709	0.987	4.528
−500	2	1.5	1.959	0.000	0.034	0.145	1.168
0			1.945	0.000	0.043	0.152	1.414
−500	1	2	1.989	0.000	0.275	0.405	1.576
0			1.945	0.000	0.627	0.842	3.355
−500	2	2	1.959	0.000	0.005	0.111	1.135
0			1.945	0.000	0.019	0.122	1.313

Table 4 Performance Comparison with Some Existing Control Charts

b	δ_1	δ_2	ARL_1						
			MEW	NCS	WLC	ELR	OPT.	VSI-ML	VSI-ML	FSI-ML
0	1	1	2.7	7.50	7.50	3.40	0.42	0.93	7.97
		1.5	2.4	2.95	2.95	2.65	0.71	0.99	4.53
		2	1.9	1.80	1.80	1.90	0.63	0.84	3.36
	2	1	1.3	1.20	1.20	1.90	0.07	0.20	1.47
		1.5	1.5	1.30	1.30	1.40	0.04	0.15	1.41
		2	1.3	1.20	1.20	1.30	0.02	0.12	1.31

6 Conclusion

In this paper, we propose the new Median Loss and (Optimal) Variable Sampling Intervals Median Loss Control Charts to simultaneously monitor changes in a loss center or in the process mean and/or variance when the distribution of a quality variable is not symmetric but rather left-skewed or right-skewed. The proposed optimal VSI and VSI ML charts both illustrate better out-of-control detection performance for the left-skewed distributed quality variable. Furthermore, the proposed VSI ML Chart shows better detection ability than the ML Chart in monitoring small to moderate shifts in process mean and/or variance. The optimal VSI ML Chart is thus recommended. A future study could consider to improve the out-of-control detection performance of the proposed Median Loss control charts with an adaptive control scheme on the effect of the contamination.

Acknowledgements The research was partially supported by a MOST 102-2118-M-004-005-MY2 research grant, Taiwan.

References

Amin R, Reynolds MR Jr, Baker S: Nonparametric quality control charts based on the sign statistic. Communications in Statistics – Theory and Methods **24**, 1597–1624 (1995)

Altukife PF: A new nonparametric control charts based on the observations exceeding the grand median. Pakistan Journal of Statistics **19(3)**, 343–351 (2003)

Azzalini, A.: A class of distributions which includes the normal ones. Scandinavian Journal of Statistics **12(2)**, 171–178 (1985)

Bakir ST: A distribution-free Shewhart quality control chart based on signed-ranks. Quality Engineering **16(4)**, 613–623 (2004)

Bakir ST: Distribution free quality control charts based in sign rank like statistics. Communication in Statistics: Theory methods **35**, 743–757 (2006)

Chen G, Cheng SW, Xie H.: Monitoring process mean and variability with one EWMA chart. Journal of Quality Technology **33**, 223–233 (2001)

Chakraborti S, Lann P, Van der Wiel MA: Nonparametric control charts: an Overview and some results. Journal of Quality Technology **33**, 304–315 (2001)

Costa AFB, Rahim MA: Monitoring process mean and variability with one non-central chi- square chart. Journal of Applied Statistics **31**, 1171–1183 (2004)

Graham MA, Human SW, Chakraborti S: A phase I nonparametric Shewhart-type control chart based on the median. Journal of Applied Statistics **37**, 1795–1813 (2010)

Graham MA, Chakraborti S, Human SW: A nonparametric exponentially weighted moving average signed-rank chart for monitoring location. Computational Statistics and Data Analysis **55(8)**, 2490–2503 (2011)

02:Li:Tang:Ng:2010 Li S, Tang L, Ng S: Nonparametric CUSUM and EWMA control charts for detecting mean shifts. Journal of Quality Technology **42(2)**, 209–226 (2010)

Ning X, Shang Y, Tsung F: Statistical process control techniques for service processes: a review. The 6th International Conference on Service Systems and Service Management, Xiamen, China, **April 2009**, 927–931 (2009)

Sullivan LP: Reducing variability: a new approach to quality, Quality Progress **July**, 15–21 (1984)

Taguchi G: Introduction to Quality Engineering: Designing Quality into Products and Processes. (1986)

Tsung F, Li Y, Jin M: Statistical process control for multistage manufacturing and service operations: A review and some extensions. International Journal of Services Operations and Informatics **3**, 191–204 (2008)

Wu Z, Wang P, Wang Q: A loss function-based adaptive control chart for monitoring the process mean and variance, Int. J. Adv. Manuf. Technol. **40**, 948–959 (2009)

Yang SF, Lin JS, Cheng S: A new nonparametric EWMA sign chart. Expert Systems with Applications **38(5)**, 6239–6243 (2011)

Yang SF, Yang CC: Optimal variable sample size and sampling interval MSE chart. The Service Industries Journal. **33(6)**, 652–665 (2013)

Yang SF: Using a new VSI EWMA average loss control chart to monitor changes in the difference between the process mean and target and/or the process variability. Applied Mathematical Modeling **37**, 7973–7982 (2013a)

Yang SF : Using a single average loss control chart to monitor process mean and variability. Communications in Statistics-Simulation and Computation **42**, 1549–1562 (2013b)

Yang SF, Lin L: Monitoring and diagnosing process loss using a weighted-loss control chart. Quality and Reliability Engineering International **30(7)**, 951–959 (2014)

Yang SF: An improved distribution-free EWMA mean chart. Communications in Statistics – Simulation and Computation. **44**, 1–18 (2015)

Yang SF, Arnold BC: A new approach for monitoring process variance. Journal of Statistical Computing and Simulation. Published online. (2015) DOI:10.1080/00949655.2015.1125901.

Zhang, S and Wu, Z: Weighted-loss-function control charts. International Journal of Advanced Manufacturing Technology **31(1)**, 107–115 (2006)

Zhang J, Chou C, Wang Z: A control chart based on likelihood ratio test for monitoring process mean and variability. Quality and Reliability Engineering International **25**, 63–73 (2010)

Zou C, Tsung F: Likelihood ratio-based distribution-free EWMA control charts. Journal of Quality Technology**42(2)**, 1–23 (2010)

Bayesian Reliability Analysis of Accelerated Gamma Degradation Processes with Random Effects and Time-scale Transformation

Tsai-Hung Fan and Ya-Ling Huang

1 Introduction

For highly reliable products, it is quite difficult to obtain their lifetimes through traditional life tests within a reasonable period of time. Alternatively, degradation tests are widely used to assess the lifetime information of highly reliable products that possess quality characteristics that degrade over time and can be related to reliability. Apart for the time to failure itself, degradation tests are also useful in providing additional information regarding the distribution and process of the product lifetime. A detailed explanation on degradation tests in reliability can be seen in Nelson (1990) and Meeker and Escobar (1998).

In degradation tests, degradation measurements of a quality characteristic of each test unit are observed in specified times. When the degradation measurement reaches a pre-stated critical level, the failure is assumed. The performance of a degradation test strongly depends on the suitability of the assumed model of a product's degradation path. For degradation paths involving independently nonnegative increments, gamma processes are more suitable for describing the deterioration of the product. Park and Padgett (2005) provided several new degradation models that incorporate an accelerated test variable based on stochastic processes including a gamma process. Some recent applications of gamma degradation models can be found in Wang (2008) and Tseng *et al.* (2009) and the references therein.

Considering the research on parameter estimation, maximum likelihood estimation (MLE) is often the tool of choice to implement parameter estimation for the stochastic process models. Nowadays, two typical situations are generally encountered in degradation analysis of modern products, i.e. (1) the degradation analysis with sparse/fragmented degradation observations, and (2) the degradation analy-

Tsai-Hung Fan
National Central University, 300 Jhomgda Rd, Jhongli 320 Taiwan, e-mail: thfanncu@gmail.com

Ya-Ling Huang
National Central University, 300 Jhomgda Rd, Jhongli 320 Taiwan, e-mail: erin98102@gmail.com

sis with evolving/updating degradation observations. The first situation is commonly introduced by the reliability analysis of the products that cannot be monitored frequently, such as the underground oil and natural gas pipelines (Qin, *et al.* (2013)). Subjective information or historical information is generally incorporated to complement the insufficiency of these sparse/fragmented degradation observations (Singpuwalla (2005) and Meeker, *et al.* (2005)). In addition, it is hard for the MLE-based method to carry out the degradation analysis under this situation. A degradation analysis for information integration is needed. The second situation is generally introduced by the system health management of the products that are subject to condition monitoring, such as the super luminescent diode analyzed in Wang, *et al.* (2013) and the GaAs Laser discussed in Wang and Xu (2010). The degradation analysis results are updated when newly observed degradation data are available. A degradation analysis method for model updating is needed as well. For the degradation analysis with subjective information and continual monitoring data, Bayesian method has become a standard toolkit. Moreover, the aspects concerning hierarchical priors for random effects information fusion and posterior analysis for degradation analysis results updating were not well studied. An improvement of the random drifts degradation process model and a further extension of the proposed Bayesian method for more general situation is needed.

In many applications, the health or quality of a system is usually quantified in terms of percentage to the initial value. For example, Chaluvadi (2008) presented a dataset of LEDs with light intensities by percent under accelerated degradation life tests, and the failure is declared when the light intensity falls below 50%. Therefore, the data received from the degradation paths must be transformed to fit in the stochastic processes considered. Moreover, the degradation data are collected by time and the stochastic process may not be linear in the true time units. In other words, the time must also be rescaled to fit in the underlying stochastic process. We will develop Bayesian inferences on different lifetime characteristics of the data behaving like the LEDs. However, due to the degradation being bounded in this way between 0 and 1, the development of inferential methods for the lifetime characteristics of LEDs becomes a challenging task. On the other hand, the degradation path may not be linear in time. Lawless and Crowder (2004) made a complex transformation in time to fit the gamma degradation process for the laser data. A simpler model in which the time scale is of power transformation may be considered alternatively.

In this article, we are interested in formulating Bayesian degradation analysis based on the gamma accelerated degradation processes with random effects in which the time scale is of power transformation. Incorporating prior distribution for all the unknown parameters of the underlying model (with or without random effects), we shall conduct a Bayesian reliability analysis for the population lifetime distribution under normal use condition. To identify if random effects model is appropriate, the DIC model selection criterion via MCMC method will be carried out to interpret the model adequacy.

This paper is organized as follows: In Section 2, it introduces the notation and the statistical formulation of the model; In Section 3, it deals with the Bayesian inference of the population lifetime distribution and sequentially predictive analysis of a new

product; In Section 4, data analysis through an illustrated example is provided; In Section 5, some concluding remarks are made.

2 Statistical Model

Assume the degradation path $Y(t)$ follows a gamma process with shape parameter $\alpha > 0$ and rate parameter $\lambda > 0$ which satisfies (i) $Y(0) = 0$; (ii) $Y(t)$ is of the independent increments, i.e. $Y(t) - Y(s)$ is independent of $Y(s)$, for $t > s > 0$ and (iii) $Y(t) - Y(s)$ has a gamma distribution with shape proportional to $\Delta\Lambda = \Lambda(t + \Delta t) - \Lambda(t)$, denoted by Gamma$(\alpha(\Delta\Lambda), \lambda)$ where $\alpha > 0$, and $\lambda > 0$ is the scale parameter. We consider the nonlinear case that $\Lambda(t) = t^c$, $c > 0$ in this paper.

Let $x_1, \ldots, x_I$ be the levels of the accelerating variable, and assume that the shape parameter of the gamma process is log linear in the stress level, and the degradation measurements are observed at $0 = t_0 < t_1 < \ldots < t_J$. Consider that K test products are put under each stress level in a constant-stress degradation test. Specifically, let $Y_k(t)$ be the gamma degradation path for product k and $y_{kij} = Y_{ki}(t_j)$ be the observed data at $t = t_j$ under stress level x_i, respectively, for $j = 1, 2, \ldots, J$, and $i = 1, \ldots, I$ and $k = 1, \ldots, K$. Define $g_{kij} = y_{ki}(t_j) - y_{ki}(t_{j-1})$, then we have g_{kij} has Gamma(α_{ij}, β), where $\alpha_{ij} = \lambda_i(t_j^c - t_{j-1}^c)$, $\lambda_i = a + bx_i$, $c, \beta > 0$ and y_{kij}'s are all independent, for for $j = 1, 2, \ldots, J$, and $i = 1, \ldots, I$ and $k = 1, \ldots, K$. Then given the observed degradation data $\mathbf{g} = \{g_{kij}\}$, and $\mathbf{x} = \{x_i\}$, the likelihood function of (a, b, c, β) is

$$L(a, b, c, \beta | \mathbf{g}, \mathbf{x}) = \prod_{k=1}^{K} \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{\beta^{-\alpha_{ij}}}{\Gamma(\alpha_{ij})} g_{kij}^{\alpha_{ij}-1} e^{-g_{kij}/\beta}. \tag{1}$$

Due to the heterogeneity of the individual degradation paths, the random effects model is taken into account by assuming unit to unit variation to be carried out through random parameter β of the inverse gamma distribution, IG(γ, δ), with pdf

$$f(\beta) = \frac{\gamma^\delta}{\Gamma(\delta)} \beta^{-\delta-1} \exp(-\frac{\gamma}{\beta}),\ \beta, \delta, \gamma > 0.$$

(Lawless and Crowder (2004).) Then the likelihood function of $\theta = (a, b, c, \delta, \gamma)$ is

$$\begin{aligned} L(\theta | \mathbf{g}, \mathbf{x}) &= \prod_{k=1}^{K} \prod_{i=1}^{I} \int_0^\infty \left(\prod_{j=1}^{J} \frac{\beta_{ki}^{-\alpha_{ij}}}{\Gamma(\alpha_{ij})} g_{kij}^{\alpha_{ij}-1} e^{-g_{kij}/\beta_{ki}} \right) \frac{\gamma^\delta}{\Gamma(\delta)} \beta_{ki}^{-\delta-1} \exp(-\frac{\gamma}{\beta_{ki}}) d\beta_{ki} \\ &= \left(\prod_{k=1}^{K} \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{g_{kij}^{-\alpha_{ij}}}{\Gamma(\alpha_{ij})} \right) \frac{\gamma^{KI\delta}}{[\Gamma(\delta)]^{KI}} \frac{\prod_{i=1}^{I} [\Gamma(\lambda_i t_j^c + \delta)]^K}{\prod_{i=1}^{I} [\prod_{k=1}^{K} (\sum_{j=1}^{J} g_{kij} + \gamma)]^{\lambda_i t_j^c + \delta}}. \end{aligned} \tag{2}$$

The major concern in reliability analysis is about the lifetime distribution under normal use condition x_0. The failure time of the product is the first time that the degradation path touches a given threshold Y_f. Thus, the cumulative distribution

function (cdf) of the failure time T under x_0 for the underlying model is

$$\begin{aligned} F_T(t|\boldsymbol{\theta}, x_0) &= P(T \le t|\boldsymbol{\theta}, x_0) \\ &= \int_{-\log Y_f}^{\infty} \frac{y^{\lambda_0 t^c - 1}\gamma^{\delta}}{(y+\gamma)^{\lambda_0 t^c + \delta}} \frac{\Gamma(\lambda_0 t^c + \delta)}{\Gamma(\lambda_0 t^c)\Gamma(\delta)}\, dy \\ &= 1 - F_{d_1,d_2}\left(\frac{\delta Y_f}{\lambda_0 \gamma t^c}\right),\ t > 0, \end{aligned} \tag{3}$$

where $F_(d_1, d_2)(\cdot)$ is the cdf of the F distribution with $d_1 = 2\lambda_0 t^c$ and $d_2 = 2\delta$ being the numerator and denominator degrees of freedom, respectively. Consequently, the failure time distribution under normal use condition with fixed effect is

$$\begin{aligned} F_T(t|\boldsymbol{\theta}, x_0) &= \frac{1}{\Gamma(\lambda_0 t^c)} \int_{-(\log Y_f)/\beta}^{\infty} \beta^{-\lambda_0 t^c} y^{\lambda_0 t^c - 1} e^{-\beta y} dy \\ &= \frac{\Gamma\left(\lambda_0 t^c, -\frac{\log Y_f}{\beta}\right)}{\Gamma(\lambda_0 t^c)},\ t > 0, \end{aligned}$$

where $\Gamma(u, v) = \int_v^{\infty} \xi^{u-1} e^{-\xi} d\xi$, the incomplete gamma function.

3 Bayesian Inference and Model Selection

We consider the Bayesian approach by using independent priors for the parameters a, b, γ and δ. Specifically, since a and b are the regression coefficients associated with the stress variable, consider a and b to have normal priors $N(\mu_a, \sigma_a^2)$ and $N(\mu_b, \sigma_b^2)$, respectively, where μ_a, μ_b and σ_a^2, σ_b^2 are the corresponding means and variances. A conjugate prior of Gamma(u, v) is used for the nuisance scale parameter γ of the random effects and $\log\delta$ has $N(\mu_d, \sigma_d^2)$. Moreover, we consider a mixture distribution of a truncated normal prior and a point mass at 1 for the power transformation parameter c; namely,

$$\pi(c) = pI_{\{1\}}(c) + (1-p)I_{\{1\}^c}(c)\pi_1(c),\ \ 0 < p < 1,$$

where $\pi_1(c) \sim$ Truncated $N(\mu_c, \sigma_c^2)$ on $c > 0$ and $c \neq 1$. Treating all the random effects $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{KI})$ as latent variables, we have the joint posterior

$$\begin{aligned} \pi(\boldsymbol{\theta}, \beta|\mathbf{g}, \mathbf{x}) \propto \prod_{k=1}^{K}\prod_{i=1}^{I}\left(\prod_{j=1}^{J} \frac{\beta_{ki}^{-\alpha_{ij}}}{\Gamma(\alpha_{ij})} g_{kij}^{\alpha_{ij}-1} e^{-g_{kij}/\beta_{ki}}\right) \\ \times \frac{\gamma^{\delta}}{\Gamma(\delta)} \beta_{ki}^{-\delta-1} \exp\left(-\frac{\gamma}{\beta_{ki}}\right) \pi(\boldsymbol{\theta}), \end{aligned}$$

where $\pi(\boldsymbol{\theta})$ is the joint posterior prior of $\boldsymbol{\theta}$ as described above.

The posterior is too complicated to make further inference, but with the aid of the Markov chain Monte Carlo (MCMC) method with Gibbs sampler for $\boldsymbol{\beta}$, γ and c; and the Metropolis-Hastings algorithm for a, b and δ, one can simulate an approximate the posterior sample of θ. To reduce the serial correlation, we take one sample point in several iterations after convergence of the MCMC sequence. Let $\boldsymbol{\theta}^{(m)} = (a^{(m)}, b^{(m)}, c^{(m)}, \gamma^{(m)}, \delta^{(m)})'$, $m = 1, 2, \ldots, M$, be the resulting approximate posterior sample of size M, then the sample mean $\bar{a} = \sum_{m=1}^{M} a^{(m)}/M$ is the usual Bayes estimates of a, for example; while its posterior variance can be approximated by the corresponding sample variance. Moreover, an approximate $100(1-p)\%$ credible interval can be obtained by $(a^{(p/2)}, a^{(1-p/2)})$, where $a^{(p)}$ is the sample p-quantile of $a^{(m)}, m = 1, 2, \ldots, M$, correspondingly. Furthermore, the distribution of the failure time, (3) or (4), under normal use condition is a function of θ, so are the mean time to failure ($MTTF_0$), the reliability function ($R(t|x_0)$) as well as p-quantile ($t_p(x_0)$) of T. Therefore, the predictive distribution of T under x_0 given $\mathbf{g}, \mathbf{x}$ and Δt is

$$F(t|\mathbf{g}, x_0) = \int_0^{\infty} F(t|\theta, x_0)\pi(\theta|\mathbf{g}, \mathbf{x})d\theta.$$

Thus the corresponding Bayesian inference can also be carried out by using the posterior sample, $\theta^{(m)}$, $m = 1, 2, \ldots, M$. For example, $MTTF_0$, $R(t)$ and t_p of the lifetime distribution under x_0 can be estimated by $\hat{E}(T|x_0) = \frac{1}{M}\sum_{l=1}^{M} E(T|\theta^{(l)}, x_0)$, $\hat{R}(t|x_0) = \frac{1}{M}\sum_{l=1}^{M} R(t|\theta^{(l)}, x_0), t > 0$, and $\hat{t}_p(x_0) = \frac{1}{M}\sum_{l=1}^{M} t_p(\theta^{(l)}, x_0)$, $0 < p < 1$, respectively.

There are two model selection issues involved. One is whether the time-scale transformation is necessary and another is the existence of random effects. The weight of the point mass in the mixture distribution of c is an indicator for the time-scale transformation. If the majority of the posterior sample yield $c \neq 1$, we conclude that the power transformation in t is needed and the inference is obtained based on those with $c \neq 1$; otherwise, the model is indeed linear in time and the inference is made through those with $c = 1$ in the posterior sample. On the other hand, we use DIC (Deviance information criterion) to identify existence of random effects among test items. The DIC can be approximated by

$$\widehat{DIC} = \frac{2}{M}\sum_{l=1}^{M}[-2\log L(\boldsymbol{\theta}^{(l)}|\mathbf{g}, \mathbf{x})] - [-2\log L(\frac{1}{M}\sum_{l=1}^{M}\boldsymbol{\theta}^{(l)}|\mathbf{g}, \mathbf{x})],$$

where $\boldsymbol{\theta}^{(l)}, l = 1, \ldots, M$, are the posterior sample. We compute the $\widehat{DIC}$ for the models based on (1) and (2), respectively, and the model with smaller DIC is preferred.

4 Data Analysis

The proposed method is applied to the LED data in Chaluvadi (2008). Under each stress level, $x_1 = 35$ and $x_2 = 40$, 12 LED bulbs were tested and their lightness was

observed in every 50 hours for 5 times. Figure 1 presents the paths of the degradation data, and Figure2 and Figure 3 show the probability plots of the increments fitted by the normal and gamma distributions, respectively. From the plots, we decided to model the degradation paths by the gamma process.

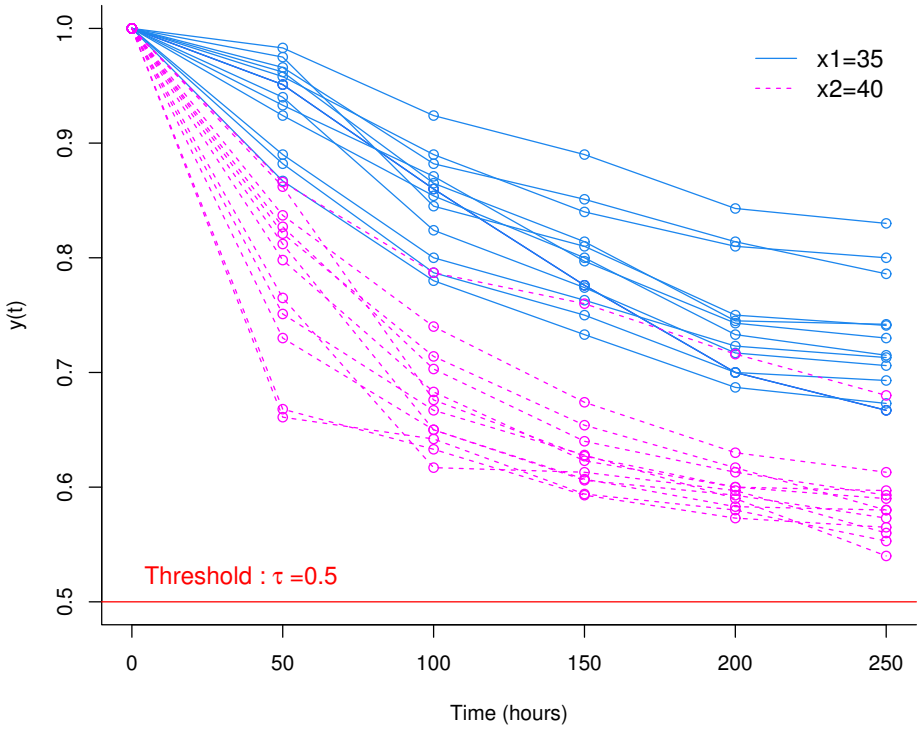


Fig. 1: LED lightness degradation data from Chaluvadi (2008).

In the Bayesian analysis, we considered $(\mu_a, \mu_b, \sigma_a^2, \sigma_b^2)$ = (-15,3, 100,25) and $(\mu_c, \mu_d, \sigma_c^2, \sigma_d^2, u, v)$ = (1, 2.4, 1, $(0.35)^2$,8,4) were used in the random effects model and $\pi(\boldsymbol{\beta}) \sim IG(100, 0.1)$ was considered for the fixed effect model. An MCMC procedure, with normal proposal densities of $\sigma_1 = 0.3$ and $\sigma_2 = 100$, was performed 20000 iterations and the convergence was assured after 5000 iterations by the Gelman-Rubin ratios plots (cf. Gelman and Rubin (1992)). Then one sample was taken in every 10 iterations to get M= 1000 posterior samples afterwards. It concludes that there exists no random effect in this data set based on the DIC criterion and the time scale is about $\Lambda(t) = t^{0.6}$. On the other hand, we applied the function

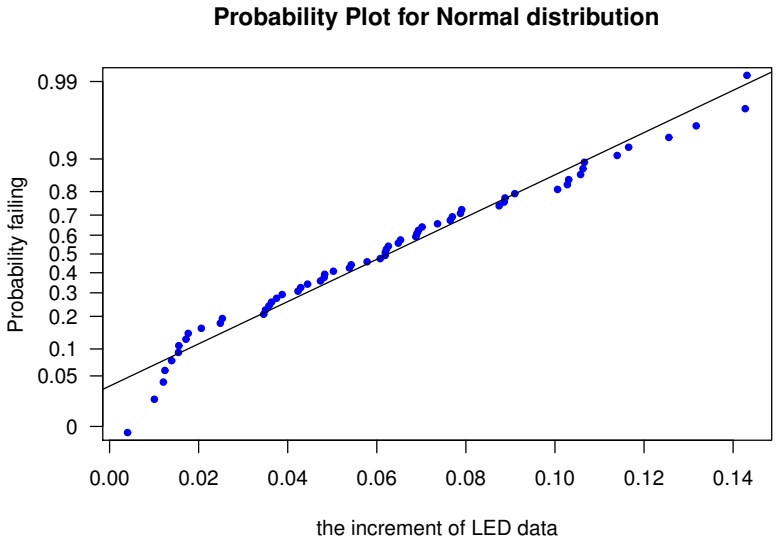


Fig. 2: Probability plots of the increments of the LED data fitted by normal distribution.

optim() in the R-software to get the maximum likelihood estimates (MLE) and the confidence interval was obtained by bootstrap method. The estimation results are presented in Table 1 and the inference of the failure time under normal use condition $x_0 = 25$ is given in Table 2.

Table 1: Parameters estimation (95% interval estimation) of the LED data.

	MLE	Fixed Effect	Random Effects
a	-13.93 (-19.54, -8.32)	-14.87(-15.46, -14.27)	-13.92(-20.03, -8.03)
b	3.63(2.05, 5.20)	3.86(3.63, 4.07)	3.61(1.95, 5.33)
c	0.60(0.53, 0.68)	0.61(0.55, 0.69)	(0.54, 0.70)
β	0.032(0.024, 0.042)	0.033(0.025, 0.042)	–
δ	–	–	251.4(39.63, 647.1)
γ	–	–	8.25(1.26, 21.27)
$\widehat{DIC}$	–	-415.6	-411.3

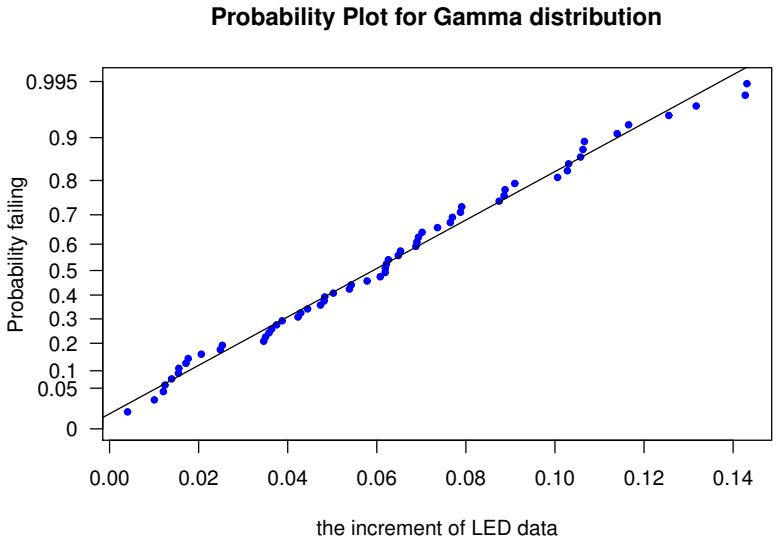


Fig. 3: Probability plots of the increments of the LED data fitted by gamma distribution.

Table 2: Inference on failure time distribution under normal use condition.

	MTTF			$t_{0.1}$		
	Estimate	SE	95% C.I.	Estimate	SE	95% C.I.
MLE	2508.43	873.11	(1458.86, 4651.48)	1470.19	541.06	(840.51, 2982.19)
Fixed Effect	2476.02	641.63	(1508.48, 4004.17)	1473.32	363.83	(901.17, 2354.82)
Random Eff.	2835.47	533.90	(1998.66, 4149.64)	1470.54	245.80	(1067.10, 2049.83)

5 Concluding Remarks

A Bayesian approach is applied to an ADT test under gamma processes with random effects. We use a mixture prior to determine if a time scaling is necessary and then use DIC to identify existence of random effects among the test items. The proposed method is also applied to analysis the LED data. It seems that Bayesian inference can make reliable inference for the failure time distribution under normal use condition under the random effects model; while the conventional ML approach may encounter unstable estimation frequently unless the sample size is large.

Other kinds of random effects may be of practical interest and how to construct a unified approach to identify the time scaling and the random effects simultaneously is under investigation.

References

Chaluvadi, V. N. H.: Accelerated life testing of electronic revenue meters. Ph.D. dissertation, Clemson Univ., Clemson, SC, USA (2008)

Lawless, J., Crowder, M.: Covariates and random effects in a gamma process model with application to degradation and failure. Lifetime Data Analysis. **10**, 213-227 (2004)

Meeker, W. Q., Escobar, L. A.: Statistical Methods for Reliability Data. John Wiley & Sons, New York (1998)

Meeker, W. Q. and Escobar, L. A. and Lu, C. J.: Accelerated degradation tests: modeling and analysis. Technometrics. **40**, 89-99 (2005)

Nelson, W.: Accelerated Testing: Statistical Models, Test Plans, and Data Analysis. John Wiley & Sons, New York (1990)

Park, C., Padgett, W. J.: Accelerated degradation models for failure based on geometric Brownian motion and gamma processes. Lifetime Data Analysis, **11**, 511-527 (2005)

Singpurwalla, N. D.: Survival in dynamic environments. Statistical Science, **10**, 86-103 (1995)

Qin, H, Zhang, S., Zhou, W.: Inverse Gaussian process-based corrosion growth modeling and its application in the reliability analysis for energy pipelines. Front Struct Civil Eng, **7**, 276–87 (2013)

Tseng, S. T., Balakrishnan, N., Tsai, C. C. Optimal step-stress accelerated degradation test plan for gamma degradation processes. IEEE Trans. on Reliability **58**, 611-618 (2009)

Wang, X.: A pseudo-likelihood estimation method for nonhomogeneous gamma process model with random effects. Statistica Sinica, **18**, 1153-1163 (2008)

Wang, L., Pan, R., Li, X., Jiang, T.: A Bayesian reliability evaluation method with integrated accelerated degradation testing and field information. Reliability Engineering and System Safety, **112**, 38-47 (2013)

Wang, X., Xu, D. An inverse Gaussian process model for degradation data. Technometrics, **52**, 188–97 (2010)

Broy, M.: Software engineering — from auxiliary to key technologies. In: Broy, M., Dener, E. (eds.) Software Pioneers, pp. 10-13. Springer, Heidelberg (2002)

Sampling inspection by variables under Weibull distribution and Type I censoring

Peter-Th. Wilrich

Abstract The lifetime (time to failure) of a product is modelled 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 $r = 1$ 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 accepance 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.

Peter-Th. Wilrich
Institut für Statistik und Ökonometrie, Freie Universität Berlin, Garystrasse 21, D-14195 Berlin, Germany, e-mail: wilrich@wiwiss.fu-berlin.de

Key words: 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 non-conforming 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 favourable to apply sampling plans for inspection by attributes. 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, Ohta and Kase (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 paper. The Maximum Likelihood

estimators of the parameters of the Gumbel distribution and the asymptotic variance of the test statistic are derived in Annex A and B, respectively.

2 The model

The lifetime of a product is modelled as a random variable T that is Weibull distributed with 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 probabilöity 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 $p \le p_1$ 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 $t_{(1)} \le t_{(2)} \le \ldots \le t_{(r)}$ 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 paramters 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 $x_L = \ln t_L$,

$$y = \hat{\mu} - k\hat{\sigma} \ge x_L \tag{9}$$

where k is the accepance factor of the sampling plan (k, n, t_C), or equivalently

$$(x_L - \hat{\mu})/\hat{\sigma} \le -k \tag{10}$$

or

$$\hat{p} = F_Z((x_L - \hat{\mu})/\hat{\sigma}) \le F_Z(-k) = p_{crit}, \tag{11}$$

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 $r = 0$ and is very unreliable if $r = 1$ the decision rules (9) and (11) are amended by the rule to accept the lot if $r = 0$ or $r = 1$; 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 probabilty 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 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 standarized 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. The 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 (53):

$$A = \frac{z_{p_1} - z_{p_2}}{u_\alpha - u_{1-\beta}} = \frac{\sqrt{v_{11} + k^2 v_{22} - 2k v_{12}}}{\sqrt{n}} = \frac{f(k, z_C)}{\sqrt{n}} \tag{19}$$

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. The smallest $n = n_{min}$ belongs to $z_C \to \infty$, i.e. the case of no censoring, and according to (54) we obtain it as

$$n_{min} = \frac{1 + \frac{6(k+1-\gamma)^2}{\pi^2}}{A^2} \tag{21}$$

where $\gamma = 0.57721566490\ldots$ is Euler's constant (see Erdéliy (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 section **??**). 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, $z_p = (x_L - \mu)/\sigma_0 = z_{p_{50\%}} = -k$ 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 OCs 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 $P_2(p_2, \beta)$.

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 lifetme 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 $\beta = 0.1$.

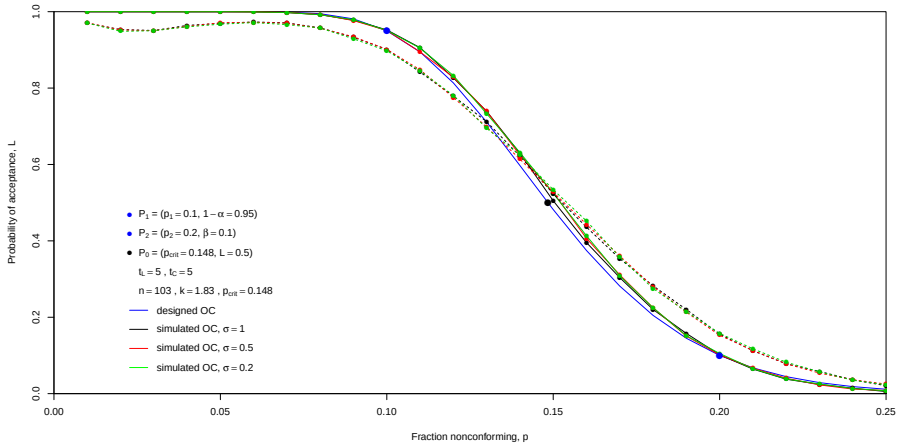


Fig. 1: The asymptotic OC curve (blue) of the sampling plan $(k, n, t_C) = (1.83, 103, 5)$ that passes through the points $P_1(p_1 = 0.1, 1-\alpha = 0.95)$ and $P_2(p_2 = 0.2, \beta = 0.1)$. The black, red, green curves lare the simulated OC curves for $\sigma = 1, 0.5, 0.2$, respectively (solid: numerical acceptance decision, dashed: graphical acceptance decision, see section **??**). Each point represents the average of 10^4 simulation runs.

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 $x_L = \ln t_L = 5$, 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 Figure 1 shows the OC curve of this sampling plan. The two blue points on this

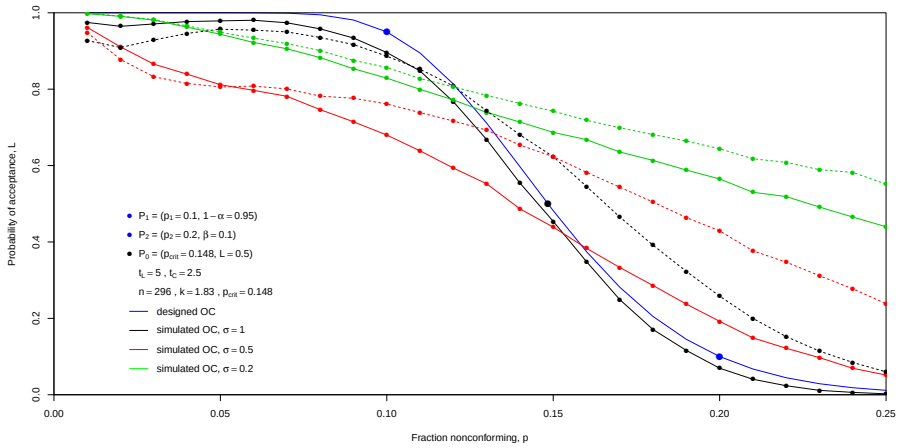


Fig. 2: The asymptotic OC curve (blue) of the sampling plan $(k, n, t_C) = (1.83, 296, 2.5)$ that passes through the points $P_1(p_1 = 0.1, 1-\alpha = 0.95)$ and $P_2(p_2 = 0.2, \beta = 0.1)$. The black, red, green curves are the simulated OC curves for $\sigma = 1, 0.5, 0.2$, respectively (solid: numerical acceptance decision, dashed: graphical acceptance decision). Each point represents the average of 10^4 simulation runs.

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 3 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 the sample size of the attributes sampling plan, $n_{att} = 109$ (orange point in Figure 3), 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 $t_L = 5$. The black, red, green curves of Figure 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 2 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 behaviour 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 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 3 demonstrates another reason for this recommendation: for censoring times decreasing from the specification limit to 0 the sample size increases sharply.

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 $\ln(-\ln(1 - F)$

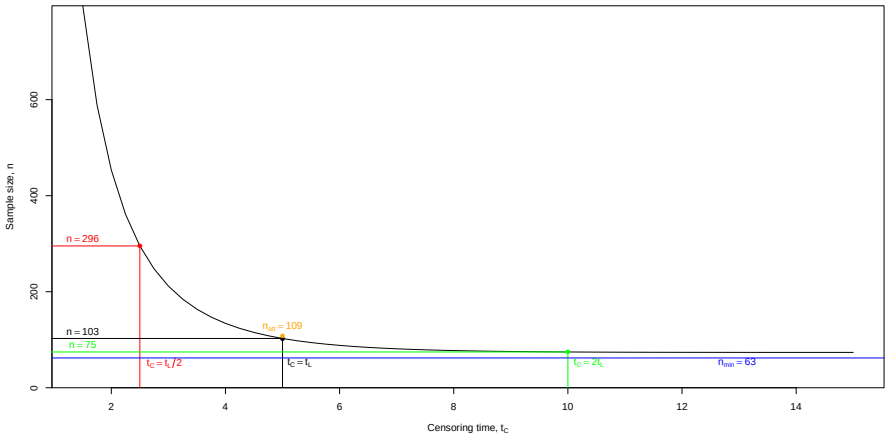


Fig. 3: The sample size n as a function of the censoring time t_C for the sampling plan the OC of which passes through the points $P_1(p_1 = 0.1, 1-\alpha = 0.95)$ and $P_2(p_2 = 0.2, \beta = 0.1)$. For the censoring times $t_C = 2t_L$ (green), $t_C = t_L$ (black), $t_C = t_L/2$ (red) we obtain the sample 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, $n_{att} = 109$.

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 the parameter τ is the lifetime t for which the cumulative distribution is equal to $1-\exp(-1) = 0.632$. Graph paper with such a coordinate system exists as Weibull probability paper.

We can use the Weibull probability paper for the application of the sampling pland 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 section 3. The dashed curves of Figure 1 are the simulated OC

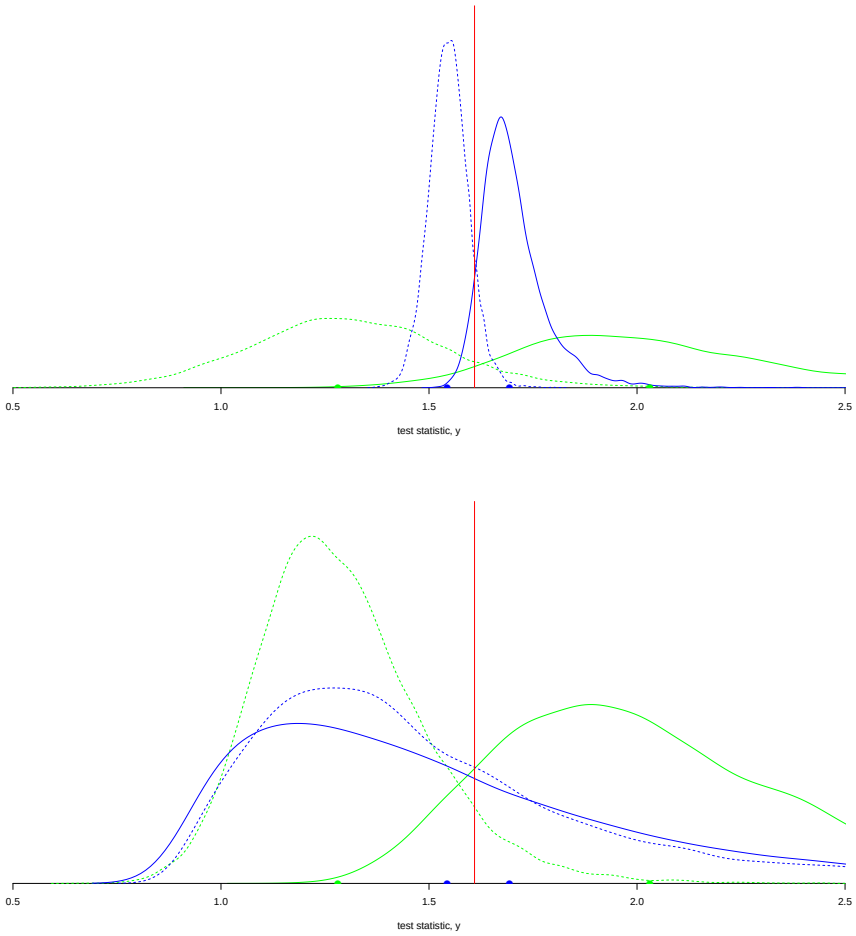


Fig. 4: The distributions of the test statistic y for censoring time $t_C = t_L = 5$ (upper graphs), $t_C = t_l/2 = 2.5$ (lower graphs), $\sigma = 1$ (green), $\sigma = 0.5$ (blue), $p_1 = 0.1$ (solid) and $p_2 = 0.2$ (dashed) obtained by 10^4 simulation runs. The expected values of the test statistic are indicated as points on the horizontal axis. The specification limit $x_L = 1.61(t_L = 5)$ is indicated as red vertical line.

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}.

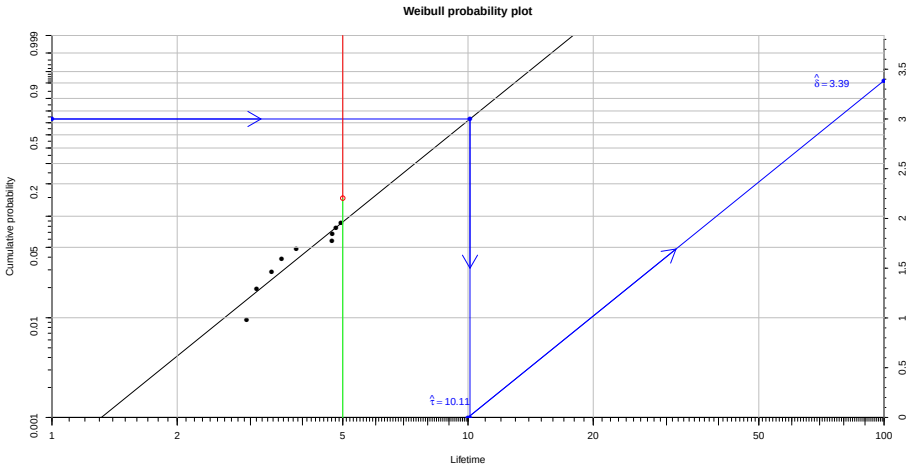


Fig. 5: In this particular example of the application of our sampling plan ($n = 103, k = 1.83, p_{crit} = 0.148, t_L = 5, t_C = 5$) 9 lifetimes $t_{(1)}, \ldots, t_{(9)}$ have been observed and are plotted against $1/(n+1), \ldots, 9/(n+1)$ (black points). The "best fit" straight line (black) intersects 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).

6 Conclusions

The lifetime (time to failure) of a product is modelled 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 $r = 1$ 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 $p \le p_1$ 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 distribtuion. Since these parameters are unknown we assume that the hazard rate of the underlying Weibull distribution is nondecreasing ($\delta \ge 1$). 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% to 30% smaller than that of the corresponding attributes sampling plan.

Annex A: Maximum likelihood estimation of the parameters of the Gumbel distribution

r lifetimes $t_{(1)} \le t_{(2)} \le \ldots \le t_{(r)}$ (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 $t_{(r+1)} \le t_{(r+2)} \le \ldots \le t_{(n)}$ 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

$$\begin{aligned}\frac{\partial l(\mu,\sigma)}{\partial\mu} &= -\frac{1}{\sigma}\left[\sum_{i=1}^{r}(1-\exp(z_i)) - (n-r)\exp(z_C)\right] \\ &= -\frac{1}{\sigma}\left[r - \exp(-\mu/\sigma)\left(\sum_{i=1}^{r}\exp(x_i/\sigma) + (n-r)\exp(x_C/\sigma)\right)\right] \end{aligned} \tag{28}$$

and

$$\begin{aligned}\frac{\partial l(\mu,\sigma)}{\partial\sigma} &= -\frac{r}{\sigma} - \frac{1}{\sigma^2}\left[\sum_{i=1}^{r}(x_i - x_i\exp(z_i)) - (n-r)x_C\exp(z_C)\right] \\ &= -\frac{r}{\sigma} - \frac{\sum_{i=1}^{r}x_i}{\sigma^2} - \frac{\exp(-\mu/\sigma)}{\sigma^2}\left[\sum_{i=1}^{r}x_i\exp(x_i/\sigma) + (n-r)x_C\exp(x_C/\sigma)\right]. \end{aligned} \tag{29}$$

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 equating (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

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{1}{\sigma^2}\left[I\exp(z) + (1-I)\exp(z_C)\right] \tag{38}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial\mu\partial\sigma} &= -\frac{1}{\sigma^2}\left[I(-1+\exp(z)) + (1-I)\exp(z_C)\right] \\ &\quad -\frac{1}{\sigma^2}\left[Iz\exp(z) + (1-I)z_C\exp(z_C)\right] \\ &= -\frac{1}{\sigma^2}\left[-\sigma\frac{\partial l}{\partial\mu} + Iz\exp(z) + (1-I)z_C\exp(z_C)\right] \end{aligned} \tag{39}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial\sigma^2} &= -\frac{1}{\sigma^2}\left[I(1+z-z\exp(z)) - (1-I)z_C\exp(z_C)\right] \\ &\quad -\frac{1}{\sigma^2}\left[I(-1-z+z\exp(z)+1+z^2\exp(z)) - (1-I)(-z_C\exp(z_C) - z_C^2\exp(z_C))\right] \\ &= -\frac{1}{\sigma^2}\left[-2\sigma\frac{\partial l}{\partial\sigma} + I(1+z^2\exp(z)) + (1-I)z_C^2\exp(z_C)\right] \end{aligned} \tag{40}$$

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 $\mathbb{E}(\frac{\partial l}{\partial\sigma}) = 0$:

$$\mathbb{E}\left(\frac{\partial^2 l}{\partial\mu^2}\right) = -\frac{1}{\sigma^2}\left[\int_{-\infty}^{z_C}\exp(z)f_Z(z)dz + (1-F_Z(z_C))\exp(z_C)\right]. \tag{41}$$

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

$$\begin{aligned}\int_{-\infty}^{z_C} \exp(z) f_Z(z) dz &= \exp(z_C) F_Z(z_C) - \int_{-\infty}^{z_C} \left(\exp(z) - f_Z(z)\right) dz \\ &= -(1 - F_Z(z_C) \exp(z_C) + F_Z(z_C)\end{aligned}$$

and hence,

$$\mathbb{E}\left(\frac{\partial^2 l}{\partial \mu^2}\right) = -\frac{1}{\sigma^2} F_Z(z_C) = -\frac{1}{\sigma^2} f_{11}. \tag{42}$$

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

$$\begin{aligned}\int_{-\infty}^{z_C} z \exp(z) f_Z(z) dz &= z_C \exp(z_C) F_Z(z_C) - \int_{-\infty}^{z_C} (1+z)\left(\exp(z) - f_Z(z)\right) dz, \\ \int_{-\infty}^{z_C} (1+z)\left(\exp(z) - f_Z(z)\right) dz &= \underbrace{\int_{-\infty}^{z_C} (1+z)\exp(z) dz}_{J_1} + \underbrace{\int_{-\infty}^{z_C} (1+z) f_Z(z) dz}_{J_2} \\ J_1 &= \exp(z_C) + z_C \exp(z_C) - \exp(z_C) \\ J_2 &= -z_C \exp(z_C)(1 - F_Z(z_C)) + \int_{-\infty}^{z_C} (1+z) f_Z(z) dz \\ &\Longrightarrow \\ \int_{-\infty}^{z_C} (1+z)(\exp(z - f_Z(z)) dz &= z_C \exp(z_C) F_Z(z_C) + \int_{-\infty}^{z_C} (1+z) f_Z(z) dz\end{aligned}$$

and hence,

$$\mathbb{E}\left(\frac{\partial^2 l}{\partial \sigma \partial \sigma}\right) = -\frac{1}{\sigma^2}\int_{-\infty}^{z_C} (1+z) f_Z(z) dz = -\frac{1}{\sigma^2} f_{12}. \tag{43}$$

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 Erdéliy (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

$$\begin{aligned}\int_{-\infty}^{z_C} z^2\exp(z)f_Z(z)dz &= z_C^2\exp(z_C)F_Z(z_C) - \int_{-\infty}^{z_C}(2z+z^2)\exp(z)F_Z(z)dz,\\ &= z_C^2\exp(z_C)F_Z(z_C) - \underbrace{\int_{-\infty}^{z_C}(2z+z^2)\exp(z)dz}_{J_3} + \int_{-\infty}^{z_C}(2z+z^2)f_Z(z)dz,\end{aligned}$$

$$\begin{aligned}J_3 &= 2\int_{-\infty}^{z_C} z\exp(z)dz + z_C^2\exp(z_C) - 2\int_{-\infty}^{z_C} z\exp(z)dz = z_C^2\exp(z_C)\\ &\Longrightarrow\\ \int_{-\infty}^{z_C} z^2\exp(z_C)f_Z(z)dz &= -(1-F_Z(z_C))z_C^2\exp(z_C) + \int_{-\infty}^{z_C}(2z+z^2)f_Z(z)dz\end{aligned}$$

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 (44) and $J_5 = \gamma^2 + \frac{\pi^2}{6}$ (see Erdéliy (1954), p. 149) we get $f_{22,\infty} = (1-\gamma)^2 + \frac{\pi^2}{6}$.

The formulae for f_{11}, f_{12} and f_{22} in (42), (44) and (45) are equivalent to formulae derived in Harter and Moore (1968). The integrals in (44) and (45) cannot be soved directly. Escobar and Meeker (1986) present series expansions

$$\begin{aligned}f_{12} &= F_Z(z_C) + \sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{j!}\left(z_C - \frac{1}{j}\right)(\exp(z_C))^j\\ f_{22} &= 2f_{12} - F_Z(z_C) + \sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{j!}\left(\left(z_C - \frac{1}{j}\right)^2 + \frac{1}{j^2}\right)(\exp(z_C))^j\end{aligned} \tag{46}$$

and recommend to use these series expansions if $z_C < 0$ and to split the integrals into

$$\begin{aligned}\int_{-\infty}^{z_C}(1+z)f_Z(z)dz &= \int_{-\infty}^{1}(1+z)f_Z(z)dz+\int_{1}^{z_C}(1+z)f_Z(z)dz\\ &= 0.2720757938345342+\int_{1}^{z_C}(1+z)f_Z(z)dz\\ \int_{-\infty}^{z_C}(1+z)^2f_Z(z)dz\ n &= \int_{-\infty}^{1}(1+z)^2f_Z(z)dz+\int_{1}^{z_C}(1+z)^2f_Z(z)dz\\ &= 1.475933122158450+\int_{1}^{z_C}(1+z)^2f_Z(z)dz \quad (47)\end{aligned}$$

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), (44), (45), respectively.

The asymptotic covariance matrix of the estimators $\hat{\mu}$ and $\hat{\sigma}$ is the inverse of the Fisher inormation matrix,

$$\mathbf{V} = \begin{pmatrix} \sigma_{\hat{\mu}}^2 & \sigma_{\hat{\mu}\hat{\sigma}} \\ \sigma_{\hat{\mu}\hat{\sigma}} & \sigma_{\hat{\sigma}}^2\end{pmatrix} = \mathbf{F}^{-1} = \frac{\sigma^2}{n}\begin{pmatrix} f_{11} & f_{12} \\ f_{12} & f_{22}\end{pmatrix}^{-1} = \frac{\sigma^2}{n}\begin{pmatrix} v_{11} & v_{12} \\ v_{12} & v_{22}\end{pmatrix}. \quad (49)$$

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

$$\mathbf{F}_\infty = \frac{n}{\sigma^2}\begin{pmatrix} f_{11,\infty} & f_{12,\infty} \\ f_{12,\infty} & f_{22,\infty}\end{pmatrix} = \frac{n}{\sigma^2}\begin{pmatrix} 1 & 1-\gamma \\ 1-\gamma & (1-\gamma)^2+\frac{\pi^2}{6}\end{pmatrix}, \quad (50)$$

and the asymptotic covariance matrix is

$$\mathbf{V}_\infty = {\mathbf{F}_\infty}^{-1} = \frac{\sigma^2}{n}\cdot\frac{6}{\pi^2}\begin{pmatrix} (1-\gamma)^2+\frac{\pi^2}{6} & \gamma-1 \\ \gamma-1 & 1\end{pmatrix}. \quad (51)$$

The asymptotic variance of the test statistic $y = \hat{\mu} - k\hat{\sigma}$ becomes

$$\sigma_y^2 = \sigma_{\hat{\mu}}^2 + k^2\sigma_{\hat{\sigma}}^2 - 2k\sigma_{\hat{\mu}\hat{\sigma}} = \frac{\sigma^2}{n}\left(v_{11}+k^2v_{22}-2kv_{12}\right) = \sigma^2A^2 \quad (52)$$

with

$$A = \frac{\sigma_y}{\sigma} = \frac{\sqrt{v_{11}+k^2v_{22}-2kv_{12}}}{\sqrt{n}} = \frac{f(k,z_C)}{\sqrt{n}} \quad (53)$$

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

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 MILSTD-105D Sampling Plans to Life and Reliabilty Testing.* Office of the Assistant Secretary of Defense (Installations and Logistics), U.S. Government Printing Office, USA.

Goode, H.P., and Kao, J.H.K. (1961). Sampling Plans Based on the Weibull Distribution. *Proceedings of the Seventh National Symposium on Reliability and Quality Control*, 24 - 40.

Goode, H.P., and Kao, J.H.K. (1962). Sampling Procedures and Tables for Life and Reliability Testing Based on the Weibull distribution (Hazard Rate Criterion). *Proceedings of the Eighth National Symposium on Reliability and Quality Control*, 37 - 58.

Goode, H.P., and Kao, J.H.K. (1963). Weibull Tables for Bio-Assaying and Fatigue Testing. *Proceedings of the Ninth National Symposium on Reliability and Quality Control*, 270 - 286.

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.* International Standardisation Organisation, Geneva.

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.* International Standardisation Organisation, Geneva.

Fertig, K.W., and Mann, N.R. (1980). Life-Test Sampling Plans for Two-Parameter Weibull Populations. *Technometrics*, 22, 165 - 177.

Hosono, Y., Ohta, H., and Kase, S. (1981). Design of Single Sampling Plans for Doubly Exponential Characteristics. In *Frontiers in Statistical Quality Control, eds. Lenz, H.J., Wetherill, G.B., and Wilrich, P.T.* Physica-Verlag, Würzburg, Germany.

Schneider, H. (1989). Failure-Censored Variables-Sampling Plans for Lognormal and Weibull Distributions. *Technometrics*, 31, 199 - 206.

Erdélyi, A. (editor) (1954). *Tables of Integral Transforms, vol. I.* McGraw-Hill. London - New York.

Harter, H.L., and Moore, A.H. (1968). Maximum-Likelihood Estimation, From Doubly Censored samples, of the Parameters of the First Asymptotic Distribution of Extreme Values. *J. Am. Statist. Ass.*, 63, 889 - 901.

Escobar, L.A., and Meeker, W.Q. (1986). Elements of the Fisher Information Matrix for the Smallest Extreme Value Distribution and Censored Data. *J. Roy. Statis. Soc. Series C (Applied Statistics)*,35, 80-86.

Design of Experiments: A Key to Successful Innovation

Douglas C. Montgomery and Rachel T. Silvestrini

Abstract Does the use of statistical methodology such as design of experiments stifle innovation? This is an important theme in this paper. 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.

1 Introduction

In June 2007 (http://www.bloomberg.com/news/articles/2007-06-10/at-3m-a-struggle-between-effic
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 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 CEO's as well as a number of business school professors

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

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

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 paper 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. Statistical programs such as Six Sigma and TQM 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 papers (2012a, 2012b), which discuss aspects of statistical engineering and how best to use statistical methods for improved results. Also see the papers by Anderson-Cook et al (2012a, 2012b), Box and Woodall (2012) and Hockman and Jensen (2016). Antony et al 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 paper 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 research department is generally the consumer and the customer of the development department is generally manufacturing or the fulfillment process.

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 paper 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. 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 the final solution. Consequently, 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.

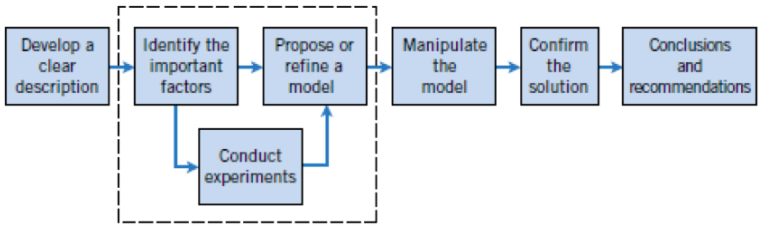


Fig. 1: The Scientific (or Engineering) Method Montgomery & Runger (2014)

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. 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 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."

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 20th 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, they flew kites and eventually gliders at Kitty Hawk, North Carolina, developed a control system 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 frameworks, 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) and Montgomery (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 there 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 will 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 provide 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 paper Jones et al (2015) 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.

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 (2015). Space-filling designs are not recommended for use in modeling response surfaces with low-order polynomials because of undesirable prediction variance properties 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 reason 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 you the ability to iterate through ideas in a very efficient manner. There is always struggle with regards to innovation and invention. The struggle will not 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., MacKay, J., Montgomery, D.C., Parker, P.A., Simpson, J.R., Snee, R.D., Steiner, S.H., Van Mullekom, J., Vining, G.G., and Wilson, A.G. (2012a), "Statistical Engineering—Forming the Foundations", *Quality Engineering*, Vol. 24, No. 2, pp. 110–132.

Anderson-Cook, C.M., Lu, L., Clark, G., DeHart, S.P., Hoerl, R., Jones, B, MacKay, J., Montgomery, D.C., Parker, P.A., Simpson, J.R., Snee, R.D., Steiner, S.H., Van Mullekom, J., Vining, G.G., and Wilson, A.G. (2012b), "Statistical Engineering—Roles for Statisticians and the Path Forward", *Quality Engineering*, Vol. 24, No. 2, pp. 133–152.

Antony, J., Coleman, S., Montgomery, D.C., Anderson, M.J., and Silverstrini, R.T. (2011), "Design of Experiments for Non-manufacturing Processes: Benefits, Challenges and Some Examples", *Journal of Engineering Manufacture*, Vol. 225, No. 11, pp. 2088–2095.

Box, G.E.P. and Woodall, W.H. (2012), "Innovation, Quality Engineering, and Statistics", *Quality Engineering*, Vol. 24, No. 1, pp. 20–29.

Coleman, D. E. and Montgomery, D.C. (1993), "A Systematic Approach to Planning for a Designed Industrial Experiment", (with discussion), *Technometrics*, Vol. 35, No. 1, pp. 1–27.

Hockman, K.K. and Jensen, W.A. (2016), "Statisticians as Innovation Leaders", *Quality Engineering*, Vol. 28, No. 2, pp. 165–174

Hoerl, R. W. and R. D. Snee (2010a), "Moving the Statistics Profession Forward to the Next Level", *The American Statistician*, February 2010, pp. 10–14.

Hoerl, R. W. and R.D. Snee (2010b), "Closing the Gap: Statistical Engineering can Bridge Statistical Thinking with Methods and Tools", *Quality Progress*, May 2010, pp. 52–53.

Johnson, R.T., Montgomery, D.C. and Jones, B. (2011), "An Empirical Study of the Prediction Performance of Space-filling Designs", *International Journal of Experimental Design and Process Optimization*, Vol. 2, pp. 1–18.

Johnson, R.T., Montgomery, D.C., Jones, B. and Parker, P.A. (2010), "Comparing Computer Experiments for Fitting High-Order Polynomial Models", *Journal of Quality Technology*, Vol. 42, No. 1, pp. 86–102.

Jones, B. and Johnson, R.T (2009), "Design and Analysis for the Gaussian Process Model", *Quality and Reliability Engineering International*, Vol. 25, pp. 515–524.

Jones, B. and Montgomery, D.C. (2010), "Alternatives to Resolution IV Screening Designs in 16 Runs", *International Journal of Experimental Design and Process Optimization*, Vol. 1, No. 4, pp. 285–295.

Jones, B. and Nachtsheim, C.J. (2011), "A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects", *Journal of Quality Technology*, Vol. 43, pp. 1–15.

Jones, B., Shinde, S.M., and Montgomery, D.C. (2015), "Alternatives to Resolution III Regular Fractional Factorial Designs for 9–14 Factors in 16 Runs", *Applied Stochastic Models in Business and Industry*, Vol. 31, pp. 50–58.

Jones, B., Silvestrini, R.T., Montgomery, D.C. and Steinberg, D.M. (2015), "Bridge Designs for Modeling Systems with Low Noise", *Technometrics*, Vol. 57, No. 2, pp. 155–163.

Krishnamoorthy, A., Montgomery, D.C., Jones, B., and Borror, C.M. (2015), "Analyzing No-confounding Designs using the Dantzig Selector", *International Journal of Experimental Design and Process Optimization*, Vol. 4, pp. 183–205.

Montgomery, D. C. (1992), "The Use of Statistical Process Control and Design of Experiments in Product and Process Development", *IIE Transactions*, Vol. 24, No. 5, pp. 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)*, Vol. 48, Part 2, pp. 159–177.

Montgomery, D. C. (2012), *Design and Analysis of Experiments*, 8th edition, Wiley, Hoboken, NJ.

Montgomery, D. C. and Runger, G.C. (2014), *Applied Statistics and Probability for Engineers*, 6th edition, John Wiley & Sons, New York.

Shinde, S.M., Montgomery, D.C., and 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*, Vol. 4, No. 1, pp. 1–26.

Silvestrini, R.T., Montgomery, D.C. and Jones, B. (2013), "Comparing Computer Experiments for the Gaussian Process Model Using Integrated Prediction Variance", *Quality Engineering*, Vol. 25, No. 2, pp. 164–174.

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. Risk-adjusted CUSUM charting procedure for monitoring surgical performances has already been developed in the literature. In this paper, we develop a risk-adjusted EWMA charting procedure based on 2 or more outcomes. The properties of this procedure is 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.

Key words: 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

Xu Tang
National University of Singapore, 6 Science Drive 2, Singapore 117546

Fah Fatt Gan
National University of Singapore, 6 Science Drive 2, Singapore 117546, e-mail: staganff@nus.edu.sg

other critical cases. The importance of effective online 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 usually 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 et al. (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 3 or more surgical outcomes, Tang, Gan and Zhang (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, Lin and Loke (2012). In order to improve the effectiveness of this procedure, Tang, Gan and Zhang (2015) developed a risk-adjusted CUSUM procedure based on more than 2 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. Their EWMA chart is only feasible for monitoring surgical performances with 2 outcomes. However, more effective procedures can be obtained by classifying the surgical outcomes into more than 2 outcomes as explained in Tang, Gan and Zhang (2015).

In this paper, we will develop a risk-adjusted EWMA chart based on 2 or more outcomes. In Section 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 is investigated and conditions are derived for it to be a reasonable monitoring statistic. In Section 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 performances of 3 surgeons based on a real data set in Section 4. Similarities and differences between these 2 procedures are compared. Conclusions are given in Section 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, $Y = 1, 2, ..., J-1$ 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 notations used by Tang, Gan and Zhang (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 a historical data set of patients' risk scores and surgical outcomes. The model assumes 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 notations used by Tang, Gan and Zhang (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, Gan and Zhang (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 $R_0 = ... = R_{J-1} = 1$ 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 equation (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 statistic for detecting improvement and deterioration can be determined using equation (2) as

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

and

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

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 2 procedures. We propose the adaptive 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 a mathematical support for the use of this adaptive 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 equations (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 adaptive statistic are given in Theorem 2. The proof of this theorem is given in Appendix 2.

Theorem 2. *Assume equations (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 paper, 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 Figure 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 a reward and

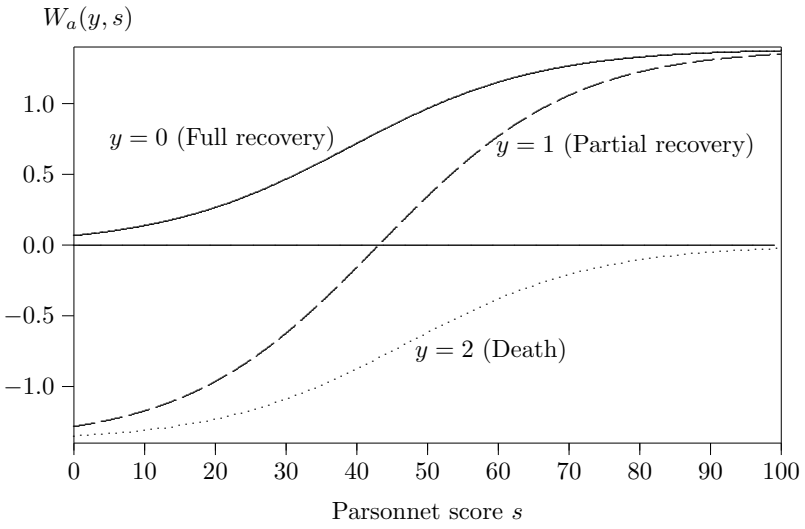


Fig. 1: Plots of $W_a(y,s)$ against the Parsonnet score s for $y = 0,1,2$ when $J = 2$, $R^+ = 2$ and $R^- = 0.5$.

$W_a(J,s)$ 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 $\pi_k(s)$, $k = 0, ..., J$, given the 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 $\pi_k^*(s)$, $k = 0, ..., J$, given the Parsonnet score s, assuming the odds ratios $R_0, R_1, ..., R_{J-1}$ for a surgeon can be determined using equation (3). The penalty-reward score $W_a(y, s)$ can then be calculated using equation (5).

Step 3. Plot $Z_n = \lambda W_a(S_n, Y_n) + (1-\lambda)Z_{n-1}$ against n and signal if $Z_n > H$ of $Z_n < h$.

4 Evaluation of the Performances of 3 Surgeons

In this section, we will construct risk-adjusted EWMA and CUSUM charts of 3 surgeons and compare their performances. These 3 surgeons are among a group of 7 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 3 outcomes is only approximate and quite likely not the best possible classification, a surgeon should be able make a more appropriate classification. For the 3-outcome data, we first fit a proportional odds logistic regression model as

$$\log\Big[\frac{\pi_0(s)}{\pi_1(s)+\pi_2(s)}\Big] = \alpha_0 + \beta s,$$

$$\log\Big[\frac{\pi_0(s)+\pi_1(s)}{\pi_2(s)}\Big] = \alpha_1 + \beta s, \tag{6}$$

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 3 outcomes can be obtained using equation (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 equation (3).

We will highlight the performances of 3 surgeons. The risk-adjusted EWMA charts constructed for surgeons A, B and C are displayed in Figures 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 approximate the first half of the patients but deteriorated for the rest of the patients.

The risk-adjusted CUSUM charts for the 3 surgeons are displayed in Figures 5, 6 and 7 respectively. The upper-sided CUSUM chart is designed to be 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. Even though the CUSUM chart is slightly more sensitive than the EWMA chart, the EWMA chart has the advantage of ease of interpretation.

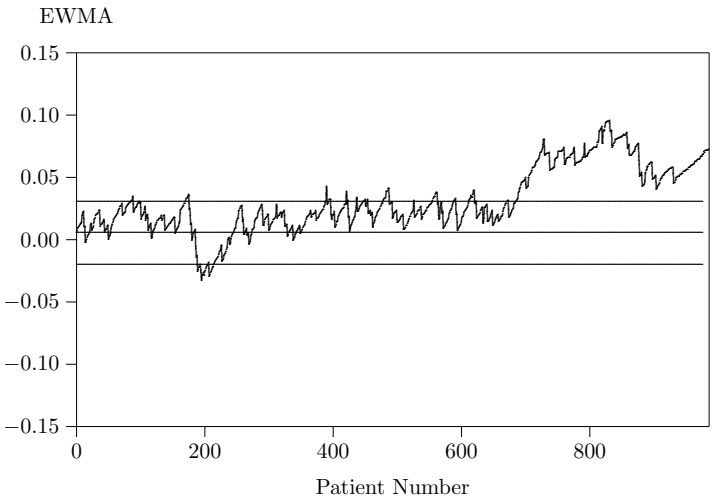


Fig. 2: Plot of risk-adjusted EWMA chart for Surgeon A.

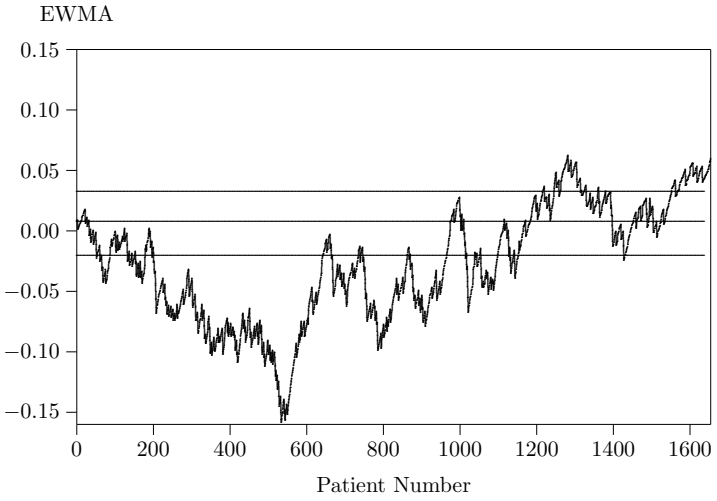


Fig. 3: Plot of risk-adjusted EWMA chart for Surgeon B.

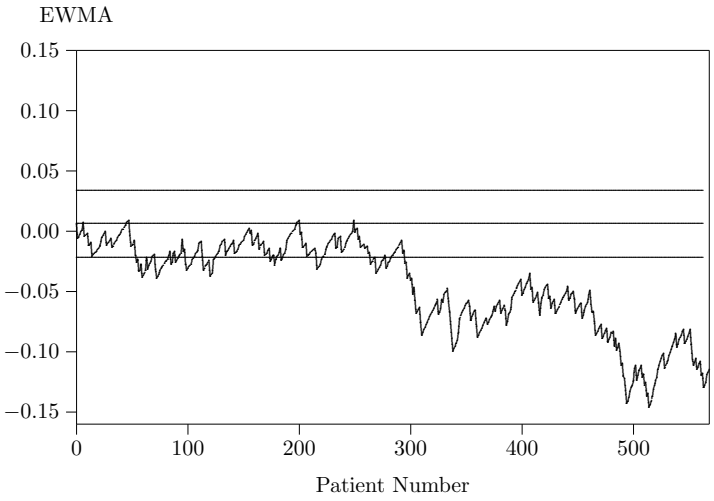
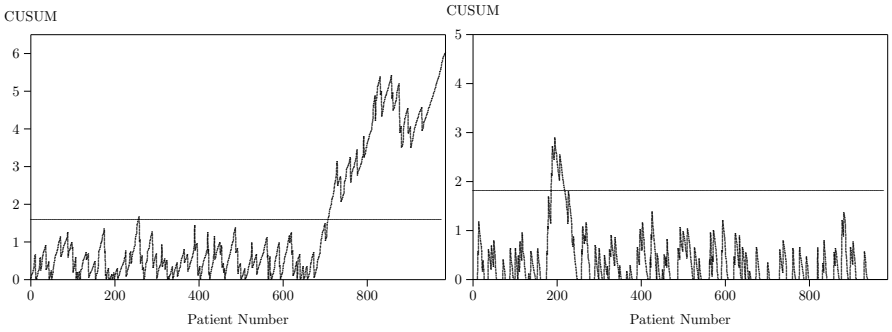


Fig. 4: Plot of risk-adjusted EWMA chart for Surgeon C.

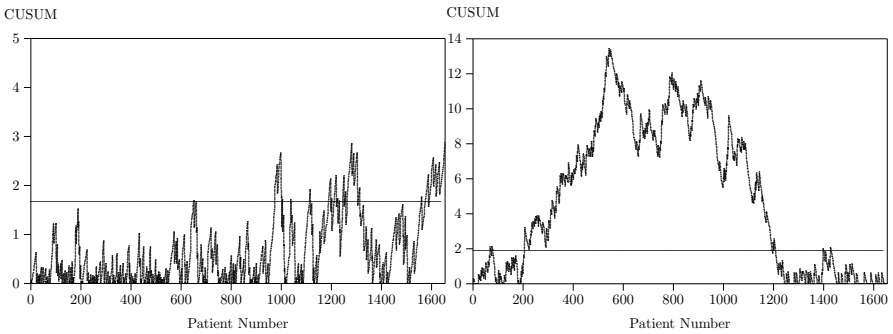


(a) Detecting Improvement ($R = 2$) (b) Detecting Deterioration ($R = 0.5$)

Fig. 5: Risk-adjusted CUSUM charts for detecting improvement ($R = 2$) and deterioration ($R = 0.5$) for Surgeon A.

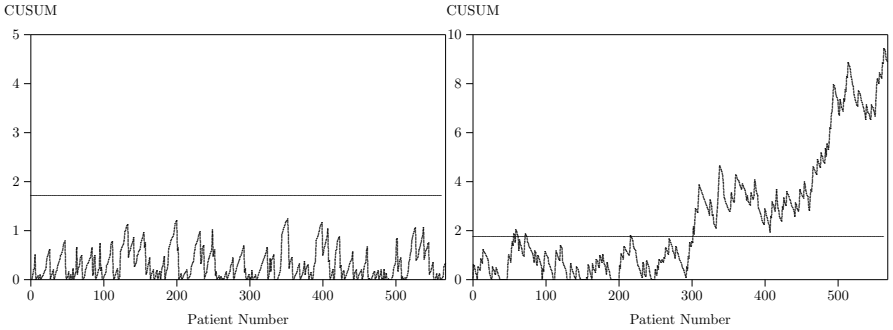
5 Conclusions

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 sur-



(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.

vival. In order to improve the effectiveness of the CUSUM procedure, Tang, Gan and Zhang (2015) developed a risk-adjusted CUSUM procedure based on 3 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 paper, we develop a risk-adjusted EWMA procedure based on 2 or more outcomes. The monitoring statistic is an adaptive statistic obtained by combining the log likelihood ratio statistics for detecting improvement and deterioration. The properties of this statistic is 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 3 surgeons' surgical data. 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 equation (3), we have

$$\log\Big(\sum_{i=0}^{k} \pi_i^+(s)/\big[1 - \sum_{i=0}^{k} \pi_i^+(s)\big]\Big) = \log(R_k^+) + \alpha_k + \beta s.$$

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

$$\begin{aligned} E = \; & \frac{1}{1+\exp(\alpha_k + \log(R_k^+) + \beta s)} - \frac{1}{1+\exp(\alpha_k + \log(R_k^-) + \beta s)} \\ & + \frac{1}{1+\exp(\alpha_{k-1} + \log(R_{k-1}^+) + \beta s)} - \frac{1}{1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)}. \end{aligned}$$

Note that $R_0^+/R_0^- = \cdots = R_{J-1}^+/R_{J-1}^- = R^+/R^- \geq 1$, this imply $E \leq 0$. 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

$$\begin{aligned} g_k(\Delta) = \; & \log(1+\exp(\alpha_{k+1} + \log(R_{k+1}^-) + \beta s)) \\ & - \log(1+\exp(\alpha_{k+1} + \log(R_{k+1}^-) + \Delta + \beta s)) \\ & - \log(1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \beta s)) \end{aligned}$$

$$+\log(1+\exp(\alpha_{k-1}+\log(R_{k-1}^{-})+\Delta+\beta s)).$$

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

$$\frac{\sum_{i=0}^{k}\pi_i^{+}(s)}{1-\sum_{i=0}^{k}\pi_i^{+}(s)} = \frac{R_k^{+}}{R_k^{-}}\frac{\sum_{i=0}^{k}\pi_i^{-}(s)}{1-\sum_{i=0}^{k}\pi_i^{-}(s)}. \tag{A1}$$

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

$$\frac{\pi_0^+(s)}{\pi_0^-(s)} \geq \frac{\pi_0^+(s)+\pi_1^+(s)}{\pi_0^-(s)+\pi_1^-(s)} \geq \cdots \geq \frac{\sum_{i=0}^{J-1}\pi_i^+(s)}{\sum_{i=0}^{J-1}\pi_i^-(s)} \geq \frac{\sum_{i=0}^{J}\pi_i^+(s)}{\sum_{i=0}^{J}\pi_i^-(s)} = 1, \tag{A2}$$

$$\frac{\pi_J^+(s)}{\pi_J^-(s)} \leq \frac{\pi_{J-1}^+(s)+\pi_J^+}{\pi_{J-1}^-(s)+\pi_J^-(s)} \leq \cdots \leq \frac{\sum_{i=1}^{J}\pi_i^+(s)}{\sum_{i=1}^{J}\pi_i^-(s)} \leq \frac{\sum_{i=0}^{J}\pi_i^+(s)}{\sum_{i=0}^{J}\pi_i^-(s)} = 1. \tag{A3}$$

Based on the odds ratio of cumulative probabilities defined in equation (A1) we can obtain

$$\frac{\sum_{i=0}^{k}\pi_i^+(s)}{\sum_{i=0}^{k}\pi_i^-(s)} = \frac{R_k^+/R_k^-}{1-\sum_{i=0}^{k}\pi_i^-(s)+R_k^+/R_k^-\sum_{i=0}^{k}\pi_i^-(s)},\ k=0,\cdots,J-1 \tag{A4}$$

$$\frac{\sum_{i=k+1}^{J}\pi_i^+(s)}{\sum_{i=k+1}^{J}\pi_i^-(s)} = \frac{1}{1-\sum_{i=0}^{k}\pi_i^-(s)+R_k^+/R_k^-\sum_{i=0}^{k}\pi_i^-(s)},\ k=0,\cdots,J-1. \tag{A5}$$

Substitute (A4) into (A2) and (A5) into (A3), we get

$$\frac{R_0^+/R_0^-}{1-\pi_0^-(s)+R_0^+/R_0^-\cdot\pi_0^-(s)} \geq \frac{R_1^+/R_1^-}{1-\sum_{i=0}^{1}\pi_i^-(s)+R_1^+/R_1^-\sum_{i=0}^{1}\pi_i^-(s)} \geq \cdots$$

$$\geq \frac{R_{J-1}^+/R_{J-1}^-}{1-\sum_{i=0}^{J-1}\pi_i^-(s)+R_{J-1}^+/R_{J-1}^-\sum_{i=0}^{J-1}\pi_i^-(s)} \geq 1 \tag{A6}$$

and

$$\frac{1}{1-\sum_{i=0}^{J-1}\pi_i^-(s)+R_{J-1}^+/R_{J-1}^-\sum_{i=0}^{J-1}\pi_i^-(s)} \leq \frac{1}{1-\sum_{i=0}^{J-2}\pi_i^-(s)+R_{J-2}^+/R_{J-2}^-\sum_{i=0}^{J-2}\pi_i^-(s)}$$

$$\leq \cdots \leq \frac{1}{1-\pi_0^-(s)+R_0^+/R_0^-\cdot\pi_0^-(s)} \leq 1. \tag{A7}$$

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 equation (A6)

$$R_0^+/R_0^- \geq R_1^+/R_1^- \geq \cdots \geq R_{J-1}^+/R_{J-1}^- \geq 1. \tag{A8}$$

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 equation (A7)

$$1 \geq R_0^-/R_0^+ \geq R_1^-/R_0^+ \geq \cdots \geq R_{J-1}^-/R_{J-1}^+. \tag{A9}$$

(A8) and (A9) imply

$$R_0^+/R_0^- = R_1^+/R_0^- = \cdots = R_{J-1}^+/R_{J-1}^- \geq 1.$$

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

$$\begin{aligned} W_a(J,s) = &\log(1+\exp(\alpha_{J-1} + \log(R_{J-1}^-) + \beta s)) \\ &- \log(1+\exp(\alpha_{J-1} + \log(R_{J-1}^-) + \Delta + \beta s)) \end{aligned}$$

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 $Y = k \in \{1, \cdots, J-1\}$,

$$\begin{aligned} W_a(k,s) = \Delta &+ \log(1+\exp(\alpha_k + \log(R_k^-) + \beta s)) \\ &- \log(1+\exp(\alpha_k + \log(R_k^-) + \Delta + \beta s)) \\ &- \log(1+\exp(\alpha_{k-1} + \log(R_{k-1}^-) + \Delta + \beta s)). \end{aligned}$$

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

$$\begin{aligned} W_a(J,s) = &\ \log(1+\exp(\alpha_{J-1} + \log(R_{k-1}^-) + \beta s)) \\ &- \log(1+\exp(\alpha_{J-1} + \log(R_{k-1}^-) + \Delta + \beta s)) \\ = &-\Delta + \log(1+\exp(\alpha_{J-1} + \log(R_{k-1}^-) + \beta s)) \\ &- \log(\exp(-\Delta) + \exp(\alpha_{J-1} + \log(R_{k-1}^-) + \beta s)). \end{aligned}$$

Let $s \to -\infty$, then $W_a(J,s) \to -\Delta > 0$.

Appendix 3: Average Run Length of EWMA Chart

Page (1954) introduced a 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 $W_a(Y,S)$. This 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, UK: The Stationery Office, 2001. Available from https://www.bristol-inquiry.org.uk/final_report./

Crowder, Stephen. 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 and Loke, C. K. (2012), "Risk-adjusted Cumulative Sum Charting Procedures", Frontiers in Statistical Quality Control 10, 207–225. Edited by H J Lenz, P Th Wilrich and W Schmid, Berlin-Heidelberg, Germany: Springer Verlag.

Grigg, O. and 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.

Levogrove, J., Vanlencia, O., Treasure, T., Sherlaw-Johnson, C. and Gallivan, S. (1997), "Monitoring the Results of Cardiac Surgery By Variable Life-adjusted Display", *Lancet,* 350, 1128–1130.

Levogrove, J., Sherlaw-Johnson, C., Vanlencia, O., Treasure, T., and Gallivan, S. (1999), "Monitoring the Performance of Cardiac Surgeons", *The Journal of the Operational Research Society*, 50, 684–689.

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. and Bernstein AD. (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. and 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., Cortina, J., David, M., Faichney, A., Gabrielle, F., Gams, E., Harjula, A., Jones, M. T., Pintor, P. P., Salamon, R. and Thulin, L. (1999), "Risk Factors and Outcome in European Cardiac Surgery: Analysis of the EuroSCORE Multinational Database of 19030 Patients", *Eur J Cardiothorac Surg*, 15, 816–822, discussion 822-823.

Steiner, S. H., Cook, R. J., Farewell, V. T. and Treasure, T. (2000), "Monitoring Surgical Performance Using Risk-Adjusted Cumulative Sum Charts", *Biostatistics*, 1, 4, 441–452.

Tang, X., Gan, F. F. and 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. and Sherlaw-Johnson, C. (2004), "Monitoring Cardiac Surgical Performance: A Commentary", *The Journal of Thoracic and Cardiovascular Surgery*, 128, 823–825.

Treasure, T., Taylor, K. and 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).

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 papers confine themselves to describe misunderstandings, misinterpretations and misuses of statistical methods. In contrast, in this paper 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 paper aims at raising the awareness of the statistical community for what is going on in medicine and hoping that this will lead to improvements.

Key words: Laboratory medicine, evidence-based medicine, significance test, probability, Jakob Bernoulli

1 Introduction

During the last five 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.

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:

University Würzburg, Sanderring 2, D-97070 Würzburg, Germany,
e-mail: elart.collani@uni-wuerzburg.de

- Physicians not only feel left alone by statistics, but that statisticians propose the applied methods and interpretations that afterwards are 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*" ISO15189 (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 still be comparable.

In view of this goal, my personal experiences with several private and clinical 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 the

differences might go up to 20% or even 30% from the same blood sample.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 can 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.

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.

In order 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 strengthen misunderstandings. What is therefore needed are simple and straightforward methods to determine the uncertainty of measurements. 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)” BIPM (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.

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 nineties (see Sackett 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 objects. This depends on the specific procedures and especially on the study objective, 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 includes all the requirements, models and statistical methods that are to be used. In contrast to the study design, the evaluation method

is less influenced by the 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 reasons for the large number of articles in medical literature that report on the big rate of medical papers 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 as the following quote from a work by Lang T. und Altman D. Lang & Douglas (2013) show 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 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 evaluation and 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 can try all possible statistical methods, until a procedure is found that leads to a "significance". This result is then 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, of course, 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. Homogeneous 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 many 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 papers 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 Timothy R. Levine et al. 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 paper "On the issue of the Most Efficient tests of Statistical Hypotheses" Neyman &

Pearson (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 19th century in the statistical literature, but the "significance test" was only introduced by R.A. Fisher 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 the predetermined levels of significance. As significance level Fisher set four values, namely 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 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 paper 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 paper 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, 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 40s 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 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 also reflected in the statistical method.

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. 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 test leads 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 a big part of them are adherents 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, Jakob Bernoulli defined the concept probability of a future event as the degree of certitude of its occurrence Bernoulli (2006). 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. It is easy to show that in case of a series of experiments Bernoulli's interpretation coincides with the frequentist interpretation.

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.

References

Jacob Bernoulli: *The Art of Conjecturing together with Letter to a Friend on Sets in Court Tennis. Translated with introduction and notes by Edith Dudley Sylla.* The Johns Hopkins University Press, Baltimore 2006.

Ronald A. Fisher: *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, London, 5th ed., 1934.

Stanton A. Glantz: *Biostatistics: how to detect, correct and prevent errors in the medical literature.* Circulation 61, 1980, 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): *Evaluation of measurement data – Guide to the expression of uncertainty in measurement.* BIPM 2008.

Thomas A. Lang and Douglas G. Altman: *Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines.* In: Smart P, Maisonneuve H, Polderman A (eds). Science Editors' Handbook, European Association of Science Editors, 2013.

Timothy R. Levine, Rene Weber, Craig Hullett, Hee Sun Park and Lisa L. Massi Lindsey: *A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research.* Human Communication Research 34 (2008) 171-187.

Jerzy Neyman and Egon Pearson: *On the Problem of the Most Efficient Tests of Statistical Hypotheses.* Philosophical Transactions of the Royal Society of London. Series A, 231. (1933), 289-337.

David L. Sackett, William M.C. Rosenberg, J.A. Muir Gray, R. Brian Haynes, W. Scott Richardson: *Evidence based medicine: what it is and what it isn't.* (Editorial) BMJ 1996;312:71 - 72.

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 paper, 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 paper 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.

Key words: 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.

Ken Nishina
Nagoya Institute of Technoligy, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan, e-mail: nishina@nitech.ac.jp

Hironobu Kawamura, Kosuke Okamoto, and Tatsuya Takahashi
Nagoya Institute of Technoligy, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan.

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 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 paper, 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 paper 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.

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 2 levels orthogonal array. Our goal in diagnosis is to isolate an unusual causal relationship. In our proposal, isolation of an unusual variable is used as the first step to narrow down the unusual causal relationship.

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 Equation (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 k^{th} eigenvalue of the correlation coefficient matrix. The T^2 and Q statistic are modified slightly from the decomposition as shown in Equation (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 Equation (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 paper, 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 paper, 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:

$$c = \frac{\theta_1\left[(Q/\theta_1)^{h_0} - 1 - \{\theta_2 h_0(h_0-1)/\theta_1^2\}\right]}{\sqrt{2\theta_2 h_0^2}},$$

$$\theta_i = \sum_{r=m+1}^{p} \lambda_r^i \quad (i = 1,2,3), \quad h_0 = 1 - (2\theta_1\theta_3/3\theta_2^2).$$

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 Equation (3), the statistic Q consists of PCSs. The k^{th} 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 j^{th} centralized (or normalized) variable and w_{kj} ($k = 1,2,\ldots,p$) is the element of eigenvector corresponding to the k^{th} largest eigenvalue λ_k. Therefore, the original contribution of the variable x_j to the statistic Q can be measured as shown in Equation (4).

$$\sum_{r=m+1}^{p} (w_{rj}x_j)^2 \quad (j = 1,2,\ldots,p) \tag{4}$$

Higashide et al. (2014) gave slight modification for the original contribution plots as shown in Equation (5).

$$C_j = \sum_{r=m+1}^{p} \{I(r) w_{rj} x_j\}^2 \tag{5}$$

$$I(r) = \begin{cases} 1 & \text{if } \operatorname{sgn}(t_r) = \operatorname{sgn}(w_{rj} x_j) \\ 0 & \text{if } \operatorname{sgn}(t_r) \neq \operatorname{sgn}(w_{rj} x_j) \end{cases} \tag{6}$$

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 2 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.

The response is the Mahalanobis distance between the average of the usual dataset and the i^{th} sample, which indicates the out of control, as follows:

$$D_i = (\boldsymbol{x}_{i(J)} - \bar{\boldsymbol{x}}_{(J)})' \, \boldsymbol{S}_{(J)}^{-1} \, (\boldsymbol{x}_{i(J)} - \bar{\boldsymbol{x}}_{(J)})$$

where $\boldsymbol{x}_{i(J)}$ and $\bar{\boldsymbol{x}}_{(J)}$ is the i^{th} 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.

Table 1: Assignment to L_8 orthognal array for diagnosis of variables

No.	x_1	x_2			x_3			x_4
1	0	0	0	0	0	0	0	
2	0	0	0	1	1	1	1	
3	0	1	1	0	0	1	1	
4	0	1	1	1	1	0	0	
5	1	0	1	0	1	0	1	
6	1	0	1	1	0	1	0	
7	1	1	0	0	1	1	0	
8	1	1	0	1	0	0	1	

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 i^{th} sample $\boldsymbol{x}_i$, which indicates the out of control in Q chart:

$$D_u = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(u)})' \, \boldsymbol{S}^{-1} \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(u)}) \, . \tag{7}$$

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. Fig. 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 Equation (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.

$$u^* = \arg\min_{u} \left[\min_{\alpha} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(u)})' \, \boldsymbol{S}(\alpha)^{-1} \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{(u)}) \right]$$

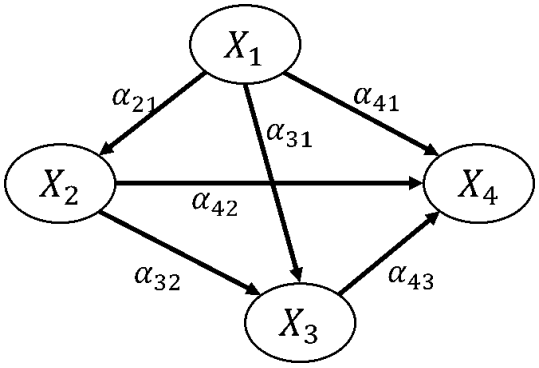


Fig. 1: An example of causal model

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:

$$
\begin{aligned}
X_1 &= \varepsilon_1 \\
X_2 &= \alpha_{21} X_1 + \varepsilon_2 \\
X_3 &= \alpha_{31} X_1 + \alpha_{32} X_2 + \varepsilon_3 \\
X_4 &= \alpha_{41} X_1 + \alpha_{42} X_2 + \alpha_{43} X_2 + \varepsilon_3
\end{aligned}
$$

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 twelve 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 unusual models in Table 2 are determined so that the Q chart can detect the unusuality with about 25% of detection power (ARL is about 4.0).

Table 2: Unusual models in our simulation study

path	pattern 1		pattern 2	
coefficient	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

Our simulation study is carried out as follows: the sample size for determining the control limit, which is shown in Section 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.

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 i^{th} 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 Subsection 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 Equation (8).

The results of the simulation study (the success rate of the isolation) are shown in Table 3. Table 3 indicates that the diagnosis of variables by MT system is better than the modified contribution plots. 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. 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 Subsection 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 $\alpha_{43}$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 Equation (9), 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 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

5 Conclusive remarks

In this paper, 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 paper, 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.

Acknowledgements This work was supported by KAKENHI (25750120).

References

[1]Nishina K., Higashide M., Hasegawa, Y., Kawamura, H. and Ishii, N. (2011): "A paradigm shift from monitoring the amount of variation into monitoring the pattern of variation in SPC," *Proceedings of ANQ Congress Ho Chi Minh City 2011*, VIETNAM.

[2]Tatebayashi, K., Teshima, M. and Hasegawa, Y. (2008): *Nyumon MT system*, Nikkagiren (in Japanese).

[3]Kourti, T. and MacGregor, J. F. (1996): "Multivariate SPC Methods for Process and Product Monitoring," *Journal of Quality Technology*, Vol. 28, No. 4, 409 – 428.

[4]Jackson, J. E. and Mudholkar, G. S. (1979): "Control Procedures for Residuals Associated With Principal Component Analysis," *Technometrics*, Vol. 21, No. 3, 341 – 349.

[5]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*, Vol. 19, No. 4, 213 – 246.

[6]Higashide, M., Nishina, K. and Kawamura, H. (2014): "A Practice of T^2 – Q control Charts in Semiconductor Manufacturing Process," *Quality*, Vol. 44, No. 3, 77 – 86 (in Japanese).

Distribution Free Bivariate Monitoring of Dispersion

Ross Sparks and Subha Chakraborti

Abstract This paper 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 the dispersion of variables from skewed distributions.

Key words: 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 & 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 of the underlying control process of the STP. In the STP application in this paper, the two variables we have near complete data on are TSS and Total Residual Chlorine(TRC), and so we are going use these variables to demonstrate processes for monitoring of 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

Ross Sparks
CSIRO Australia, Data61, 11 Julius Ave, Sydney, Australia, e-mail: Ross.Sparks@csiro.au

Subha Chakraborti
Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa, USA, e-mail: schakrab@cba.ua.edu

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 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.,1992). 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 transformation and then applying the S-chart has been demonstrated as unreliable, particularly for flagging changes in dispersion (see Sparks and Chakraborti, 2016). Therefore alternatives need to be investigated that are more reliable.

This paper therefore 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 skew distribution. In this paper we explore Lui'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 observation, 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 the $E(x_k^{(j)}) = \mu_k^{(j)}$ 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, and therefore $\alpha_k = 0$ 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 paper but note that traditional rational subgroups are size 20 are fairly rare in practice. Our focus on dispersion means we

examine how larger β_k is compared to its expected value of one. If it is significantly larger than one, then the rational subgroup has a larger standard deviation than the in-control data. If it is significantly smaller than one, then the rational subgroup has a smaller standard deviation than the in-control data.

2 Bivariate controls 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 twenty observations were generated from one the distributions. This training data was used to estimate the number of in-control points that are expected to have Lui (1990)'s data depth score of zero where a depth of zero indicates an outlier. If we know that these outlying points don't cluster in a small region in bivariate space then this is likely to be an outbreak in dispersion. In other words, too many extreme points that don't cluster in two dimensional space indicates an increase in dispersion. The Lui (1990)'s depth score is not useful for assessing decreases in dispersion, but therefore 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 Lui (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 comparative 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: twelve of 10 000 rational subgroups had one observation with a depth of zero and one with two zero. Therefore decision rule for out-of-control is taken as either:

1. Any rational subgroup with two or more observations from the rational subgroups of 20 observations have depth equal to zero or
2. Two or more consecutive rational subgroup samples with one depth equal to zero.

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 works reasonable 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 of these across all 1 000 rational subgroups for both bivariate data ($k = 1, 2$), 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 sample 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 has increased.

For each of the ten thousand rational sub-groups (i) estimate the parameters of the following the simple linear model 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 $i = 1, \ldots, 10000$. The median ordered rational subgroup values in the 10 000 simulated data are calculated for both variables $k = 1, 2$. These are used to establish significant increases in dispersion. Next we outline the threshold necessary for delivering an acceptable false discovery rate for flagging significant bivariate changes in dispersion. Denote

$$\mathbb{X}_q = \begin{pmatrix} \mu_1^{(1)} & \mu_2^{(1)} \\ \mu_1^{(2)} & \mu_2^{(2)} \\ \vdots & \vdots \\ \mu_1^{(n)} & \mu_2^{(n)} \end{pmatrix} \quad \text{and} \quad \mathbb{X}_i = \begin{pmatrix} x_{i,1}^{(1)} & x_{i,2}^{(1)} \\ x_{i,1}^{(2)} & x_{i,2}^{(2)} \\ \vdots & \vdots \\ x_{i,1}^{(n)} & x_{i,2}^{(n)} \end{pmatrix}$$

Using the usual quadratic form, we flag significant changes in dispersion when

$$P = (\hat{\beta}_{i1} - 1 \ \hat{\beta}_{i2} - 1)\mathbb{X}_i'\mathbb{X}_i \begin{pmatrix} \hat{\beta}_{i1} - 1 \\ \hat{\beta}_{i2} - 1 \end{pmatrix} / \operatorname{tr}(\mathbb{X}_i'\mathbb{X}_i) \ > \ h\,.$$

An alternative is flag significant changes in dispersion when

$$Q = (\hat{\beta}_{i1} - 1 \ \hat{\beta}_{i2} - 1)\mathbb{X}_q'\mathbb{X}_q \begin{pmatrix} \hat{\beta}_{i1} - 1 \\ \hat{\beta}_{i2} - 1 \end{pmatrix} / \operatorname{tr}(\mathbb{X}_i'\mathbb{X}_i) \ > \ h_q\,.$$

These statistics (P/Q) 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 an ellipse as described in Sparks (1992) then diagnosing the nature of the significant change is easy.

2.3 Transformation to a normal distribution

Here we consider using a Box-Cox transformation to normality and then use charts derived from the normal distribution. The main advantage of this approach is that the plan simply involves finding the appropriate Box-Cox transformation, and then the design of the chart for the normal distribution applies. 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-normal distributed. Then notice that X_2 is log-normal distributed. This means that the thresholds for $\log(X_1)$ and $\log(X_2) = \log(X_1^{\beta} Z) = \beta \log(X_1) + \log(Z)$ is 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 simulate the data $X_1 = X$ and $X_2 = 0.5X + Z$ but this is hidden (unknown), then find $E(f_1(X_1)) = \mu_1$, $E(f_2(X_2)) = \mu_2$, $Var(f_1(X_1)) = \sigma_1^2$, $Var(f_2(X_2)) = \sigma_2^2$ and $Cov(f_1(X_1), f_2(X_2)) = \sigma_{12}$ 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.

Mathematically if we knew the in-control covariance matrix Σ_0 and the transformed sample covariance matrix is $\mathbb{S}$ 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 the determinant of Σ_0 denoted $|\Sigma_0|$ then this is a measure of the volume of space 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. 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 $[\mathbb{X}_1 - \bar{x}_1 \;\; \mathbb{X}_2 - \bar{x}_2]$ 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 charts that were tried in this section are illustrated using Roussouw's bagplot in the appendix. These are distributed from a two parameter distribution denoted $d(a, b)$. The bagplots in the appendix are constructed using 10 000 simulating data using $x \sim d(a, b1)$ and $y \sim 0.5x + z$ where $z \sim d(a, b2)$, and a, $b1$ and $b2$ are defined in the 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 10 000 simulations. The bivariate data (X_1, X_2) is simulated as follows:

$X_1 = X$ and $X_2 = 0.5X + Z$ where X and Z are simulated using a skewed distribution $d(a, b)$

The out-of-control simulated data are in black ink with either both control variables changing when the distribution of X departs from the in-control distribution, or for the second variable (X_2) when only the Z variable changes from its in-control distribution. The thresholds where trained using a bootstrap sample from a Phase I dataset of 10 000 observation from the known but hidden distribution except for the data depth method where the rules defined earlier were used. The in-control false alarms where then checked using in-control data and these are reported in red ink in Table 1, e.g. for log-normal data the false alarms are very similar as 9, 11 and 8 in 10 000 simulations. For the log-normal we assumed that we knew that $X_1 = X$ and $X_2 = 0.5X + Z$ and this helps improve the design of the bivariate control charts otherwise it is difficult to get acceptable false alarm rates (e.g., if we don't assume this knowledge 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 consider 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.

We have only simulated out-of-control data with increased dispersion for the rational subgroup because this is the usual out-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 discovery 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 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.

The robust regression and depth measures seem to be good alternatives. Note that when the change is consistent with the correlation structure (i.e., in the X variable) then generally the robust regression methods is more efficient. While if it is inconsistent 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 are exceptions to this rule. It seems as if a robust plan should involve a combination of depth and robust regression. The advantage 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 Q 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.

3 Example of Application

An application of bivariate control charts is in effluent monitoring of total suspended solids (TSS) 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 TSS and TRC from 1 July 1996 to 24 June 1999. 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 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 TSS was to add 0.1 to this variable and invert it, and the best transform of TRC is the cube root of this measure. The correlation between these two variables is not high at 0.065, but nevertheless

Table 1: The number of flags for increased (decreased) dispersion for the various approaches; log-normal distribution

$d(a,b)$	Box-Cox Transformation		Robust regression	Robust regression (q)	Data depth (2 or more depths=0 or two consecutive one depth=0)
	Log Volume	Log Perimeter	h = 6.259469	h_q = 5.33291	
	thresholds				
	Upper (Lower)	Upper (Lower)			
	27.5601	10.9545			
$LN(\mu,\sigma)$	(7.5393)	(5.6227)			
$X \sim LN(0,1)$, $Z \sim LN(0,\sqrt{0.75})$	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**

Table 2: The number of flags for increased (decreased) dispersion for the various approaches; inverse Gaussian distribution

$d(a,b)$	Box-Cox Transformation		Robust	Robust	Data depth
	-0.06	-0.03	regression	regression (q)	
	Volume	Perimeter	$h = 3.09924$	$h_q = 2.595764$	
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)			
	1.0464	6.8207			
$IG(\mu,\sigma)$	(0.3790)	(6.4754)			
$X \sim IG(1,1)$,	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)			

Table 3: The number of flags for increased (decreased) dispersion for the various approaches; Weibull distribution

$d(a,b)$	Box-Cox Transformation		Robust	Robust	Data depth
	0.27	0.22	regression	regression (q)	
	Volume	Perimeter	$h = 1.707337$	$h_q = 1.470695$	
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)			
	7.8774	8.0077			
$WEI(\mu,\sigma)$	(2.8452)	(6.1948)			
$X \sim WEI(1,1)$,	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)			

positively correlated. We construct bootstrap samples of TSS $= 1/x - 0.13 + 0.023z$ where $x \sim n(3.6, 1.259)$ and $z \sim n(0,1)$ and TRC $= log(y + 1.1) + 0.023z$ where $y \sim n(1.304, 0.49)$ and the small positive correlation is induced by the normally distributed variable z.

This 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 higher false alarm rate than we would like). The two robust regression procedures lead to identical conclusions and therefore only one is reported. In these cases we were able to train the methods to have a false alarm rate of 0.0027. The results are reported in Figure

Table 4: The number of flags for increased (decreased) dispersion for the various approaches; Gamma distribution

$d(a,b)$	Box-Cox Transformation		Robust	Robust	Data depth
	0.32	0.225	regression	regression (q)	
	Volume	Perimeter	$h = 0.6425606$	$h_q = 0.5919475$	
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)			
	7.5292	8.8335			
$Ga(shape, rate)$	(2.6899)	(7.5664)			
$X \sim Ga(3,2)$,	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)			

3. 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 Figure 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).

Table 5: The number of flags for increased (decreased) dispersion for the various approaches; Inverse Gamma distribution

$d(a,b)$	Box-Cox Transformation		Robust	Robust	Data depth
	-0.15	-0.2	regression	regression (q)	
	Volume	Perimeter	h = 740.709	h_q = 659.1923	
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)			
	2.3849	6.1934			
$IGa(\mu,\sigma)$	(0.8289)	(4.4893)			
$X \sim IGa(2,1)$, $Z \sim IGa(2,\sqrt{0.75})$	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**

Table 6: The number of flags for increased (decreased) dispersion for the various approaches; Pareto distribution

$d(a,b)$	Box-Cox Transformation		Robust regression	Robust regression (q)	Data depth
	log Volume	log Perimeter	h = 74930.24	h_q = 39975.8	
	thresholds				
	Upper (Lower)	Upper (Lower)			
$PARETO2 = Pa$	76.9411	19.9683			
$Pa(\mu,\sigma)$	(20.0488)	(9.7517)			
$X \sim Pa(2,1)$, $Z \sim Pa(2,\sqrt{0.5})$	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

4 Concluding remarks

The distribution free method proposed in this paper 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 discover 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

Table 7: The number of flags for increased (decreased) dispersion for the various approaches; Reverse Gumbel distribution

$d(a,b)$	Box-Cox Transformation		Robust	Robust	Data depth
	-2.68	-2.61	regression	regression (q)	
	Volume	Perimeter	$h = 0.7894239$	$h_q = 0.6935926$	
	thresholds				
	Upper	Upper			
	(Lower)	(Lower)			
	1.2774	0.4498			
$RG(\mu,\sigma)$	(0.4452)	(0.3397)			
$X \sim RG(13,1)$, $Z \sim RG(13,\sqrt{1.5})$	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

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. 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

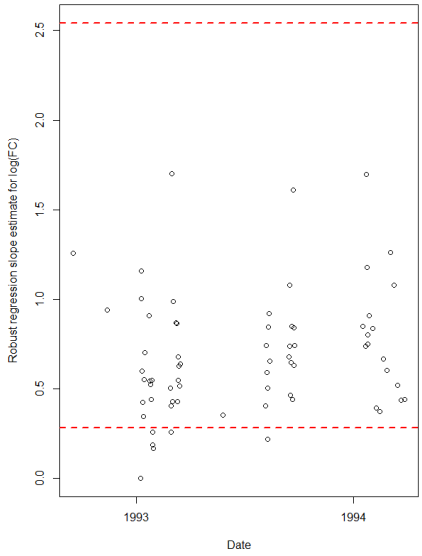


Fig. 1: Robust regression control chart applied to Faecal Coliform measures in Sydney Harbour

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. and Cox, D. R. (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, **26**, 211-252.

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.

Morrison, J. (1958). The lognormal distribution in quality control. *Applied Statistics*, 160-172.

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,

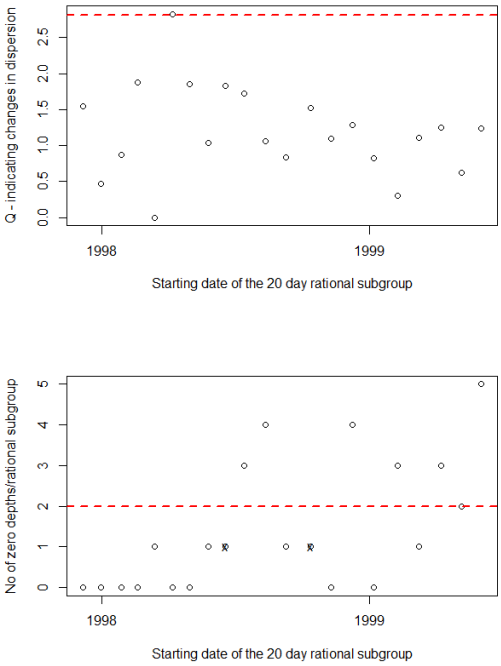


Fig. 2: Bivariate application flagging changes in dispersion of TSS and TRC

and phosphorus. *Water Research*, **32**(5), 1455-1462.

Ferrell, E.B. (1958). Control charts for lognormal universe. *Industrial Quality Control*, **15**, 4-6.

Cheng, S. W., & Xie, H. (2000). Control charts for lognormal data. *Tamkang Journal of Science and Engineering*, **3**(3), 131-138.

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.

Park, G.S. (2007). The role and distribution of total suspended solids in the macrotidal coastal waters of Korea. *Environ Monit Assess*, **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.

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. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Fourth edition. Springer.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel (1986). *Robust Statistics: The Approach based on Influence Functions*. Wiley.

Gibbons, J.D. and Chakraborti, S. (2010). *Nonparametric Statistical Inference.* Fifth Edition. Chapman & Hall.

Hajek, J. (1969). *A course in nonparametric statistics.* Holden-Day. San Francisco.

Liu, R.Y. (1990). On a notion of data depth based on random simplices. *Ann. Stat.*, **18**(1), 405-414.

5 Bivariate Distributions and their bagplots

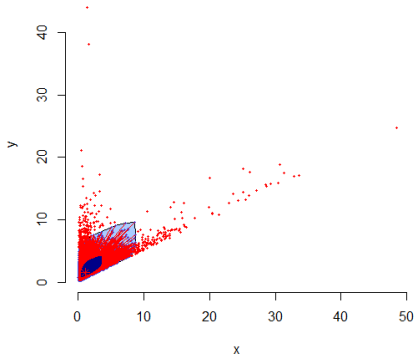


Fig. 3: Log-normal distribution ($X_1 \sim \text{LN}(\mu = 0, \sigma = 1)$ & $X_2 = 0.5X_1 + Z$ where $Z \sim \text{LN}(\mu = 0, \sigma = \sqrt{0.75})$

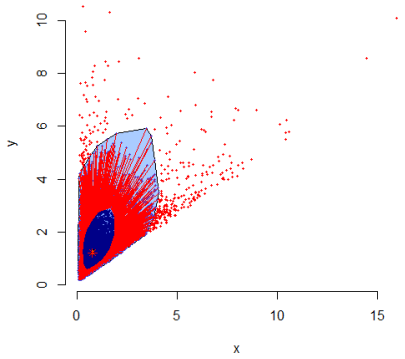


Fig. 4: Inverse Gaussian ($X_1 \sim \text{IG}(\mu = 1, \sigma = 1)$ & $X_2 = 0.5X_1 + Z$ where $Z \sim \text{IG}(\mu = 1, \sigma = \sqrt{0.75})$

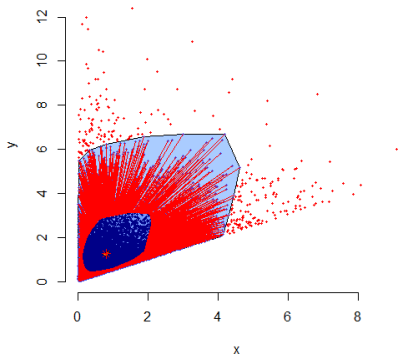


Fig. 5: Weibull distribution ($X_1 \sim \text{WEI}(\mu = 1, \sigma = 1)$ & $X_2 = 0.5X_1 + Z$ where $Z \sim \text{WEI}(\mu = 1, \sigma = \sqrt{0.75})$

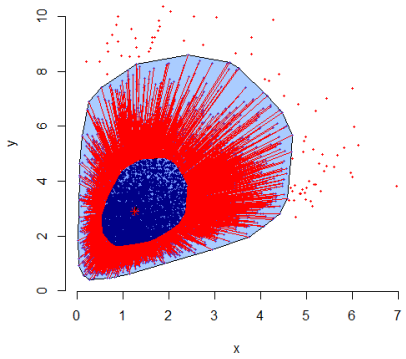


Fig. 6: Gamma distribution ($X_1 \sim \text{Ga}(shape = 3, rate = 2)$ & $X_2 = 0.5X_1 + Z$ where $Z \sim \text{Ga}(shape = 3, rate = \sqrt{1.75})$

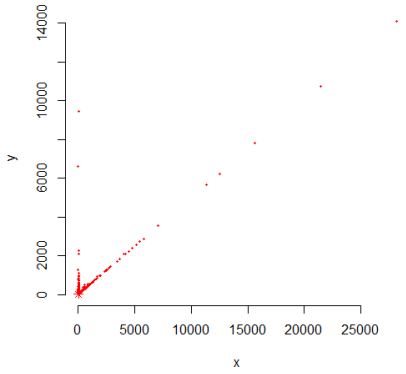


Fig. 7: Inverse Gamma distribution ($X_1 \sim \text{IGa}(\mu = 2, \sigma = 2)$ & $X_2 = 0.5X_1 + Z$ where $Z \sim \text{IGa}(\mu = 2, \sigma = \sqrt{0.75})$

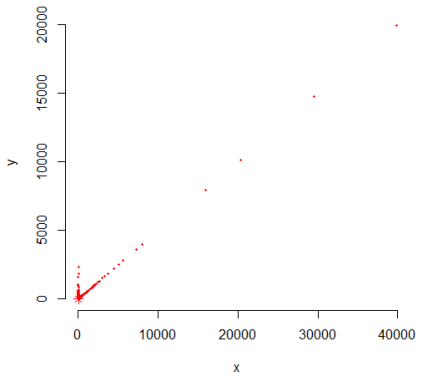


Fig. 8: Pareto distribution ($X_1 \sim \text{PARETO2}(\mu = 2, \sigma = 1)$ & $X_2 = 0.5X_1 + Z$ where $Z \sim \text{PARETO2}(\mu = 2, \sigma = \sqrt{1.5})$

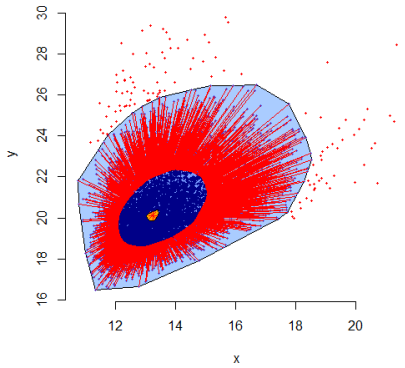


Fig. 9: Reverse Gumbel distribution ($X_1 \sim \text{RG}(\mu = 13, \sigma = 1)$ & $X_2 = 0.5X_1 + Z$ where $Z \sim \text{RG}(\mu = 13, \sigma = \sqrt{1.5})$

Monitoring of short series of dependent observations using a control chart approach and data mining techniques

Olgierd Hryniewicz and Katarzyna Kaczmarek

Abstract Many different control chart have been proposed during the last 30 years for monitoring of 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 paper we consider the case when the information from the available data is not sufficient for good estimation of the model. Therefore, we use the Bayesian concept of model averaging in order to improve model prediction. The novelty of the proposed method consists in the usage of computational intelligence methodology for the construction of alternative models and the calculation of their prior probabilities (weights).

1 Introduction

Control charts were originally devised for the monitoring of 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. Pioneering works in the area of process control in presence of dependent (autocorrelated) data, such as Box, Jenkins & MacGregor (1974), were published in the 1970th. Since that time many papers devoted to this problem have been published, and they can be, in general, divided into two groups. Authors of the first group of papers, such as, e.g.,

Olgierd Hryniewicz
Systems Research Institute, Newelska 6, 01-447 Warszawa, Poland, e-mail: hryniewi@ibspan.waw.pl

Katarzyna Kaczmarek
Systems Research Institute, Newelska 6, 01-447 Warszawa, Poland, e-mail: K.Kaczmarek@ibspan.waw.pl

Vasilopoulos and Stamboulis Vasilopoulos & Stamboulis (1978), Montgomery and Mastrangelo Montgomery & Mastrangelo (1991), Maragah and Woodall Maragah & Woodall (1992), Yashchin Yashchin (1993), Schmid Schmid (1995) or Zhang 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 papers is the paper by Alwan and Roberts Alwan & 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 proposed by many authors, such as, e.g., Wardell et al. Wardell, Moskowitz & Plante (1994), Zhang Zhang (1997), Kramer and Schmid Kramer & Schmid (2000). Both approaches have been compared in many papers, such as, e.g., Lu and Reynolds Lu & 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 paper by Runger 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. Jiang, Tsui & Woodall (2000), the chart proposed by Chin and Apley Chin & Apley (2006) based on second-order linear filters, the chart proposed by Apley and Chin Apley & Chin (2007) based on general linear filters or the PCA-based procedure for the monitoring multidimensional processes proposed by De Ketelaere et al. De Ketelaere, Hubert & Schmitt (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 paper by Kramer and Schmid Kramer & Schmid (2000) or the paper by Apley and Lee Apley & 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 paper.

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 Geweke (2005). What is specific in this approach is the concept of model averaging. The Bayesian model in this approach contains not only the prior knowledge about model parameters, but also a 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 Kaczmarek Hryniewicz & Kaczmarek (2014) 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 paper we try to adopt a similar approach for the construction of Shewhart control charts for residuals.

The paper 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 paper only some representative results have been described in Section 4. The paper 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

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.

2.1 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 represented 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 paper 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 and Jenkins Box, Jenkins & Reinsel (2008) or a popular textbook by Brockwell and Davis Brockwell & 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 of the model p. 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 its average. Then, the parameters $a_1, \ldots, a_p$ or the $AR(p)$ model are calculated by solving the Yule-Walker equations (see, Brockwell & 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 Akaike (1978) defined as

$$BIC = (n-p)\ln[n\hat{\sigma}^2/(n-p)] + n(1+\ln\sqrt{2\pi}) + p\ln[(\sum_{t=1}^{n} x_t^2 - n\hat{\sigma}^2)/p], \quad (6)$$

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 Section 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 Geweke (2005) we define a set $M = \{M_1, M_2, ..., M_J\}$ of multiple competitive 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)

$$p(\omega|y, M) = \sum_{j=1}^{J} p(M_j|y, M)p(\omega|y, M_j) \quad (7)$$

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 paper we will use the concept of model averaging for the construction of a control chart. Different $AR(p)$ models will be used as competitive probabilistic models of a monitored process, and their prior probabilities (weights) will be computed using a methodology described in Section 3.

2.2 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 Alwan & Roberts (1988). Their methodology is applicable for any class of processes, so it is also applicable for the $AR(p)$ process considered in this paper. According to the methodology proposed by Alwan and Roberts Alwan & 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 Subsection 2.1 from a sample on 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 competitive models $M_j, j = 1,\ldots,k$, each described by a vector of parameters $(a_{1,j},\ldots,a^0_{p_j,j})$. In general, any model with known parameters can be used as a competitive one, but in this paper we restrict ourselves to the models chosen according to the algorithm described in Section 3. Let $w_1^{'},\ldots,w_k^{'}$ denote the weights assigned to models $M_1,\ldots,M_k$ by the algorithm described in Section 3. In the construction of our control chart, coined XWAM (X Weighted Average Model chart), to each competitive (alternative) model we will assign a weight $w_j = (1-w_0)w_j^{'}, j = 1,\ldots,k$.

When we model our process using $k+1$ alternative models each process observation generates $k+1$ residuals. In the case considered in this paper they are calculated using the following formula

$$z_{i,j} = x_i - (a_{1,j}x_{i-1} + \ldots + a_{p_j,j}x_{i-p_j}), j = 0,\ldots,k; i = p_j+1,\ldots. \tag{8}$$

Let $i_{min} = \max(p_0,\ldots,p_k)+1$. For the calculation of the parameters of the XWAM control chart we use $n - i_{min} + 1$ weighted residuals calculated from the formula

$$z_i^{\star} - \sum_{j=0}^{k} w_j z_{i,j}, i = i_{min},\ldots,n. \tag{9}$$

The central line of the chart is calculated as the mean of $z_i^{\star}$, and the control limits are equal to 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

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 the 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 Geweke (2005).

3.1 Similarity measures of series of observations

The similiarity of two time series is evaluated by calculating the distance between them. Within the proposed approach, the DTW (*Dynamic Time Warping*) measure as introduced by Berndt and Clifford Berndt (1994) is adapted. DTW is the classical elastic measure and enables to calculate the smallest distance between two series of observations taking into account dilatation in time.

Let $X = \{x_1, x_2, ..., x_N\}$ and $Z = \{z_1, z_2, ..., z_M\}$ denote time series to be compared. The distance d between two points x_i and z_j, the so called local cost function, is defined as follows

$$d(i,j) = f(x_i, z_j) \geq 0 \tag{10}$$

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.

The DTW distance is based on the following recursive relation, which defines a cumulative distance $g(i,j)$ for $i \in \{1, ..., N\}$ and $j \in \{1, ..., M\}$

$$g(i,j) = d(i,j) + min[g(i-1,j), g(i-1,j-1), g(i,j-1)] \tag{11}$$

The cumulative distance 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 grid are called neighboring if $(|i - i*| = 1 \; and \; |j - j*| = 0)$ or $(|i - i*| = 0 \; and \; |j - j*| = 1)$.

When two compared series are of the same length the value of $g(N,M)$ defines the distance between them. However, when $N \neq M$ the situation is more complicated, and the elements of both series have to be aligned in some way. The alignment of the elements from X and Z such that the distance between them is minimized is called a *warping path*. The DTW problem is defined as a minimization of cumulative distances over potential warping paths based on the cumulative distance for each

path. This problem is solved using dynamic programming, and its complexity is $O(NM)$, and its solution is considered as the distance between two compared time series.

In Figure **??**, the performance of the Euclidean and DTW distances is compared for exemplary series of observations from AR(-0,9), AR(-0,5) and AR(0,0) processes.

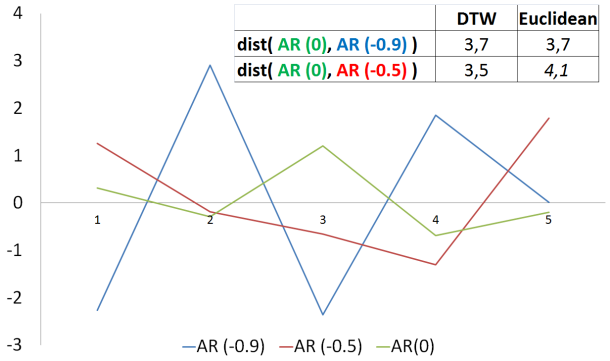


Fig. 1: Euclidean and DTW distances for exemplary series of observations

As observed, 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. On the other hand, the Euclidean distance between series from AR(0,0) and AR(-0,9) is also 3,7, but the distance between series from AR(0,0) and AR(-0,5) results 4,1, which is contradictory to intuition.

Our experiments confirm the good properties of the DTW measure, especially for time series with identified dilatation in time, because DTW seems to preserve trends. For further reading, we refer to, e.g., the recent survey and experimental comparison of representation methods and distance measures for time series data provided by Wang et al. Wang et al. (2013). Wang et al. conclude that especially on small data sets, elastic measures like DTW can be significantly more accurate than Euclidean distance and other lock-step measures.

3.2 Construction of prior probabilities (weights)

Having defined the distance between 2 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 $\{M_1, ..., M_k\}$ to be considered in predictions of the monitored process and their respective weights $\{w_1, ..., w_k\}$ such that $\sum_{h=1}^{k} w_h = 1$.

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**:
$M_1, ..., M_k$ - alternative models to be considered for the monitored process,
$w_1, ..., w_k$ - 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
        $M_1, ..., M_k \leftarrow$ selectAlternativeModels($dist_m$, k)
    $w_1, ..., w_k \leftarrow$ scaleWeights(M, k)
    return $M_1, ..., M_k, w_1, ..., w_k$

Algorithm 1 depicts a high-level description of the proposed approach. It consists of the following steps:

Step 1. Generation of template database $Y_{J,s}$.

The template database consists of models $\{M_1, ..., M_J\}$ 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

$$dist_{m,i} = DTW(y_{m,i}, y) \tag{12}$$

Step 3. Aggregating similarities to establish weights corresponding to models $\{M_1, ..., M_k\}$.
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

$$dist_m = \frac{\sum_{i=1}^{s} dist_{m,i}}{s} \tag{13}$$

Having evaluated the average distance for each of the template models $\{M_1, ..., M_J\}$, the k models with smallest distance are selected. Then, the prior weights $\{w_1, ..., w_k\}$ are calculated

$$w_i = \frac{dist_i}{\sum_{h=1}^{k} dist_h} \tag{14}$$

4 Numerical experiments

In this paper we consider a Shewhart control chart whose parameters are designed using information from relatively small samples. The effect of parameter estimation on the properties of the classical Shewhart control chart has been investigated by many authors. They found that estimated control limits, in general, are too wide. Thus, the values of ARL are larger than expected, and special corrections are needed, such as, e.g., proposed by Albers and Kallenberg Albers & Kallenberg (2004). The same effect has been observed in the case of positively autocorrelated data. However, for negatively autocorrelated data the control limits are too narrow, and the rate of false alarms is too high. 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 Kramer & 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 10000 charts was designed, and for each of them 500 process runs were simulated. W 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 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 Section 2. In order to compare both charts we have computed four characteristics: average ARL (AvgARL), median ARL (MedARL), average MRL (AvgMRL), and median MRL (MedMRL), where MRL

is the median of observed run lengths. The results of the experiment are presented in Table 1.

Table 1: Properties of control charts with independent observations

	X-chart				X-chart (residuals)			
n	AvgARL	MedARL	AvgMRL	MedMRL	AvgARL	MedARL	AvgMRL	MedMRL
20	1811,2	306,1	1366,5	213,5	398,9	67,8	300,5	48,0
30	1067,2	332,2	753,7	231,0	363,3	114,5	254,4	80,0
40	740,2	335,4	515,7	233,75	295,1	137,2	205,6	96
50	642,1	349,5	446,8	242,0	302,9	162,7	210,9	114,0
100	472,8	358,1	328,5	249,5	329,0	245,9	229,3	171,0
500	388,3	366,7	269,8	255,0	360,1	340,1	250,9	237,25
1000	379,7	368,7	263,5	256,0	366,6	357,0	255,3	248,5
2000	374,2	369,1	259,8	256,0	371,4	366,5	258,7	255

The results of simulations presented in Table 1 confirm many of well known facts. First, consider the case of the X chart for direct, and independent, 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 5–9) the situation is different. In this case the uncertainty related to imprecisely calculated control limits (positive bias) in combined with the uncertainty related to the computation of residuals (negative, as it was proved in Kramer & Schmid (2000)). The total 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.

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 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 paper 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.

The properties of the XWAM chart 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. 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 has been generated from an autoregressive process of the second order, AR(2), with the parameters 0,7 and -0,9.

$$\begin{array}{cccccccccc} -0,94 & -1,35 & -1,4 & -0,34 & 0,71 & 0,14 & -0,49 & -1,08 & 1,41 & 2,25 \\ -0,05 & -2,74 & -1,02 & 2,95 & 4,03 & 0,43 & -2,67 & -2,45 & 1,98 & 4,25 \end{array}$$

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 Section 3 we have found 5 best alternative models. All of them are 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, 500 runs of the process. The values of its characteristics, ARL and MRL, are presented in Table 2 for different values of the weight w assigned to the estimated model.

Table 2: Chart in-control characteristics for different weights assigned to the estimated model

w	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0
ARL	21,064	28,208	38,23	58,062	94,53	178,13	341,562	776,1	2110	5496	16894
MRL	15,0	19,5	28,0	42,5	68,0	128,0	270,0	546,5	1402	4045	10961

The results presented in Table 2 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. 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 0,4. For bigger weights the number of false alarms is to high, and on the other hand, when this weight is too small the control limits are too wide, and alarms indicating deterioration of a process would be triggered too late.

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 must 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 Section 3 we have found 5 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. However, 4 out of 5 alternative models look more similar to the original (the fifth is visibly different).

In Table 3 we present the results of a similar simulation experiment as in the case of the first considered sample. The values of ARL and MRL are, in this case, very far from expected when only the estimated model is taken into account. The best, i.e., the closest to the expected ones, values are obtained when we completely neglect ($w = 0$) the model estimated from the sample.

Table 3: Chart in-control characteristics for different weights assigned to the estimated model - extreme sample

w	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0
ARL	26055	23463	21180	19640	17608	16224	14585	13210	12431	11590	10352
MRL	17688	15960	14540	13545	12230	11540	10256	9492	9133	8067	7083

The results presented in Tables 2–3 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 4–5 we present the average and median values of ARL's evaluated from 1000 generated control charts, and 500 process runs generated for each control chart, i.e., from all together 500000 simulated process runs. In all these runs the simulated process was in in-control state.

Table 4: Average in-control ARL for different weights assigned to the estimated model, n=20

Model/w:	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)	81,1	87,3	96,5	110,2	131,7	167,8	235,1	373,6	708,3	1498	2995
AR(-0,5)	327,8	327,9	332,3	341,7	356,1	377,7	407,5	450,9	509,6	587,0	693,3
AR(0)	368,9	397,3	431,4	476,3	531,2	761,9	867,4	1007,9	1198	1422	1683
AR(0,5)	215,7	255,8	309,2	388,3	509,8	707	1035	1536	2286	3315	4545
AR(0,9)	87,5	147,9	656,9	1839	4108	6809	9812	13132	16619	20279	23821
AR(0,7 , -0,9)	77,5	91,3	154,2	712,4	1975	4408	8654	14067	20742	28182	35668

The results presented in Tables 4 – 5 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 4) are significantly different from the respective values of medians of averages (Table 5). This difference tells 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

Table 5: Median in-control ARL for different weights assigned to the estimated model, n=20

Model/w:	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)	38,5	42,0	47,0	52,3	58,2	66,0	75,7	89,0	104,5	123,0	144,0
AR(-0,5)	52,9	58,2	64,0	71,1	80,8	94,3	108,7	126,1	149,7	184,2	220,2
AR(0)	70,7	76,4	82,1	89,9	98,4	117	129	147,1	171,4	191,7	217
AR(0,5)	50,1	56,8	63,1	71,9	84,6	99,4	116,2	143,3	164,2	206,4	248,9
AR(0,9)	30,0	33,5	38,5	44,4	52,6	72,2	98,4	145,7	198,3	298,3	417,3
AR(0,7 , -0,9)	25,8	31,3	41,2	59,8	92,8	164,1	285,8	570,6	1172	2698	7056

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 this paper have, as for now, rather qualitative character.

The results presented in Tables 4–5 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's the ability of a chart to detect shifts of a monitored process. In this paper we consider only the shifts of the average value, measured in units of standard deviation. In Table 6 we show the values of ARL for different shifts when we use a sample of 20 elements, and the observations positively, but not very strongly, correlated ($\rho = 0,5$).

Table 6: Average ARL for different weights assigned to the estimated model and different shifts of the process level, $\rho = 0,5$, n=20

Shift/w:	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0
-3	10,4	9,7	9,1	8,6	8,2	7,9	7,7	7,6	7,5	7,5	7,6
-2	26,5	25,9	25,6	25,3	25,3	25,6	26,2	27,3	28,9	31,2	34,5
-1	95,6	100,4	106,5	114,0	124,7	138,9	158,6	186,6	225,3	285,0	370,1
0	215,7	241,7	272,4	310,2	357,1	418,3	498,7	611,6	779,0	1053,1	1491,6
1	92,3	98,2	105,1	114,8	127,3	144,7	169,6	205,4	265,3	355,8	495,6
2	25,8	25,1	24,8	24,8	25,0	25,5	26,4	27,8	30,2	33,5	39,0
3	10,1	9,4	8,8	8,3	7,9	7,7	7,5	4,4	7,3	7,4	7,7

From Table 6 it can be seen quite clearly that 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 7. However, too large values of the ARL for shifts of small and medium sizes are hardly acceptable. Therefore, we can set parameter w to such value that the average time to a false alarm is not smaller than a given value (e.g., equal to 370). In the considered case, such an "optimal" value of w is equal to 0,6. 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 also automatically 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.

Table 7: Discriminative power of the XWAM chart for different shifts of the process level, $\rho = 0,5$, n=20

Shift/w:	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0
-3	0,05	0,04	0,03	0,03	0,02	0,02	0,02	0,01	0,01	0,01	0,01
-2	0,12	0,11	0,09	0,08	0,07	0,06	0,05	0,04	0,04	0,03	0,02
-1	0,44	0,42	0,39	0,37	0,29	0,33	0,32	0,31	0,29	0,27	0,25
0	1	1	1	1	1	1	1	1	1	1	1
1	0,43	0,41	0,39	0,37	0,36	0,35	0,34	0,34	0,34	0,34	0,33
2	0,12	0,10	0,09	0,08	0,07	0,06	0,05	0,05	0,04	0,03	0,03
3	0,05	0,04	0,03	0,03	0,02	0,02	0,01	0,01	0,01	0,01	0,01

Interesting case is presented in Tables 8–9. In these Tables we consider the case of negative dependence of medium strength ($\rho = -0,5$). The "optimal" value of w is the same as in the case of the positive dependence of similar strength. However, the discriminative power is in this case much higher, and - what is somewhat surprising - does not depend upon the value of w. Thus, by changing this value we act as if we only change the y-axis scale of a chart. What is more interesting in this case, however, is better discrimination of this chart in comparison to a classical Shewhart X-chart for individual independent observations.

Table 8: Average ARL for different weights assigned to the estimated model and different shifts of the process level, $\rho = -0,5$, n=20

Shift/w:	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0
-3	3,2	3,2	3,2	3,2	3,2	3,2	3,2	3,2	3,3	3,3	3,3
-2	4,5	4,5	4,5	4,6	4,7	4,8	4,9	5,1	5,4	5,7	6,2
-1	21,0	22,4	21,9	22,7	23,7	25,1	27,0	29,4	32,5	37,0	43,2
0	327,8	327,9	332,3	341,7	356,1	377,7	407,5	450,9	509,6	587,0	693,3
1	21,1	22,7	21,9	22,7	23,7	25,1	27,0	29,4	32,5	37,0	43,2
2	4,6	4,6	4,6	4,7	4,7	4,8	5,0	5,2	5,4	5,8	6,2
3	3,2	3,2	3,2	3,2	3,2	3,2	3,2	3,2	3,3	3,3	3,4

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 10 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 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 10 confirm already known results that classical X charts for residuals perform very badly. The

Table 9: Discriminative power of the XWAM chart for different shifts of the process level, $\rho = -0,5$, n=20

Shift/w:	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0
-3	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,005	0,005	0,005	0,005
-2	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
-1	0,06	0,07	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06
0	1	1	1	1	1	1	1	1	1	1	1
1	0,06	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,06	0,06	0,06
2	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
3	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,005	0,005	0,005	0,005

Table 10: Values of ARL for different sample sizes, $\rho = 0,9$

	w=1.0		w=0,8		w=0,7		w=0,6		w=0,5		w=0,4	
Shift/n	20	50	20	50	20	50	20	50	20	50	20	50
-3.0	64,5	36,9	66,4	36,9	**106,3**	38,7	263,7	44,0	453,4	61,0	570,6	**144,1**
-2.0	67,7	45,8	85,7	50,3	**197,4**	55,1	414,0	66,8	634,1	107,3	932,7	**280,2**
-1.0	80,9	54,2	126,0	66,7	**285,1**	75,1	525,4	97,2	961,5	172,3	1662,8	**430,9**
0,0	87,5	58,5	146,2	73,1	**376,5**	88,0	695,4	121,7	1386,7	232,9	2630,9	**422,2**
1,0	78,0	55,4	135,1	66,8	**395,4**	78,1	718,9	100,9	1489,9	165,9	2576,1	**330,0**
2,0	59,6	46,6	102,7	50,9	**307,3**	55,8	597,6	64,7	1027,9	84,6	1804,9	**146,0**
3,0	39,9	37,0	58,7	36,3	**176,6**	37,1	445,4	39,2	706,2	44,2	1095,9	**57,7**

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 11 where respective coefficients of discriminative power are displayed.

Table 11: Discriminative power of the XWAM chart for different sample sizes, $\rho = 0,9$

	w=1.0		w=0,8		w=0,7		w=0,6		w=0,5		w=0,4	
Shift/n	20	50	20	50	20	50	20	50	20	50	20	50
-3,0	0,74	0,63	0,45	0,51	**0,28**	0,44	0,38	0,36	0,33	0,26	0,22	**0,33**
-2,0	0,77	0,78	0,59	0,69	**0,52**	0,63	0,60	0,55	0,46	0,46	0,35	**0,66**
-1,0	0,92	0,93	0,86	0,91	**0,76**	0,85	0,76	0,80	0,69	0,74	0,63	**1,02**
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	0,89	0,95	0,92	0,91	**1,05**	0,89	1,03	0,83	1,07	0,71	0,98	**0,78**
2,0	0,68	0,80	0,70	0,70	**0,82**	0,63	0,86	0,53	0,74	0,36	0,69	**0,35**
3,0	0,46	0,63	0,40	0,50	**0,47**	0,42	0,64	0,32	0,51	0,20	0,42	**0,14**

The performance of respective XWAM charts in this very unfavorable case is not much better. If we look at the values of their ARL's, and their respective coefficients of discriminative power (displayed in bold for the "optimal" choice of w) we can see that the discrimination power, in general, has been improved. However, the increase of sample size has resulted in visibly better performance only in the case of positive shifts of the process level. When such shifts are negative XWAM charts with smaller

sample sizes perform, quite unexpectedly, better. The explanation of this rather strange phenomenon needs further investigations. It has to be noted, however, that in more favorable cases (not presented in this paper) positive effects of the increase of sample size is more visible.

5 Conclusions

In the paper 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 Data 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 these 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 paper, 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.

References

Akaike, H.: Time Series Analysis and Control Through Parametric Model. In: Findley, D.F. (ed.) Applied Time Series Analysis, Academic Press, New York (1978)

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

Alwan, L.C., Roberts, H.V.: Time-Series Modeling for Statistical Process Control, Journal of Business & Economic Statistics **6**, 87–95 (1988)

Apley, D.W., Chin, C.: An Optimal Filter Design Approach to Statistical Process Control. Journ. of Qual. Techn. **39**, 93–117 (2007)

Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. AAAI-94 Workshop on Knowledge Discovery in Databases, 359–370 (1994)

Apley, D.W., Lee, H.C.: Robustness Comparison of Exponentially Weighted Moving-Average Charts on Autocorrelated Data and on Residuals. Journ. of Qual. Techn. **40**, 428–447 (2008)

Box, G.E.P., Jenkins, G.M., MacGregor, J.F.: Some Recent Advances in Forecasting and Control, Part II. Journ. of the Roy. Stat. Soc., Ser. C, **23**, 158–179 (1974)

Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis. Forecasting and Control (4th Ed.). J.Wiley, Hoboken NJ (2008)

Brockwell, P.J., Davis, R.A.: Introduction to Time Series and Forecasting (2nd Ed.). Springer, New York (2002)

Chin, C-H., Apley, D.W.: Optimal Design of Second-Order Linear Filters for Control Charting. Technometrics, **48**, 337–348 (2006)

De Ketelaere, B., Hubert, M., Schmitt, E.: Overview of PCA-Based Statistical Process-Monitoring Methods for Time-Dependent, High-Dimensional Data. Journ. of Qual. Techn. **47**, 318–335 (2015)

Geweke, J.: Contemporary Bayesian econometrics and statistics. J. Wiley, Hoboken NJ (2005)

Hryniewicz, O., Kaczmarek, K.: Bayesian analysis of time series using granular computing approach (in press). Appl. Soft Comput. J. (2014) doi: 10.1016/j.asoc.2014.11.024

Jiang, W., Tsui, K., Woodall, W.H.: A New SPC Monitoring Method: The ARMA Chart. Technometrics,**42**, 399–410 (2000)

Köksal, G., Kantar, B., Ula, T.A., Testik, M.C.: The effect of Phase I sample size on the run length performance of control charts for autocorrelated data. Journ. of Appl. Stat., **35**, 67–87 (2008)

Kramer, H., Schmid, W.: The influence of parameter estimation on the ARL of Shewhart type charts for time series. Statistical Papers,**41**, 173–196 (2000)

Lu, C.W., Reynolds, M.R.Jr.: Control Charts for Monitoring the Mean and Variance of Autocorrelated Processes. Journ. of Qual. Techn., **31**, 259–274 (1999)

Maragah, H.D., Woodall, W.H.: The effect of autocorrelation on the retrospective X-chart. Journ. Stat. Simul. and Comput., **40**, 29–42 (1992)

Montgomery, D.C., Mastrangelo, C.M.: Some statistical process control methods for autocorrelated data (with discussion). Journ. of Qual. Techn., **23**, 179–204 (1991)

Schmid, W.: On the run length of Shewhart chart for correlated data, Stat. Papers, **36**, 111–130 (1995)

Runger, G.C.: Assignable Causes and Autocorrelation: Control Charts for Observations or Residuals? Journ. of Qual. Techn., **34**, 165–170 (2002)

Vasilopoulos, A.V., Stamboulis, A.P.: Modification of Control Chart Limits in the Presence of data correlation. Journ. of Qual. Techn., **10**, 20–30 (1978)

Wang, Xi, Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E.: Experimental comparison of representation methods and distance measures for time series data. Data Min Knowl Disc, **26**:275–309 (2013)

Wardell, D.G., Moskowitz, H., Plante, R.D.: Run-Length Distributions of Special-Cause Control Charts for Correlated Processes (with disscussion). Technometrics, **36**, 3–27 (1994)

Yashchin, E.: Performance of CUSUM Control Schemes for Serially Correlated Observations. Technometrics, **35**, 37–52 (1993)

Zhang, N.F.: Detection Capability of Residual Chart for Autocorrelated Data. Journ. of Appl. Stat., **24**, 475–492 (1997)

Zhang, N.F.: A Statistical Control Chart for Stationary Process Data. Technometrics, **40**, 24–38 (1998)

A Primer on SPC and Web Data

Erwin Saniga, Darwin Davis, James Lucas

Abstract In this paper 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.

1 Introduction

In this paper 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.

Erwin Saniga
Department of Business Administration, Alfred Lerner College of Business & Economics, University of Delaware, Newark, DE 19716, USA, e-mail: saniga@udel.edu

Darwin Davis
Department of Business Administration, Alfred Lerner College of Business & Economics, University of Delaware, Newark, DE 19716, USA, e-mail: dd@udel.edu

James Lucas
James Lucas and Associates, e-mail: James.Lucas@verizon.net

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 four 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 seven 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, w3 counter, statcounter and WP SLimstat Analytics

We implemented three software analytics packages on the large commercial site (Google Analytics, Clicky and Statcounter) over a three 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 2 of the 3 outcomes outside the three sigma control 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.

A second reason for disparity might be that there are IP's (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 S. Chimphlee, Salim, Ngadiman and W. Chimphlee (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 five 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.

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, Davis and Lucas (2009) illustrate the use of these plots in an actual example. 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, Salim, Ngadiman, and Chimphlee (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 with visit a web page within some period of time.

They do not, however, provide any tools for tracking changes in web site performance.

Zhu (2002) 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 ith order chain versus the alternative that it is a jth 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 users flow through the system.

Many different types of measures of user visits to a website are presented in the various analytics packages we tested in this paper. 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 by session duration.

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 paper.

References

Agresti, A. (2013), *Categorical Data Analysis*, Wiley, NY.

Anderson, T. W., and 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. and Chimphlee, W. (2006), in Sabh, T. and Elleithy, K. (eds.), *Advances in Systems, Computing Sciences and Software Engineering*, 371-376.

Clicky, https://clicky.com/

Google Analytics, www.google.com/analytics/

Hawkins, D. and Olwell, D. (1998), *Cumulative Sum Charts and Charting for Quality Improvement*, Springer-Verlag, NY.

Huang, Q., Yang, Q., Huang, J. Z. and Ng, M. K. (2004), Mining of web-page visiting patterns with continuous-time markov models, In *Advances in Knowledge Discovery and Data Mining*, 549-558, Springer Berlin, Heidelberg.

Li, Z., and Tian, J. (2003), Testing the suitability of Markov chains as Web usage models, In *Computer Software and Applications Conference, 2003. COMPSAC 2003. Proceedings. 27th Annual International*, 356-361. IEEE.

Lucas, J. (1985), Counted Data Cusums, *Technometrics*, **27**(2), 129-144.

Marques, A. and 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*, 5^{th} ed., Wiley, NY.

Saniga,E., Davis, D. and 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/

Statcounter, https://statcounter.com/

Topiladou, E. and Psarakis, S. (2009), Review of Multiattribute and Multinomial Control Charts, *Quality and Reliability Engineering International*, **25**, 773-809.

Zhu, J., Hong, J. and Hughes, J. G. (2002a), Using Markov Chains for Link Prediction in Adaptive Web Sites, In *Soft-Ware 2002: Computing in an Imperfect World*, 60-73. Springer Berlin, Heidelberg.

Zhu, J., Hong, J. and Hughes, J. G. (2002), Using Markov Models for Web Site Link Prediction, In *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*, 169-170. ACM.

W3counter, https://www.w3counter.com/

Wpslimstat, http://www.wp-slimstat.com/

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

Table 1: Visits to a Small Website as Recorded by Various Analytics Packages

		Google Analytics	w3counter	Statscounter	WPSlimstat Analytics
Mar 13 - 20	*New Users*	9	11	11	Installed on 21st March
	Returning Users	7	0	3	Installed on 21st March
Mar 21 - 27	*New Users*	26	28	18	15
	Returning Users	4	0	5	1
Mar 28 - Apr 3	*New Users*	17	16	13	11
	Returning Users	3	0	3	2
Apr 4 - 10	*New Users*	40	29	30	5
	Returning Users	0	7	9	39
Apr 11 - 17	*New Users*	36	23	16	6
	Returning Users	3	3	3	28
Apr 18 - 24	*New Users*	12	21	13	20
	Returning Users	6	3	6	3
Apr 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

Table 3.1a: Raw Data for April 25 - May 1, 2016

Current State \ Future 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**

Table 3.1b: Transition Probability Matrix for April 25 - May 1, 2016

Current State \ Future 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**

Table 3.2: Transition Probability Matrix for May 2 - May 8, 2016

Current State \ Future 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 yellow rows and columns.

Table 3.3: Transition Probability Matrix for May 9 - May 15, 2016

Current State \ Future 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**

= There was no activity on these pages so the row probabilites cannot be estimated.
A transition probability matrix can be created by removing the yellow rows and columns.

Table 3.4: Transition Probability Matrix for May 16 - May 22, 2016

Current State \ Future 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**

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	Topalidou 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	Topalidou and Psarakis (2009)
Visit length	Multinomial	Multinomial	Topalidou and Psarakis (2009)
Visit length	Mean	CUSUM for a Mean	Hawkins and Olwell (1998)
Browsers	Multinomial	Multinomial	Topalidou and Psarakis (2009)

On the Phase I Shewhart Control Chart Limits for Minimizing Mean Squared Error When the Data are Contaminated

Murat Caner Testik, Christian H. Weiß, Yesim Koca, and Ozlem Muge Testik

Abstract A Shewhart-type control statistic lacks memory of previous observations on a monitored quality characteristic. Therefore, Shewhart-type control charts are known to be relatively insensitive to moderate-to-small sized shifts in the process parameters. Yet, these charts are recommended in the literature for the design stage of a control chart, namely Phase I, where unknown process parameters are estimated and the control limits are set for online process monitoring in Phase II. There are several studies that considered time-weighted control charts for the Phase I application, as well as alternative estimators, to improve the design for monitoring in Phase II. Some studies also considered the design of Shewhart control charts based on false alarm rates or overall false alarm probabilities, where it is concluded that the control limits may be widened. In this study, we simulate assignable causes of variation by contaminating sets of Phase I observations and investigate the use of traditional Shewhart control chart for parameter estimation when the observations are normally distributed. By varying the distance of the control limits from the center-line, Phase I Shewhart control charts are evaluated in terms of the true and false alarm percentages, number of iterations performed to finalize the Phase I analysis, and the mean squared error of the parameter estimates. Some practical recommendations are provided.

Key words: Statistical process control; Phase I; control limit widths; estimation error; contaminated data; mean squared error.

Murat Caner Testik, Yeşim Koca, Özlem Müge Testik
Hacettepe University, Department of Industrial Engineering, 06800 Beytepe-Ankara, Turkey,
e-mail: mtestik@hacettepe.edu.tr
e-mail: yesimkoca@hacettepe.edu.tr
e-mail: ozlemaydin@hacettepe.edu.tr

Christian H. Weiß
Helmut Schmidt University, Department of Mathematics and Statistics, 22008 Hamburg, Germany,
e-mail: weissc@hsu-hh.de

1 Introduction

The first stage of a control chart implementation is the design of a control chart for online process monitoring. While many alternative charts can be used for online process monitoring, called Phase II implementation of control charts, often Shewhart control charts are suggested in Phase I implementations, where process parameters are estimated to design a Phase II control chart.

The estimation process is a retrospective study where observations are investigated all at once for a characterization of the process stability. Using an initial Phase I set of subgroups of observations, parameter estimates are obtained and trial control limits of the Shewhart control chart are determined. Out-of-control states of the process are identified by using the signals of the chart, and observations corresponding to the signals are eliminated. Parameter estimates are then revised by using the remaining observations. Calculating the revised Shewhart control limits, further out-of-control states of the process are identified (by using the signals of the revised control chart) and removed. This process is iterated until all of the observations are within the control limits. The final estimates are then used in designing the control chart for the Phase II implementation.

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 & Testik (2011), Testik (2007), Testik et al. (2006), Jones et al. (2001)). 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. On the other hand, some studies considered robust estimators in Phase I (see, for example, Schoonhoven & Does (2012), Zwetsloot et al. (2014)), while others proposed adjustments to Phase II control limits (see, for example, Albers & Kallenberg (2004, 2005)). However, research on methods to obtain an in-control reference set of observations has received less emphasis (Jones-Farmer et al., 2014), which is the topic considered in the following.

Simulations of the Shewhart control chart in Phase I implementations were performed. To study the effects of the presence of assignable causes of variation in the process, the initial Phase I data set for estimation was contaminated, such that the means of a given percent of the subgroups were shifted. Iterations of Phase I implementations were simulated for obtaining statistically in-control reference sets of observations for parameter estimation. As a control factor, the control limits' width of Shewhart control charts was altered. Performance metrics related to Phase I implementations as well as parameter estimates are provided in the following.

The organization of the paper is as follows. The methodology, together with the assumptions of the simulations, are described in Section 2. Simulation results and interpretations are provided in Section 3, and the study is concluded in Section 4.

2 Methodology and Assumptions for Phase I Simulations

In this study, the Phase I implementation steps were simulated by using the two-sided $\overline{x}$ and s control charts. Suppose that m subgroups of size n observations on a characteristic x are used at an iteration of a Phase I application. Let $\overline{x}$ be the subgroup average,

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

and s be the subgroup standard deviation,

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

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

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

where L is the distance of a control limit from the center line in terms of standard deviation units (this is often selected to be 3 in conventional use), $\overline{\overline{x}}$ is the average of m subgroup averages,

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

$\overline{s}$ 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 $\sigma\sqrt{1-c_4^2}$; UCL, CL, and LCL for 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.

In the following, it is assumed that the in-control process observations can be modeled as independent and identically distributed standard normal random variables with mean $\mu_0 = 0$ and standard deviation $\sigma_0 = 1$.

As a strategy of sampling, we consider rational subgroups and assume that consecutive samples are taken to minimize the chance of variability due to assignable causes of variation within a subgroup, and to maximize the chance of variability between subgroups when assignable causes of variation are present. Consequently, in the simulations, process changes due to the presence of assignable causes were represented as shifts in the mean between two successive subgroups, which affect all the observations within the subgroup after the change. Note that this is the snapshot approach as discussed in Montgomery (2009). A second approach, called the random sample approach, that assumes a random sample of all process output over a sampling interval will be considered in another research, where a change in the process will be allowed to affect only some of the observations in a subgroup.

In the simulations, an out-of-control process was modeled by a change only in the mean from μ_0 to a new level μ_1, while the standard deviation was kept constant. The values tested for the out-of-control process subgroup means were 1,2, and 3. These out-of-control states of the process were modeled as sustained shifts over time. Note that, although it is common to assume large shifts of the mean for the out-of-control process states in Phase I, here we also consider the effect of smaller mean shifts.

Considering m subgroups each having a size of n observations and a total of $N = mn$ observations in the Phase I data set, a percent c of the total number of observations were contaminated, such that the shifted subgroup means in the initial set of Phase I subgroups was 0, 4, or 8 %. As a controllable factor, the distance L of a control limit from the center line is varied from 1 to 5 with increments 0.1.

3 Performance Metrics and Simulation Results

Several metrics can be considered for the performance of a control chart in Phase I implementations. Here, the following metrics were computed using the simulation results for 100,000 replications:

- *Average Number of Iterations (ANI)*, which was obtained by counting the number of iterations performed to reach the final parameter estimates in each replication and by taking the average of these.
- *True Alarm Percentage (TAP)*, which was calculated as the ratio of the number of true signals to the number of contaminated samples in each replication, averaged over all replications.
- *False Alarm Percentage (FAP)*, which was calculated as the ratio of the number of false signals to the number of clean samples in each replication, averaged over all replications.

- *Mean squared error (MSE)*, which was calculated as the square of the difference of the final parameter estimate from its true value in a replication, averaged over all replications.

Due to the similarity of the interpretations and to save some space, we provide only some representative simulation results for ANI, TAP, FAP, and MSE for various cases in the sequel.

We start with a discussion of Tables 1 to 6, which refer to the case of the number of subgroups being $m = 50$. The ANI has a peak between the L values 1 to 2 for the no shift case (Table 1) and all of the shifted mean cases (Tables 2–6). This is because of the tight control limits, exceeded by many points. When the shift values μ_1 are 0 or 1, ANI approaches 1 with L in the interval 3 to 5, so that the estimation is generally completed in one iteration, on average. Considering the shift value $\mu_1 = 2$ but with $n = 10$ (Table 6), the minimum of ANI is achieved in the interval 3 to 4 for L, where the minimum is approximately 2.

Now consider the simulation results for the TAP measuring the power of detecting the shift, first with the shift value $\mu_1 = 1$ (Table 2). While more than 60 % of the out-of-control subgroups are detected in a Phase I study with an L value in the interval 1 to 2, the true alarm percentage is around 21 % for $L = 3$. This signifies that with the conventional 3σ limits, small shifts are mostly undetected. As L approaches 5, the TAP approaches 0 indicating that the charts lose their capability to detect the shift. On the other hand, for the shift value $\mu_1 = 2$, more than 99 % of the out-of-control subgroups are detected with the L values in the interval 1 to 2 when $n = 5$ (Table 3). With the conventional 3σ limits, this is greater than 91 % for both of the contamination levels tested (Tables 3, 5). Yet, when $n = 10$ (Table 6), one can detect almost all of the out-of-control subgroups. There is a clear performance advantage in true signals when $n = 10$ compared to $n = 5$. As the shift size increases, say 3 (Table 4), the detection performance increases significantly. In the worst case with $L = 5$, TAP is greater than 93 %.

The FAP of the charts (size) under the different cases considered are all similar. The FAP decreases monotonically as L gets larger. At $L = 1$, FAP is between 55–60 %, and with $L \geq 3$, TAP is less than 1 %, indicating that false signals are rare.

The sensitivity of the MSE of the mean estimates to L can be observed from the tables. It exhibits some kind of U-shaped pattern: we have 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). Larger shifts in the mean result in smaller MSE values for a given value of L. This is because of the increase of the detection performance of the chart with larger mean shifts. 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.

The MSEs of the standard deviation estimates increase as L decreases. This is because less subgroups are used in the estimation. Since rational subgroups are considered and the mean shifts are assumed to be between subgroups, standard deviation estimates are not affected severely from shift sizes, for given L, n, and c.

Table 1: ANI, TAP, FAP and MSE metrics with $m = 50$, $n = 5$, and shift = 0 ($c = 0$) for Phase I data sets.

m	n	μ_1	c	L	FAP	TAP	ANI	MSE_μ	MSE_σ
50	5	0	0	5.0	0.002		1.001	4.023	2.640
				4.9	0.002		1.001	4.032	2.644
				4.8	0.002		1.001	4.000	2.656
				4.7	0.003		1.002	4.014	2.652
				4.6	0.005		1.002	4.010	2.637
				4.5	0.007		1.003	3.997	2.656
				4.4	0.009		1.004	3.983	2.637
				4.3	0.014		1.007	4.015	2.659
				4.2	0.017		1.009	4.001	2.634
				4.1	0.024		1.012	4.020	2.670
				4.0	0.033		1.017	4.018	2.666
				3.9	0.046		1.023	4.020	2.667
				3.8	0.063		1.031	4.053	2.657
				3.7	0.087		1.043	4.013	2.704
				3.6	0.113		1.055	4.037	2.686
				3.5	0.156		1.075	4.019	2.729
				3.4	0.210		1.100	4.064	2.753
				3.3	0.285		1.133	4.052	2.770
				3.2	0.378		1.172	4.069	2.776
				3.1	0.515		1.228	4.106	2.816
				3.0	0.684		1.292	4.177	2.889
				2.9	0.908		1.373	4.180	2.950
				2.8	1.196		1.467	4.262	3.020
				2.7	1.579		1.580	4.320	3.134
				2.6	2.077		1.708	4.468	3.251
				2.5	2.708		1.842	4.539	3.441
				2.4	3.542		1.995	4.686	3.648
				2.3	4.638		2.154	4.821	3.922
				2.2	6.013		2.309	5.067	4.238
				2.1	7.760		2.460	5.319	4.621
				2.0	9.908		2.594	5.622	5.071
				1.9	12.556		2.728	5.999	5.585
				1.8	15.687		2.847	6.381	6.013
				1.7	19.279		2.955	6.785	6.486
				1.6	23.397		3.052	7.331	6.925
				1.5	28.045		3.130	7.818	7.219
				1.4	33.058		3.176	8.266	7.319
				1.3	38.662		3.222	8.628	7.610
				1.2	44.484		3.226	9.046	7.495
				1.1	50.593		3.206	9.209	7.376
				1.0	56.888		3.143	9.431	7.072

MSE_μ/MSE_σ = MSE of mean/standard deviation estimates $\times 1,000$.

Table 2: ANI, TAP, FAP and MSE metrics with $m = 50$, $n = 5$, $c = 4\,\%$ and shift = 1 for Phase I data sets.

m	n	μ_1	c	L	FAP	TAP	ANI	MSE_μ	MSE_σ
50	5	1	4	5.0	0.002	0.290	1.007	5.618	2.643
				4.9	0.002	0.344	1.008	5.573	2.629
				4.8	0.002	0.484	1.011	5.562	2.655
				4.7	0.004	0.614	1.014	5.614	2.654
				4.6	0.005	0.802	1.018	5.569	2.661
				4.5	0.007	1.014	1.023	5.566	2.647
				4.4	0.010	1.312	1.031	5.577	2.640
				4.3	0.014	1.746	1.041	5.552	2.655
				4.2	0.018	2.136	1.051	5.550	2.662
				4.1	0.025	2.730	1.065	5.520	2.660
				4.0	0.035	3.374	1.081	5.453	2.666
				3.9	0.048	4.237	1.103	5.525	2.683
				3.8	0.064	5.232	1.128	5.452	2.674
				3.7	0.087	6.350	1.158	5.433	2.692
				3.6	0.119	7.774	1.196	5.416	2.713
				3.5	0.164	9.138	1.236	5.393	2.744
				3.4	0.218	11.160	1.288	5.369	2.751
				3.3	0.291	13.103	1.343	5.316	2.777
				3.2	0.395	15.514	1.411	5.315	2.815
				3.1	0.531	18.133	1.486	5.275	2.860
				3.0	0.705	20.908	1.567	5.318	2.914
				2.9	0.944	24.144	1.660	5.280	2.992
				2.8	1.243	27.377	1.754	5.315	3.056
				2.7	1.647	30.968	1.857	5.331	3.183
				2.6	2.140	35.202	1.964	5.360	3.275
				2.5	2.794	39.251	2.073	5.459	3.509
				2.4	3.639	43.472	2.186	5.538	3.709
				2.3	4.748	47.959	2.302	5.675	3.989
				2.2	6.148	52.401	2.417	5.828	4.303
				2.1	7.885	56.833	2.532	6.078	4.727
				2.0	10.058	61.265	2.647	6.318	5.162
				1.9	12.731	65.505	2.769	6.693	5.675
				1.8	15.822	69.529	2.874	7.087	6.134
				1.7	19.423	73.245	2.979	7.479	6.540
				1.6	23.581	77.073	3.063	7.935	7.047
				1.5	28.211	80.412	3.136	8.417	7.337
				1.4	33.356	83.625	3.190	8.971	7.550
				1.3	38.808	86.227	3.213	9.405	7.658
				1.2	44.655	88.698	3.219	9.757	7.607
				1.1	50.763	90.952	3.189	10.084	7.498
				1.0	57.042	92.615	3.135	10.166	7.198

$\text{MSE}_\mu/\text{MSE}_\sigma$ = MSE of mean/standard deviation estimates $\times 1,000$.

Table 3: ANI, TAP, FAP and MSE metrics with $m = 50$, $n = 5$, $c = 4\,\%$ and shift = 2 for Phase I data sets.

m	n	μ_1	c	L	FAP	TAP	ANI	MSE_μ	MSE_σ
50	5	2	4	5.0	0.002	25.615	1.447	8.044	2.668
				4.9	0.002	28.743	1.493	7.749	2.659
				4.8	0.002	32.228	1.544	7.513	2.690
				4.7	0.005	35.950	1.596	7.293	2.692
				4.6	0.005	39.786	1.645	6.974	2.700
				4.5	0.008	43.713	1.695	6.724	2.690
				4.4	0.011	47.729	1.743	6.493	2.689
				4.3	0.016	51.692	1.785	6.246	2.711
				4.2	0.021	55.814	1.828	6.027	2.720
				4.1	0.028	59.729	1.863	5.819	2.716
				4.0	0.039	63.840	1.899	5.570	2.734
				3.9	0.055	67.385	1.926	5.462	2.751
				3.8	0.072	71.234	1.956	5.238	2.746
				3.7	0.099	74.671	1.976	5.082	2.767
				3.6	0.135	77.853	1.993	4.997	2.793
				3.5	0.185	80.785	2.009	4.859	2.826
				3.4	0.248	83.258	2.021	4.817	2.834
				3.3	0.327	85.885	2.033	4.663	2.864
				3.2	0.446	88.012	2.044	4.639	2.900
				3.1	0.594	89.955	2.057	4.602	2.945
				3.0	0.788	91.701	2.071	4.622	2.999
				2.9	1.045	93.255	2.089	4.593	3.076
				2.8	1.370	94.458	2.113	4.640	3.140
				2.7	1.800	95.611	2.140	4.683	3.270
				2.6	2.328	96.467	2.181	4.746	3.357
				2.5	3.012	97.212	2.229	4.868	3.594
				2.4	3.887	97.705	2.293	4.981	3.786
				2.3	5.036	98.312	2.370	5.088	4.073
				2.2	6.472	98.693	2.458	5.307	4.383
				2.1	8.245	98.993	2.557	5.584	4.805
				2.0	10.457	99.251	2.668	5.833	5.241
				1.9	13.150	99.460	2.780	6.232	5.745
				1.8	16.276	99.589	2.885	6.654	6.201
				1.7	19.912	99.709	2.990	7.126	6.607
				1.6	24.083	99.788	3.073	7.624	7.114
				1.5	28.719	99.843	3.144	8.276	7.404
				1.4	33.868	99.895	3.193	9.052	7.601
				1.3	39.300	99.922	3.213	9.699	7.713
				1.2	45.119	99.946	3.215	10.329	7.681
				1.1	51.211	99.964	3.183	11.067	7.542
				1.0	57.449	99.979	3.127	11.625	7.254

$\text{MSE}_\mu/\text{MSE}_\sigma$ = MSE of mean/standard deviation estimates $\times 1,000$.

Table 4: ANI, TAP, FAP and MSE metrics with $m = 50$, $n = 5$, $c = 4\,\%$ and shift = 3 for Phase I data sets.

m	n	μ_1	c	L	FAP	TAP	ANI	MSE_μ	MSE_σ
50	5	3	4	5.0	0.002	93.588	2.024	4.673	2.746
				4.9	0.002	94.754	2.024	4.529	2.737
				4.8	0.003	95.867	2.021	4.449	2.760
				4.7	0.005	96.621	2.019	4.429	2.756
				4.6	0.006	97.362	2.016	4.348	2.760
				4.5	0.009	97.954	2.014	4.281	2.746
				4.4	0.012	98.388	2.012	4.274	2.744
				4.3	0.017	98.732	2.010	4.265	2.767
				4.2	0.023	99.019	2.009	4.252	2.767
				4.1	0.031	99.277	2.008	4.230	2.763
				4.0	0.043	99.457	2.007	4.186	2.771
				3.9	0.061	99.571	2.007	4.241	2.787
				3.8	0.081	99.715	2.007	4.213	2.780
				3.7	0.110	99.786	2.009	4.197	2.794
				3.6	0.149	99.830	2.010	4.240	2.819
				3.5	0.205	99.896	2.014	4.217	2.845
				3.4	0.274	99.916	2.018	4.287	2.858
				3.3	0.365	99.949	2.023	4.228	2.883
				3.2	0.492	99.960	2.032	4.277	2.913
				3.1	0.656	99.973	2.045	4.306	2.956
				3.0	0.867	99.981	2.061	4.392	3.012
				2.9	1.144	99.983	2.080	4.421	3.088
				2.8	1.499	99.991	2.109	4.506	3.150
				2.7	1.963	99.994	2.143	4.594	3.282
				2.6	2.528	99.998	2.191	4.693	3.369
				2.5	3.258	99.997	2.245	4.851	3.607
				2.4	4.184	100.000	2.315	4.989	3.801
				2.3	5.398	100.000	2.400	5.143	4.092
				2.2	6.888	100.000	2.491	5.377	4.400
				2.1	8.732	100.000	2.594	5.682	4.825
				2.0	11.011	100.000	2.708	5.959	5.274
				1.9	13.763	100.000	2.818	6.394	5.774
				1.8	16.960	100.000	2.927	6.869	6.241
				1.7	20.643	100.000	3.028	7.417	6.644
				1.6	24.858	100.000	3.106	8.039	7.157
				1.5	29.514	100.000	3.174	8.857	7.445
				1.4	34.655	100.000	3.219	9.883	7.634
				1.3	40.102	100.000	3.237	10.865	7.760
				1.2	45.877	100.000	3.225	11.925	7.721
				1.1	51.911	100.000	3.188	13.204	7.564
				1.0	58.108	100.000	3.131	14.480	7.296

$\text{MSE}_\mu/\text{MSE}_\sigma$ = MSE of mean/standard deviation estimates $\times 1,000$.

Table 5: ANI, TAP, FAP and MSE metrics with $m = 50$, $n = 5$, $c = 8\,\%$ and shift = 2 for Phase I data sets.

m	n	μ_1	c	L	FAP	TAP	ANI	MSE_μ	MSE_σ
50	5	2	8	5.0	0.002	21.297	1.643	19.960	2.681
				4.9	0.002	24.705	1.715	18.713	2.659
				4.8	0.003	28.051	1.784	17.568	2.705
				4.7	0.004	31.828	1.855	16.486	2.703
				4.6	0.006	35.832	1.927	15.202	2.740
				4.5	0.009	39.882	1.987	14.083	2.740
				4.4	0.012	44.200	2.048	12.850	2.736
				4.3	0.017	48.321	2.096	11.919	2.730
				4.2	0.024	52.772	2.142	10.889	2.788
				4.1	0.035	57.050	2.181	9.977	2.775
				4.0	0.049	61.282	2.212	9.116	2.803
				3.9	0.065	65.320	2.230	8.460	2.830
				3.8	0.090	69.223	2.245	7.813	2.833
				3.7	0.120	73.054	2.258	7.164	2.862
				3.6	0.164	76.410	2.256	6.694	2.889
				3.5	0.229	79.744	2.260	6.312	2.901
				3.4	0.304	82.572	2.254	5.961	2.937
				3.3	0.408	85.260	2.251	5.718	2.961
				3.2	0.548	87.542	2.240	5.478	2.995
				3.1	0.730	89.608	2.235	5.394	3.051
				3.0	0.960	91.407	2.236	5.247	3.106
				2.9	1.273	92.929	2.233	5.209	3.192
				2.8	1.659	94.270	2.245	5.126	3.291
				2.7	2.146	95.351	2.261	5.205	3.378
				2.6	2.782	96.271	2.290	5.257	3.570
				2.5	3.574	97.045	2.330	5.346	3.754
				2.4	4.595	97.692	2.389	5.465	3.992
				2.3	5.836	98.182	2.453	5.666	4.241
				2.2	7.410	98.589	2.538	5.937	4.562
				2.1	9.360	98.911	2.631	6.252	4.979
				2.0	11.735	99.168	2.735	6.590	5.512
				1.9	14.543	99.375	2.835	7.093	5.958
				1.8	17.845	99.516	2.938	7.669	6.432
				1.7	21.606	99.648	3.032	8.386	6.981
				1.6	25.880	99.731	3.114	9.274	7.365
				1.5	30.588	99.793	3.174	10.303	7.690
				1.4	35.638	99.844	3.196	11.463	7.882
				1.3	41.189	99.891	3.221	13.104	8.003
				1.2	46.905	99.918	3.200	14.812	7.876
				1.1	52.900	99.940	3.163	16.743	7.813
				1.0	59.015	99.959	3.097	19.088	7.545

$\text{MSE}_\mu/\text{MSE}_\sigma$ = MSE of mean/standard deviation estimates $\times 1,000$.

Table 6: ANI, TAP, FAP and MSE metrics with $m = 50$, $n = 10$, $c = 4\,\%$ and shift = 2 for Phase I data sets.

m	n	μ_1	c	L	FAP	TAP	ANI	MSE_μ	MSE_σ
50	10	2	4	5.0	0.001	88.124	2.026	2.496	1.179
				4.9	0.001	89.961	2.026	2.405	1.185
				4.8	0.002	91.526	2.028	2.356	1.188
				4.7	0.002	93.113	2.025	2.294	1.189
				4.6	0.003	94.395	2.024	2.251	1.175
				4.5	0.004	95.403	2.022	2.218	1.185
				4.4	0.006	96.339	2.019	2.195	1.194
				4.3	0.009	97.161	2.017	2.160	1.193
				4.2	0.015	97.718	2.015	2.159	1.189
				4.1	0.018	98.205	2.013	2.141	1.195
				4.0	0.026	98.595	2.012	2.129	1.193
				3.9	0.038	98.899	2.010	2.112	1.187
				3.8	0.056	99.194	2.010	2.124	1.201
				3.7	0.076	99.401	2.010	2.117	1.195
				3.6	0.108	99.513	2.010	2.096	1.208
				3.5	0.149	99.678	2.012	2.120	1.202
				3.4	0.207	99.750	2.014	2.119	1.214
				3.3	0.282	99.819	2.019	2.146	1.222
				3.2	0.392	99.866	2.026	2.133	1.230
				3.1	0.530	99.910	2.034	2.153	1.248
				3.0	0.718	99.935	2.047	2.180	1.257
				2.9	0.961	99.954	2.062	2.207	1.275
				2.8	1.289	99.968	2.082	2.235	1.301
				2.7	1.722	99.979	2.112	2.291	1.338
				2.6	2.285	99.988	2.144	2.321	1.372
				2.5	2.987	99.994	2.187	2.383	1.417
				2.4	3.909	99.996	2.237	2.479	1.475
				2.3	5.080	99.996	2.297	2.569	1.539
				2.2	6.531	99.998	2.364	2.676	1.629
				2.1	8.316	99.997	2.438	2.798	1.720
				2.0	10.465	100.000	2.518	2.950	1.825
				1.9	13.014	100.000	2.596	3.144	1.933
				1.8	16.057	100.000	2.675	3.362	2.045
				1.7	19.602	100.000	2.756	3.585	2.180
				1.6	23.622	100.000	2.831	3.910	2.312
				1.5	28.225	100.000	2.895	4.252	2.441
				1.4	33.240	100.000	2.949	4.623	2.546
				1.3	38.650	100.000	2.979	5.088	2.621
				1.2	44.515	100.000	3.002	5.662	2.681
				1.1	50.665	100.000	2.998	6.202	2.717
				1.0	57.024	100.000	2.974	6.849	2.737

MSE_μ/MSE_σ = MSE of mean/standard deviation estimates $\times 1,000$.

Table 7: L values corresponding to "minimum and 5 % upper" bound for the MSE of the mean estimates with various m, c, and shift values μ_1 where $n = 5$.

n	m	μ_1	c	Minimum MSE_μ	5 % bound MSE_μ	L for Minimum MSE_μ	L for 5 % bound MSE_μ
5	25	0	0	7.935	8.331	3.1	5.0
	50	0	0	3.983	4.182	2.9	5.0
	100	0	0	1.989	2.089	2.7	5.0
	25	1	4	9.515	9.991	2.8	5.0
	50	1	4	5.275	5.539	2.4	4.2
	100	1	4	2.915	3.061	2.2	3.3
	25	2	4	9.215	9.676	2.5	3.5
	50	2	4	4.593	4.823	2.6	3.4
	100	2	4	2.298	2.413	2.5	3.4
	25	3	4	8.401	8.821	2.9	4.6
	50	3	4	4.186	4.395	2.9	4.8
	100	3	4	2.100	2.205	2.9	4.8
	25	1	8	12.878	13.521	2.2	3.4
	50	1	8	7.440	7.812	2.0	2.8
	100	1	8	4.419	4.640	1.9	2.4
	25	2	8	10.286	10.801	2.4	3.1
	50	2	8	5.126	5.382	2.5	3.1
	100	2	8	2.625	2.757	2.4	3.2
	25	3	8	8.890	9.334	3.0	4.3
	50	3	8	4.414	4.634	3.1	4.4
	100	3	8	2.209	2.319	3.0	4.5
10	25	0	0	3.978	4.176	2.8	5.0
	50	0	0	1.989	4.408	2.7	5.0
	100	0	0	0.986	1.035	2.9	5.0
	25	1	4	4.961	5.209	2.4	3.5
	50	1	4	2.576	2.704	2.3	3.2
	100	1	4	1.327	1.394	2.2	3.0
	25	2	4	4.209	4.420	2.8	4.4
	50	2	4	2.096	2.201	2.8	4.5
	100	2	4	1.056	1.109	2.8	4.7
	25	3	4	4.144	4.351	2.9	5.0
	50	3	4	2.069	2.172	2.9	5.0
	100	3	4	1.034	1.086	2.9	5.0
	25	1	8	6.129	6.435	2.1	2.8
	50	1	8	3.273	3.437	2.1	2.7
	100	1	8	1.789	1.879	2.0	2.6
	25	2	8	4.466	4.689	2.9	4.1
	50	2	8	2.213	2.324	3.0	4.2
	100	2	8	1.112	1.168	2.8	4.4
	25	3	8	4.336	4.553	3.2	5.0
	50	3	8	2.160	2.267	3.2	5.0
	100	3	8	1.082	1.136	3.4	5.0

MSE_μ = MSE of mean estimates $\times 1,000$.

Table 8: L values corresponding to minimum and 5 % upper bound for the MSE of the standard deviation estimates with various m, c, and shift values μ_1 where $n = 5$.

n	m	μ_1	c	Minimum MSE_σ	5 % bound MSE_σ	L for Minimum MSE_σ	L for 5 % bound MSE_σ
5	25	0	0	5.258	5.521	3.3	5.0
	50	0	0	2.634	2.765	3.1	5.0
	100	0	0	1.317	1.383	3.1	5.0
	25	1	4	5.275	5.539	3.4	5.0
	50	1	4	2.629	2.760	3.2	5.0
	100	1	4	1.315	1.381	3.2	5.0
	25	2	4	5.344	5.611	3.5	5.0
	50	2	4	2.659	2.792	3.4	5.0
	100	2	4	1.338	1.405	3.3	5.0
	25	3	4	5.476	5.750	3.2	5.0
	50	3	4	2.737	2.874	3.2	5.0
	100	3	4	1.369	1.437	3.3	5.0
	25	1	8	5.272	5.535	3.4	5.0
	50	1	8	2.610	2.740	3.5	5.0
	100	1	8	1.314	1.379	3.2	5.0
	25	2	8	5.378	5.646	3.9	5.0
	50	2	8	2.659	2.792	3.8	5.0
	100	2	8	1.342	1.409	3.7	5.0
	25	3	8	5.709	5.995	3.3	5.0
	50	3	8	2.824	2.965	3.2	5.0
	100	3	8	1.427	1.499	3.2	5.0
10	25	0	0	2.253	2.366	2.9	5.0
	50	0	0	1.137	1.194	2.9	5.0
	100	0	0	0.569	0.597	2.7	5.0
	25	1	4	2.260	2.373	3.2	5.0
	50	1	4	1.134	1.191	3.0	5.0
	100	1	4	0.568	0.596	2.8	5.0
	25	2	4	2.336	2.453	3.0	5.0
	50	2	4	1.175	1.234	3.1	5.0
	100	2	4	0.587	0.616	3.0	5.0
	25	3	4	2.350	2.467	3.0	5.0
	50	3	4	1.179	1.238	3.1	5.0
	100	3	4	0.587	0.617	3.0	5.0
	25	1	8	2.262	2.375	3.3	5.0
	50	1	8	1.141	1.198	3.1	5.0
	100	1	8	0.569	0.597	2.9	5.0
	25	2	8	2.438	2.560	3.0	5.0
	50	2	8	1.227	1.289	2.9	5.0
	100	2	8	0.611	0.641	3.2	5.0
	25	3	8	2.455	2.578	3.1	5.0
	50	3	8	1.236	1.298	2.9	5.0
	100	3	8	0.616	0.647	3.2	5.0

MSE_σ = MSE of standard deviation estimates $\times 1,000$.

Table 9: L values selected for approximately MSE optimal Phase I design for mean estimation.

$n = 5$	L	# intersecting intervals (out of 7)
$m = 25$	3.1	7
$m = 50$	3.1	6
$m = 100$	3.0	6

$n = 10$	L	# intersecting intervals (out of 7)
$m = 25$	3.2	6
$m = 50$	3.2	6
$m = 100$	2.9 or 3.0	5

The L values corresponding to the "minimum and 5 % upper deviation" bounds for the MSE are presented in Table 7 for the mean estimates and in Table 8 for the standard deviation estimates, for the combinations of $m = 25, 50, 100$, $n = 5, 10$, $c = 0, 4, 8$ %, and shift values $\mu_1 = 0, 1, 2, 3$. The "upper 5 % deviation" bound is considered here to identify intervals for L for a given m and n pair that yield close to minimum MSE. To provide suggestions for practitioners, we searched for L values that are robust over the considered contamination percentages $c = 4$ and 8 %, shifts 0, 1, 2, and 3, as well as their combinations. Intervals for L values were obtained for the 7 cases considered (1 in-control, and 3 out-of-control cases each with 2 contamination percentages) for each of the m and n pairs. In order to select a single L value that is robust, a rule was developed such that the L value that satisfies most if not all of the intervals was selected. Considering joint operation of $\overline{x}$ and s charts with the use of same L value for both charts, L values that provide approximately MSE optimal estimates of the mean for rational subgroups are provided in Table 9.

Overall, it becomes clear that the MSE optimal L values are often close to the popular choice of 3σ 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 7 indicates that the MSE optimal L should be slightly lower than 3.0 for situations where the mean shift is expected to be of at most moderate extent.

4 Conclusions

Considering rational subgroups as the sampling strategy, Shewhart control charts for the mean and standard deviation were considered in Phase I implementations. The normal distribution was assumed as the model for the observations. Different scenarios to represent practical situations were simulated and the results for the selected performance metrics were provided. These performance metrics quantify the computational efforts, false and true signals, and the accuracy of the estimates. By varying the control chart design parameter L, robust L values that would perform close to optimal in terms of the MSE criterion were searched. As a follow up study, the evaluation of the Phase II performance of the control charts with the mean and standard deviation estimates obtained by using alternative L values in Phase I implementations is being conducted.

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.

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 edition, John Wiley & Sons, Inc., New York.

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

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.

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.

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 & Crosier (1982) 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 the 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.

1 Introduction

The general theme of this work is the 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). Set up to detect a possible change in the baseline mean of a series of independent samples $X_1, X_2, \ldots$ drawn from a normal unit-variance population at regular time intervals, the classical SR chart 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

A.S. Polunchenko
Department of Mathematical Sciences, State University of New York at Binghamton, Binghamton, New York 13902–6000, USA e-mail: aleksey@binghamton.edu

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

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 & 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 & Pollak (1991). See also, e.g., Tartakovsky & 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 & Pollak (1996). Kenett & 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 & Rubin (1952). The chart then remained unnoticed until recently Pollak & Tartakovsky (2009) and Shiryaev & 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 & Polunchenko, 2016), and economic design of control charts. This brought the SR chart back into the spotlight. Polunchenko et al. (2016) 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 Section 2 for a formal definition. As a matter of fact Polunchenko & Tartakovsky (2010) and Tartakovsky & 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 essentially nonexistent. In part, this may be due to the lack of existing resources with 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 & 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 Section 2. The optimization itself is carried out in Section 3 using the numerical framework developed by Moustakides et al. (2011) and then improved upon by Polunchenko et al. (2014b, 2014a). The obtained optimal headstart and control limit values are reported in Section 3 as well. Conclusions follow in Section 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 $\{R^r_n\}_{n\geq 0}$ via the recurrence

$$R^r_n = (1 + R^r_{n-1})\exp\{S_n\},\ n = 1, 2, \dots \text{ with } R^r_0 = r \geq 0, \tag{2}$$

where S_n is the score function defined in (1); the initial score $R^r_0 = r \geq 0$ is a design parameter also referred to as the headstart, which is the original terminology of Lucas & Crosier (1982) who suggested to headstart the CUSUM scheme. The corresponding run length is as follows:

$$\mathcal{S}^r_A \triangleq \min\{n \geq 1 :\ R^r_n \geq A\}, \tag{3}$$

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)—are indifferent to the direction of the mean shift, i.e., the sign of $\mu \neq 0$ is irrelevant.

It has been the custom in the quality control literature to assess the operating characteristics of a control chart, with run length T, by means of only two indices: the in-control Average Run Length (ARL) and the out-of-control ARL. 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).

The in-control characteristics of a control chart T are usually gauged by virtue of the Average Run Length (ARL) to false alarm $\mathrm{ARL}(T) \triangleq \mathbb{E}_\infty[T]$ which 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 *steady-state* ARL.

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 (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 the process 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 out-of-control ARL. However, things are not as simple when the chart has a positive headstart, and 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 $\{\mathrm{ADD}_k(\mathcal{S}_A^r)\}_{k\ge0}$ 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 preset 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 & Tartakovsky (2010) and Tartakovsky & 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 $\mathrm{ARL}(\mathcal{S}_A^r)\ge\gamma$ with $\gamma>1$ given, i.e., $\mathcal{S}_A^r\in\Delta(\gamma)$, then

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

provided, however, that $r/A\to0$ 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 chat (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 & Tartakovsky (2009) and Shiryaev & 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 k-average of the delays $\mathbb{E}_k[\max\{0, T-k\}]$, $k = 0, 1, 2, \ldots$, assuming that change-point k 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, $\mathrm{RIADD}(T)$ can also be regarded as the k-average of the $\mathrm{ADD}_k(T)$'s 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 & Tartakovsky (2009).

The RIADD-optimality of the classical SR chart was generalized in (Polunchenko & 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 so-called Stationary ADD (STADD)

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

inside class $\Delta(\gamma)$, for any $\gamma > 1$; recall that $\mathrm{IADD}(T)$ 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 $\mathrm{STADD}(T)$ is not the same as the limit $\lim_{k\to\infty} \mathrm{ADD}_k(T)$.

An important "by-product" of (Polunchenko & Tartakovsky, 2010, Lemma 1) is that the quantity $\mathrm{STADD}(\mathcal{S}_A^r)$ turns out to also provide 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 & Tartakovsky, 2010, Theorem 1). Specifically, introducing $\underline{\mathrm{SADD}}(\mathcal{S}_A^r) \equiv \mathrm{STADD}(\mathcal{S}_A^r)$, the following double inequality holds:

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

for any $A > 0$ and $r \ge 0$ such that $\mathrm{ARL}(\mathcal{S}_A^r) \ge \gamma$, 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 $\mathrm{STADD}(T)$ 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 $r \ge 0$ 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 $\mathrm{ARL}(\mathcal{S}_A^r) \ge \gamma$. More formally, the optimal detection threshold A^* and headstart r^* values are to be selected as follows:

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

   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.

The constrained optimization problem (10) can be solved numerically, e.g., with the aid of the numerical method proposed by Moustakides et al. (2011) and subsequently improved upon by Polunchenko et al. (2014b, 2014a). This is precisely the object of the next section.

3 Experimental Results

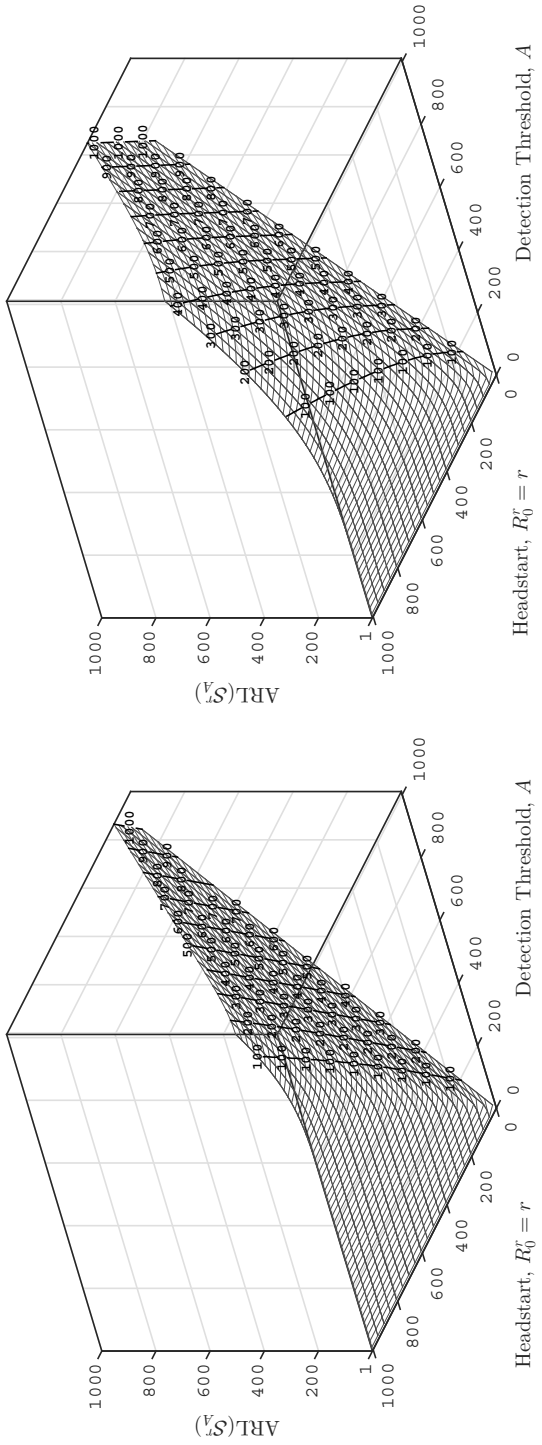
The plan now is to employ the numerical framework of Moustakides et al. (2011) and its improved version due to Polunchenko et al. (2014b, 2014a), and analyze the performance of the GSR chart given by (2) and (3) under different parameter

settings, including (and especially) the optimal choice given by the solution of the constrained optimization problem (10).

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. Figures 1 depict $\mathrm{ARL}(\mathcal{S}_A^r)$ as a function of $r \in [0, A]$ and $A \in [0, 1\,000]$. Specifically, Figure 1a is for $\mu = 0.2$ and Figure 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. Figures 1 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 $\gamma = \{100, 200, \ldots, 900, 1\,000\}$. 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, $\gamma = \{100, 500, 1\,000\}$.

Let us next look at Figures 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 $\gamma = \{100, 500, 1\,000\}$. Specifically, Figures 2 assume $\mu = 0.2$ while Figures 3 assume $\mu = 0.5$. With regard to the level $\gamma > 1$ of the ARL to false alarm, Figures 2a and 3a assume $\gamma = 100$, Figures 2b and 3b are for $\gamma = 500$, and Figures 2c and 3c assume $\gamma = 1\,000$. 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 ADDs to the steady-state regime is faster for $\mu = 0.5$ than for $\mu = 0.2$, which is consistent with one's intuition.

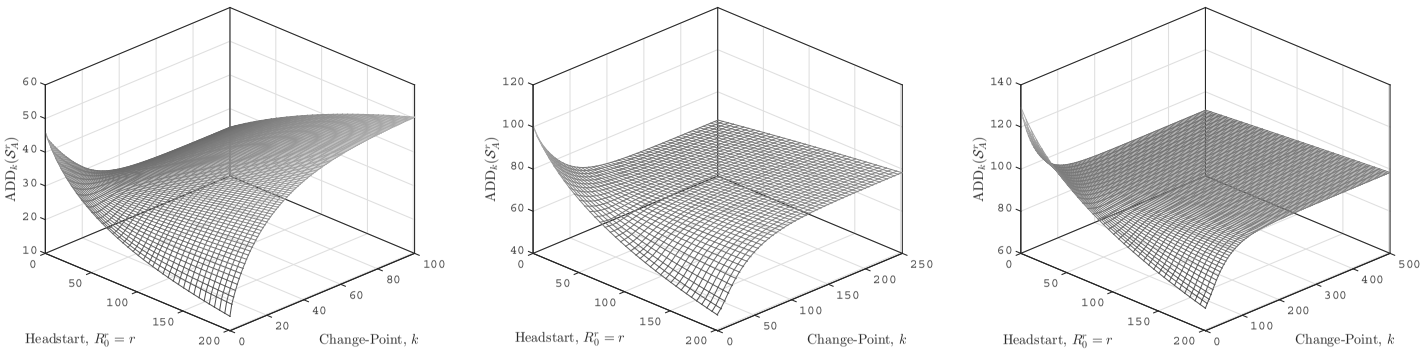
To better illustrate the FIR feature at work, let us look at Figures 4 and 5, which are effectively the projections of the 3D surfaces shown in Figures 2 and 3 onto the $(k, \mathrm{ADD}_k(\mathcal{S}_A^r))$-plane, made for a selection of values of r. Specifically, Figures 4 assume $\mu = 0.2$ and Figures 4 are 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



(a) $\mu = 0.2$.

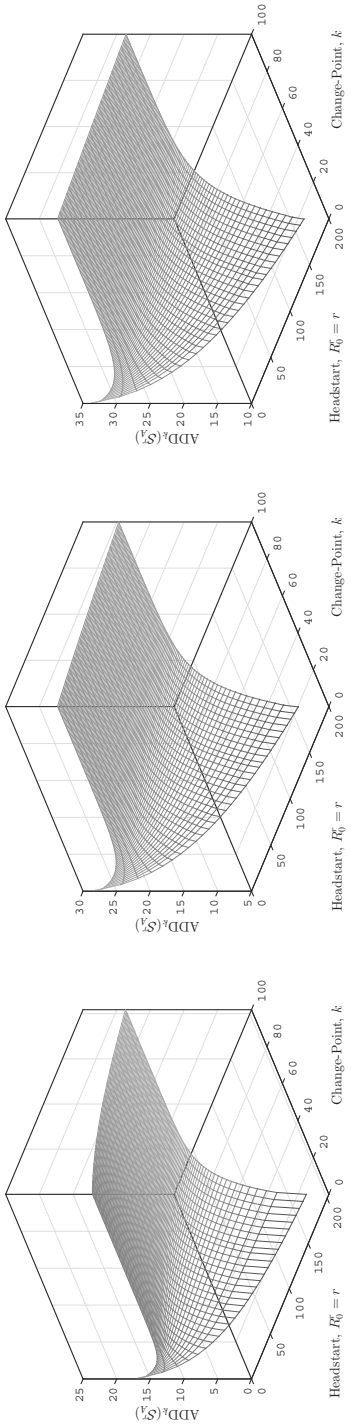
(b) $\mu = 0.5$.

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) $\gamma = 100$. (b) $\gamma = 500$. (c) $\gamma = 1\,000$.

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) $\gamma = 1\,000$.

Fig. 3: $\mathrm{ADD}_k(S_A^r)$ 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 $\mathrm{ARL}(S_A^r) = \gamma > 1$ for $\mu = 0.5$.

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 $\mathrm{SADD}(\mathcal{S}_A^r)$ involved, as is done in Figures 6 and 7.

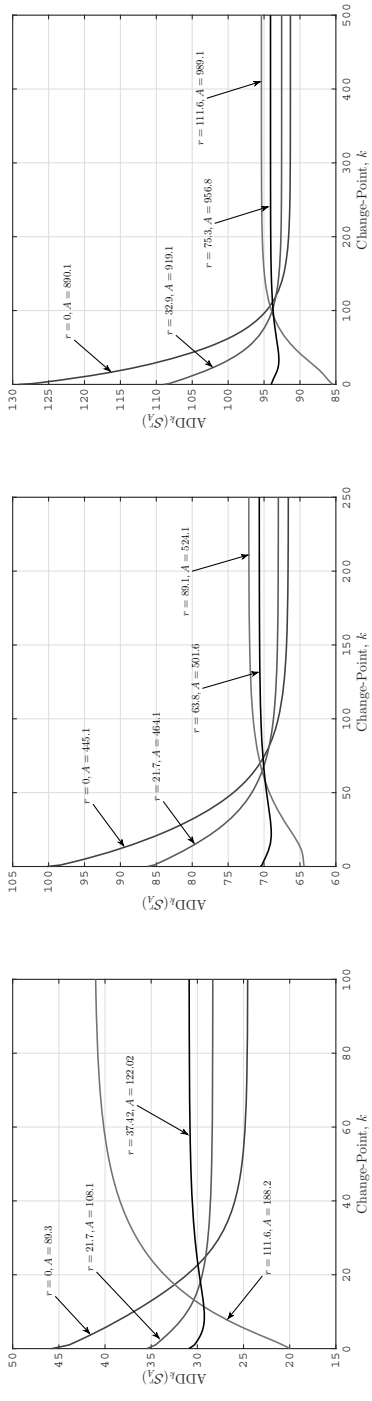
Specifically, Figures 6 and 7 provide an idea as to the manner in which $\mathrm{SADD}(\mathcal{S}_A^r)$ and $\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$ each depend on the headstart, assuming, as before, that every change in the headstart is accompanied by the appropriate adjustment of the detection threshold, so that the ARL to false alarm constraint is kept intact. More specifically, Figures 6 correspond to $\mu = 0.2$ and Figures 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 $\gamma = \{100, 200, \ldots, 900, 1\,000\}$ 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 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.

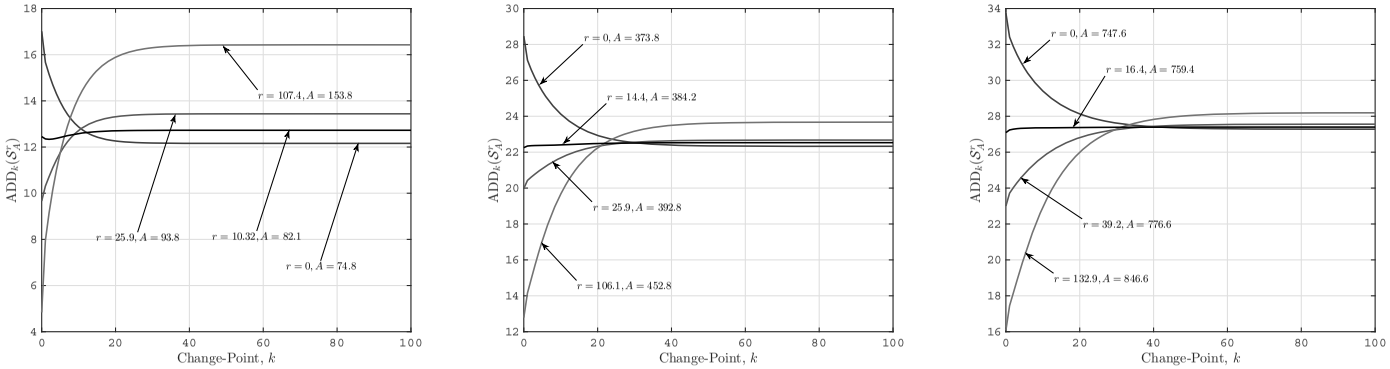
4 Concluding Remarks

In summary we see that



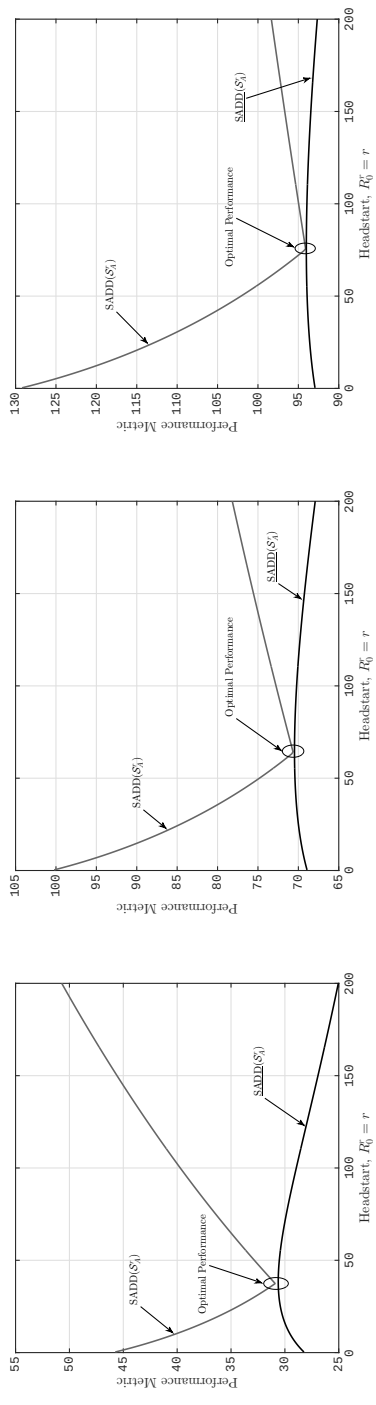
(a) $\gamma = 100$. (b) $\gamma = 500$. (c) $\gamma = 1\,000$.

Fig. 4: $\mathrm{ADD}_k(S_A^r)$ 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 $\mathrm{ARL}(S_A^r) = \gamma > 1$ for $\mu = 0.2$.



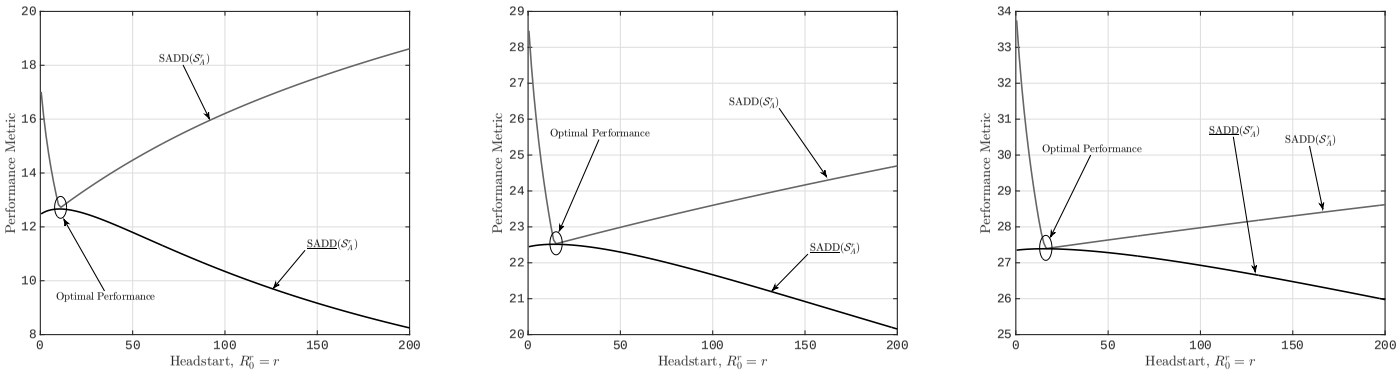
(a) $\gamma = 100$. (b) $\gamma = 500$. (c) $\gamma = 1\,000$.

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) $\gamma = 1\,000$.

Fig. 6: SADD($\mathcal{S}'_A$) and $\underline{\text{SADD}}(\mathcal{S}'_A)$ as functions of the headstart $R'_0 = r$ and the ARL to false alarm level $\text{ARL}(\mathcal{S}'_A) = \gamma > 1$ for $\mu = 0.2$.



(a) $\gamma = 100$. (b) $\gamma = 500$. (c) $\gamma = 1\,000$.

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 $\mu = 0.5$.

Table 1: Optimal headstart, $r^* > 0$, control limit, $A^* > 0$, maximal ADD, 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 $\gamma = \{100, 200, 300, 400, 500\}$.

$\text{ARL}(\mathcal{S}_A^r) = \gamma$	**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
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
	$\text{SADD}(\mathcal{S}_A^r)$	49.65	30.9	21.6	16.17	12.68	10.28	8.55	7.25	6.25	5.46
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	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
	$\text{SADD}(\mathcal{S}_A^r)$	79.79	45.39	30.1	21.75	16.61	13.19	10.79	9.03	7.69	6.65
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	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
	$\text{SADD}(\mathcal{S}_A^r)$	103.23	55.71	35.87	25.41	19.13	15.04	12.19	10.13	8.58	7.38
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	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
	$\text{SADD}(\mathcal{S}_A^r)$	122.8	63.86	40.29	28.17	21.02	16.4	13.22	10.93	9.22	7.91
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	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
	$\text{SADD}(\mathcal{S}_A^r)$	139.75	70.63	43.86	30.39	22.52	17.48	14.05	11.57	9.73	8.33
	$\underline{\text{SADD}}(\mathcal{S}_A^r)$	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, $\mathrm{SADD}(\mathcal{S}_A^r)$, 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 $\gamma = \{600, 700, 800, 900, 1\,000\}$.

$\mathrm{ARL}(\mathcal{S}_A^r) = \gamma$	**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
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
	$\mathrm{SADD}(\mathcal{S}_A^r)$	153.8	76.46	46.95	32.26	23.77	18.38	14.71	12.09	10.15	8.67
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	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
	$\mathrm{SADD}(\mathcal{S}_A^r)$	168.37	81.58	49.59	33.87	24.85	19.15	15.29	12.54	10.51	8.96
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	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
	$\mathrm{SADD}(\mathcal{S}_A^r)$	180.76	86.16	51.94	35.29	25.79	19.82	15.79	12.93	10.82	9.22
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	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^*	1 040.31	866.3	787.3	731.22	684.37	642.75	604.77	569.66	536.98	506.46
	$\mathrm{SADD}(\mathcal{S}_A^r)$	192.18	90.3	54.04	36.55	26.64	20.42	16.24	13.28	11.1	9.44
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	191.21	90.21	54.03	36.55	26.63	20.42	16.24	13.28	11.1	9.44
1 000	r^*	210.04	75.34	38.62	23.71	16.14	11.73	8.93	7.04	5.68	4.66
	A^*	1 141.3	956.81	871.86	810.77	759.35	713.44	671.46	632.6	596.38	562.54
	$\mathrm{SADD}(\mathcal{S}_A^r)$	202.79	94.09	55.96	37.7	27.39	20.96	16.64	13.59	11.35	9.65
	$\underline{\mathrm{SADD}}(\mathcal{S}_A^r)$	201.86	94.01	55.94	37.69	27.39	20.95	16.64	13.59	11.35	9.64

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.
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 & 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 & Tartakovsky (2009), Shiryaev & Zryumov (2009), Tartakovsky & Polunchenko (2010), Polunchenko & 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's effort was partially supported by the Simons Foundation via a Collaboration Grant in Mathematics under Award # 304574.

References

Girshick MA and Rubin H (1952). A Bayes approach to a quality control model. *Ann Math Statist* 23(1), 114-125. doi:10.1214/aoms/1177729489.

Kenett R and Pollak M (1986). A semi-parametric approach to testing for reliability growth, with application to software systems. *IEEE Trans Rel* 35(3), 304-311. doi:10.1109/TR.1986.4335439.

Kenett R and Pollak M (1996). Data-analytic aspects of the Shiryayev-Roberts control chart: Surveillance of a non-homogeneous Poisson process. *J Appl Stat* 23(1), 125-138. doi:10.1080/02664769624413.

Lucas JM and Crosier RB (1982). Fast initial response for CUSUM quality-control schemes: Give your CUSUM a head start. *Technometrics* 24(3), 199-205. doi:10.2307/1268679.

Mevorach Y and Pollak M (1991). A small sample size comparison of the CUSUM and Shiryayev-Roberts approaches to changepoint detection. *Amer J Math Management Sci* 11(3&4), 277-298. doi:10.1080/01966324.1991.10737312.

Moustakides GV, Polunchenko AS and Tartakovsky AG (2011). A numerical approach to performance analysis of quickest change-point detection procedures. *Statist Sinica* 21(2), 571-596.

Moustakides GV, Polunchenko AS and Tartakovsky AG (2009). Numerical comparison of CUSUM and Shiryaev-Roberts procedures for detecting changes in distributions. *Commun Stat Theory Methods* 38(16), 3225-3239.

Page ES (1954). Continuous inspection schemes. *Biometrika* 41(1&2), 100-115. doi:10.2307/2333009.

Pepelyshev A and Polunchenko AS (2016). Real-time financial surveillance via quickest change-point detection methods. *Stat Interface* (in press). Available via ArXiv: https://arxiv.org/abs/1509.01570. Cited 3 Jul 2016.

Pollak M (1985). Optimal detection of a change in distribution. *Ann Statist* 13(1), 206-222. doi:10.1214/aos/1176346587.

Pollak M (1987). Average Run Lengths of an optimal method of detecting a change in distribution. *Ann Statist* 15(2), 749-779. doi:10.1214/aos/1176350373.

Pollak M (2009). The Shiryaev-Roberts changepoint detection procedure in retrospect – Theory and practice. In: *Proce 2nd Int'l Workshop Sequential Methodol*, University of Technology of Troyes, Troyes, France, 15-17 Jun, 2009.

Pollak M and Tartakovsky AG (2009). Optimality properties of the Shiryaev-Roberts procedure. *Statist Sinica* 19, 1729-1739.

Polunchenko AS and Sokolov G (2014). Toward optimal design of the Generalized Shiryaev-Roberts procedure for quickest change-point detection under exponential observations. In: *Proc 2014 Int'l Conf Engineering & Telecommunications*, Moscow Institute of Physics and Technology, Moscow, Russia, 26-28 Nov 2014; pp. 51-55. doi:10.1109/EnT.2014.37.

Polunchenko AS, Sokolov G and Du W (2014). Efficient performance evaluation of the generalized Shiryaev-Roberts detection procedure in a multi-cyclic setup. *Appl Stoch Models Bus Ind*, 30(6), 723-739. doi:10.1002/asmb.2026.

Polunchenko AS, Sokolov G and Du W (2014). An accurate method for determining the pre-change Run-Length distribution of the Generalized Shiryaev-Roberts detection procedure. *Sequential Anal* 33(1), 112-134. doi:10.1080/07474946.2014.856642.

Polunchenko AS, Sokolov G and Du W (2016). On robustness of the Shiryaev-Roberts change-point detection procedure under parameter misspecification in the post-change distribution. *Commun Stat Simul Comput*, (in press). doi:10.1080/03610918.2015.1039131. Available via ArXiv: https://arxiv.org/abs/1504.04722. Cited 3 Jul 2016.

Polunchenko AS and Tartakovsky AG (2010). On optimality of the Shiryaev-Roberts procedure for detecting a change in distribution. *Ann Statist* 38(6), 3445-3457. doi:10.1214/09-AOS775.

Polunchenko AS and Tartakovsky AG (2012). State-of-the-art in sequential change-point detection. *Methodol Comput Appl Probab* 44(3), 649-684. doi:10.1007/s11009-011-9256-5.

Roberts SW (1959). Control chart tests based on geometric moving averages. *Technometrics* 1(3), 239-250. doi:10.2307/1271439.

Roberts SW (1966). A comparison of some control chart procedures. *Technometrics* 8(3), 411-430. doi:10.2307/1266688.

Shiryaev AN (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Math Dokl* 2, 795-799.

Shiryaev AN (1963). On optimum methods in quickest detection problems. *Theory of Probab Appl* 8(1), 22-46. doi:10.1137/1108002.

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

Tartakovsky AG and Ivanova IV (1992). Comparison of some sequential rules for detecting changes in distributions. *Probl Inf Transm* 28(2), 117-124.

Tartakovsky AG, Pollak M and Polunchenko AS (2012). Third-order asymptotic optimality of the Generalized Shiryaev-Roberts changepoint detection procedure. *Theory Probab Appl* 56(3), 457-484. doi:10.1137/S0040585X97985534.

Tartakovsky AG, Polunchenko AS and Sokolov G (2013). Efficient computer network anomaly detection by changepoint detection methods. *IEEE J Sel Top Sign Proces* 7(1), 4-11. doi:10.1109/JSTSP.2012.2233713.

Tartakovsky AG and Polunchenko AS (2010). Minimax optimality the Shiryaev-Roberts procedure. In: *Proc 5th Int'l Workshop Applied Probab*, Universidad Carlos III de Madrid, Colmenarejo Campus, Spain, 5-8 Jul, 2010.

Tartakovsky AG, Polunchenko AS and Moustakides GV (2009). Design and comparison of Shiryaev-Roberts- and CUSUM-type change-point detection procedures. In: *Proc 2nd Int'l Workshop Sequential Methodol*, University of Technology of Troyes, Troyes, France, 15-17 Jun, 2009.

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

Optimal Designs of Unbalanced Nested Designs for Determination of Measurement Precision

Seiichi Yasui and Yoshikazu Ojima

Abstract Precision of measurement results can be recognized as variance components of random effect models. The variance components are estimated from measurement results that are taken by conducting a collaborative assessment experiment. The measurement results follow a statistical model of a nested design. Although balanced nested designs are widely used, staggered nested designs, which are one of unbalanced nested designs, have the statistical advantage that degrees of freedom in all stages except for the top stage are equal. Thus, the balanced nested designs do not necessarily have the better performance from the statistical viewpoints. In this study, the *D*-optimal designs are identified in the general nested designs, which including both balanced and unbalanced ones, with considering the practical feasibility of collaborative assessment experiments as well.

1 Introduction

Nested designs are used to statistically determine precision of precision of measurement results in ISO 5725-1 (1994). In this standard, precision of measurement is defined as "the closeness of agreement between independent test results obtained under stipulated condition". This definition implies that the precision depends on the condition in which objects are measured. The important conditions are repeatability and reproducibility conditions. Repeatability conditions define that "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

Seiichi Yasui
Tokyo University of Science, 2641,Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: yasui@rs.tus.ac.jp

Yoshikazu Ojima
Tokyo University of Science, 2641,Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: ojima@rs.tus.ac.jp

short intervals of time". Reproducibility conditions define that "conditions where test results are obtained with the same method on identical test items in the different laboratories by the different operators using the different equipment". In other word, the repeatability condition means the condition in which dispersion of measurement results is minimal, and the reproducibility condition means the condition in which dispersion of measurement results is maximal in the range of our interest. The precisions under such conditions are called repeatability precision and reproducibility condition, respectively.

These precisions are determined as the variance of measurement results that are obtained from nested designs. Measurement results from nested designs are expressed as hierarchically random effect models, and the precisions are usually estimated by the linear combination of the variance component estimators based on analysis of variance.

Although balanced nested designs are widely used, staggered nested designs, which are one of unbalanced nested designs, have the statistical advantage that degrees of freedom in all stages except for the top stage are equal. Thus, the balanced nested designs do not necessarily have the better performance from the statistical viewpoints, and there are favourable situations for unbalanced nested designs. In our study, we focus on three stage nested designs, and the D-optimal designs with respect to the estimation of repeatability, intermediate, and reproducibility precisions are investigated in some situations regarding some magnitudes of variance components and given sample sizes.

Goldsmith and Gaylor (1970) researched the optimal three stage nested designs for the estimation precision of variance components with respect to A-, D-, and adjusted (scaled) A-optimality. Goldsmith and Gaylor (1970), however, found optimal designs under the quite restricted situation that sample size is multiples of twelve. It is assumed that the three-stage nested design with twelve observations is replicated as a block. We find D-designs for any sample size under the some variance component configurations by our developing the effective algorithm to search all the unbalanced nested designs in which all the degrees of freedom are non-zero.

In Section 2, we discuss the appropriate estimands to precision of measurements. In Section 3, D-optimal designs for some sample sizes are shown under the some variance component configurations, and Section 4 is the conclusion.

2 D-Optimality for Determination of Measurement Precisions

The statistical model for unbalanced nested designs with three stages is

$$\begin{gathered} y_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk} \\ i = 1, \ldots, a,\ j = 1, \ldots, b_i,\ k = 1, \ldots, r_{ij} \\ \alpha_i \sim i.i.d.N(0, \sigma_A^2),\ \beta_{ij} \sim i.i.d.N(0, \sigma_B^2),\ \varepsilon_{ijk} \sim i.i.d.N(0, \sigma_E^2)\ , \end{gathered} \tag{1}$$

where μ is a general mean (constant). The variances σ_A^2, σ_B^2, and σ_E^2 are called variance components, and the aim of using the nested designs is to estimate the variance components.

The variance components are estimated through an analysis of variance (ANOVA) which is widely used in practice. Such a 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 as follows:

$$\begin{pmatrix} MSA \\ MSB \\ MSE \end{pmatrix} = \begin{pmatrix} l_{AA} & l_{AB} & 1 \\ 0 & l_{BB} & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\sigma}_A^2 \\ \hat{\sigma}_B^2 \\ \hat{\sigma}_E^2 \end{pmatrix}. \tag{2}$$

Let the coefficient matrix of the equation be $\mathbf{L}^{-1}$. Then, the ANOVA estimator is $\mathbf{Lv}$ where $\mathbf{v} = (MSA, MSB, MSE)'$, and the $\mathbf{L}$ is the inverse of the $\mathbf{L}^{-1}$.

In the experiments to determine the precision of the measurement results, repeatability precision, intermediate precision, and reproducibility precision are the important quantities as well as variance components. reproducibility precision, intermediate precision, and repeatability precision are statistically defined as follows:

$$\begin{aligned} \sigma_A^2 + \sigma_B^2 + \sigma_E^2 &\text{ (reproducibility precision)}, \\ \sigma_B^2 + \sigma_E^2 &\text{ (intermediate precision)}, \\ \sigma_E^2 &\text{ (repeatability precision)}, \end{aligned}$$

respectively. Their ANOVA estimators are provided by replacing $(\sigma_A^2, \sigma_B^2, \sigma_E^2)$ to their estimators $(\hat{\sigma}_A^2, \hat{\sigma}_B^2, \hat{\sigma}_E^2)$. Thus, the estimator of these precisions can be expressed as $\mathbf{CLv}$ where

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3}$$

Each element of $\mathbf{CLv}$ in order from the top is the estimator of repeatability, intermediate precision, and reproducibility, respectevely. Note that the estimator of variance components is expressed by matrix form as $\mathbf{C} = \mathbf{I}$, where the $\mathbf{I}$ is an identity matrix.

The variance-covariance matrix of $\mathbf{CLv}$ is $\mathbf{CL}V(\mathbf{v})\mathbf{L}'\mathbf{C}'$. The determinate of the matrix is

$$|\mathbf{CL}V(\mathbf{v})\mathbf{L}'\mathbf{C}'| = \frac{1}{l_{AA}^2 l_{BB}^2} |V(\mathbf{v})|, \tag{4}$$

due to

$$\mathbf{L} = \begin{pmatrix} 1/l_{AA} & -l_{AB}/(l_{AA}l_{BB}) & (l_{AB} - l_{BB})/(l_{AA}l_{BB}) \\ 0 & 1/l_{BB} & -1/l_{BB} \\ 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

Goldsmith and Gaylor (1970) provided the variances and the covariances of estimators of variance components $\mathbf{L}V(\mathbf{v})\mathbf{L}'$ in three stage unbalanced nested designs. Ojima (1984) demonstrated the derivation of the variances and the covariances of sums of squares based on the canonical form induced by the orthogonal transformation in three stage unbalanced nested design. 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 for the precisions 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 $\mathbf{C}$ is able to be generalized in the assumption of the nonsingular. Then, since the determinant of the variance-covariance matrix of the precision estimators is the proportional to the equation (4), the D-optimal design for the general nonsingular $\mathbf{C}$ is the same as that for the matrix (3). However, the matrix

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

cannot be made 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 in "without replication" in a certain stage. For example, If the matrix is

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

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).

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

3 *D*-Optimal Designs of Three Stage Nested Designs

The *D*-Optimal Designs in the general case where there is no restriction regarding a, b_i's, and r_{ij}'s are interesting in mathematics. However, in such a situation, the unfeasible or unrealistic designs might be picked up as the optimal designs. In collaborative experiments to determine the precision of the measurements, the a is the number of the participating laboratories, the b_i is the number of the measurement operators in the laboratory, and the 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 one-sided to be practical, in which mainly a cost problem could occurs. Hence, we consider the restricted designs that consist of the fundamental structures (d_1, d_2, d_3, d_4) shown in Figure 1.

The *D*-Optimal Design exists in the all the possible combinations of fundamental structures such that all the degrees of freedom are positive (non-zero). For each given number of observations n, such combinations are generated, and 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 $\mathcal{D} = (m_1, m_2, m_3, m_4)$, where m_i is the number of the fundamental structure d_i in the $\mathcal{D}$. Thus, the total number of observations n is equal to $4m_1 + 3m_2 + 2m_3 + m_4$. In order that it is possible to estimate all the variance components, at least either d_1 or d_2 is necessary in the $\mathcal{D}$ and two or more structures are drawn, which is $m_1 + m_2 \geq 1$ and $\sum_{i=1}^{4} m_i \geq 2$. Hence, $n = 4$ is the minimum number of observations, and there is only $\mathcal{D} = (0, 1, 0, 1)$. In case of $n = 5$, there are three possible designs which are $\mathcal{D}_1 = (1, 0, 0, 1)$, $\mathcal{D}_2 = (0, 1, 1, 0)$, and $\mathcal{D}_3 = (0, 1, 0, 2)$. Let $opt_D(D_l)$ be the value of the formula (6) for the design $\mathcal{D}_l$. We calculate $opt_D(D_1) = 55.73, opt_D(D_2) = 61.41, opt_D(D_3) = 80.58$ in $\sigma_A^2 = \sigma_B^2 = \sigma_E^2 = 1^2$, and $\mathcal{D}_1$ is identified as the optimal design with $n = 5$ under the situation where all the variances in stages are one.

Figure 2 shows the list of optimal designs for each n $(= 5, 10, 20, 30, 60)$ under the 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 nearly balanced design for are optimal in situations where $\rho_A \leq 2$. For $n = 20$, the balanced design is optimal in any situation except for $(\rho_A, \rho_B) = (8, 8)$ In the situation of $(\rho_A, \rho_B) = (8, 8)$, the staggered nested design is preferred.

The degrees of freedom for each optimal design are shown in Figure 3. The triplet in Figure 3 denotes degrees of freedom (ϕ_A, ϕ_B, ϕ_E) of the optimal design. In case of $n = 10$, though there are two different design in $\rho_A \leq 8$ and $\rho_B \leq 4$, the sum of degrees of freedom $\phi_A + \phi_B$ is the same. The ϕ_A is close to ϕ_B in all the situations, and if the ρ_A and ρ_B are larger, there is less difference among the degrees of freedom.

4 Conclusion

We obtain *D*-optimal designs for precious determination of precision of the measurement under the 49 variance component configurations. In most of optimal designs,

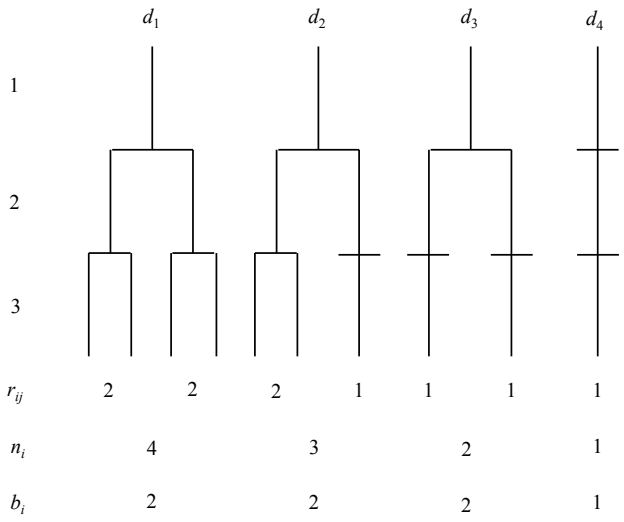


Fig. 1: Fundamental Structures

balanced fundamental structures are dominant in the wide range of the configuration. If variance components for both the first and second upper stages have so much larger than the third stage variance component, the staggered nested designs are optimal.

The optimal designs in more general unbalanced nested designs should be found and investigated in the future work. In general unbalanced cases, the number of candidate designs rapidly increase according to sample size. Combinatorial optimization would invoked to solve the problem.

References

C. H. Goldsmith and D. W. Gaylor (1970). Three Stage Nested Designs for Estimating Variance Components. *Technometrics* **12**(3), 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.

F. C. Leone, L. S. Nelson, N. L. Johnson, and S. Eisenstat (1968). Sampling Distributions of Variance Components II. Empirical Studies of Unbalanced Nested Designs. *Technometrics* **10**, 719-737.

Y. Ojima (1984)Y. The Use of Canonical Forms for Estimating Variance Components in Unbalanced Nested Designs. *Reports of Statistical Application Research*, **31**, 1-18.

		ρ_b						
	$n=5$	0.125	0.25	0.5	1	2	4	8
ρ_a	0.125	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	**(0,1,1,0)**	**(0,1,1,0)**	**(0,1,1,0)**	**(0,1,1,0)**
	0.25	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	**(0,1,1,0)**	**(0,1,1,0)**	**(0,1,1,0)**	**(0,1,1,0)**
	0.5	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	**(0,1,1,0)**	**(0,1,1,0)**	**(0,1,1,0)**	**(0,1,1,0)**
	1	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	**(0,1,1,0)**	**(0,1,1,0)**	**(0,1,1,0)**
	2	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	**(0,1,1,0)**	**(0,1,1,0)**	**(0,1,1,0)**
	4	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	**(0,1,1,0)**	**(0,1,1,0)**
	8	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	(1,0,0,1)	**(0,1,1,0)**	**(0,1,1,0)**

		ρ_b						
	$n=10$	0.125	0.25	0.5	1	2	4	8
ρ_a	0.125	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	***(0,2,2,0)***
	0.25	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	***(0,2,2,0)***
	0.5	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	***(0,2,2,0)***
	1	(2,0,0,2)	(2,0,0,2)	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	***(0,2,2,0)***
	2	(2,0,0,2)	(2,0,0,2)	(2,0,0,2)	**(2,0,1,0)**	**(2,0,1,0)**	**(2,0,1,0)**	***(0,2,2,0)***
	4	(2,0,0,2)	(2,0,0,2)	(2,0,0,2)	(2,0,0,2)	**(2,0,1,0)**	**(2,0,1,0)**	(1,2,0,0)
	8	(2,0,0,2)	(2,0,0,2)	(2,0,0,2)	(2,0,0,2)	**(2,0,1,0)**	**(2,0,1,0)**	(1,2,0,0)

		ρ_b						
	$n=20$	0.125	0.25	0.5	1	2	4	8
ρ_a	0.125	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	**(4,0,2,0)**	**(4,0,2,0)**
	0.25	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	**(4,0,2,0)**	**(4,0,2,0)**
	0.5	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	**(4,0,2,0)**	**(4,0,2,0)**
	1	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	**(4,0,2,0)**	**(4,0,2,0)**
	2	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	**(4,0,2,0)**	**(4,0,2,0)**
	4	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	**(4,0,2,0)**	*(3,2,1,0)*
	8	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	(5,0,0,0)	*(3,2,1,0)*	(0,6,1,0)

		ρ_b						
	$n=30$	0.125	0.25	0.5	1	2	4	8
ρ_a	0.125	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**
	0.25	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**
	0.5	(7,0,0,2)	(7,0,0,2)	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**
	1	(7,0,0,2)	(7,0,0,2)	(7,0,0,2)	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**
	2	(7,0,0,2)	(7,0,0,2)	(7,0,0,2)	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**
	4	(7,0,0,2)	(7,0,0,2)	(7,0,0,2)	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**
	8	(7,0,0,2)	(7,0,0,2)	(7,0,0,2)	**(7,0,1,0)**	**(7,0,1,0)**	**(7,0,1,0)**	*(0,10,0,0)*

		ρ_b						
	$n=60$	0.125	0.25	0.5	1	2	4	8
ρ_a	0.125	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)
	0.25	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)
	0.5	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)
	1	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)
	2	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)
	4	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)
	8	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	(15,0,0,0)	**(0,20,0,0)**

Fig. 2: Optimal Designs

		ρ_b						
	$n=5$	0.125	0.25	0.5	1	2	4	8
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	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**
	2	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**	**(1,2,1)**
	4	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	(1,1,2)	**(1,2,1)**	**(1,2,1)**
	8	(1,1,2)	(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	2	4	8
ρ_a	0.125	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	***(3,4,2)***
	0.25	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	***(3,4,2)***
	0.5	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	***(3,4,2)***
	1	(3,2,4)	(3,2,4)	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	***(3,4,2)***
	2	(3,2,4)	(3,2,4)	(3,2,4)	**(2,3,4)**	**(2,3,4)**	**(2,3,4)**	***(3,4,2)***
	4	(3,2,4)	(3,2,4)	(3,2,4)	(3,2,4)	**(2,3,4)**	**(2,3,4)**	(2,3,4)
	8	(3,2,4)	(3,2,4)	(3,2,4)	(3,2,4)	**(2,3,4)**	**(2,3,4)**	(2,3,4)

		ρ_b						
	$n=20$	0.125	0.25	0.5	1	2	4	8
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	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	**(5,6,8)**
	2	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	**(5,6,8)**
	4	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	(4,5,10)	**(5,6,8)**	*(5,6,8)*
	8	(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	2	4	8
ρ_a	0.125	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**
	0.25	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**
	0.5	(8,7,14)	(7,0,0,2)	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**
	1	(8,7,14)	(8,7,14)	(8,7,14)	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**
	2	(8,7,14)	(8,7,14)	(8,7,14)	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**
	4	(8,7,14)	(8,7,14)	(8,7,14)	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**	**(7,8,14)**
	8	(8,7,14)	(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	2	4	8
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	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	2	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	4	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)
	8	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	(14,15,30)	**(19,20,20)**

Fig. 3: Degrees of freedom for each optimal design

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 (1972) or more recently in Chen and Zhou (2015).
For all intents and purposes, this paper focus on three control statistics chosen by Morais and Pacheco (2015a) for their simplicity, recursive and Markovian character:

- 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.

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., 1995), in the sense that they detect any shifts in the traffic intensity sooner than they trigger a false alarm.
In this paper, we focus on the design of these type of charts for the traffic intensity of the three single server queues mentioned above.

Key words: Statistical process control, Dependent control statistics, *ARL-unbiased* charts

Manuel Cabral Morais
CEMAT & Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa, Portugal, e-mail: maj@math.ist.utl.pt

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

1 Basic facts

The first contributions on queueing theory (QT) can be traced back to three pioneering papers 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 papers 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 a *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 (2015b) led Morais and Pacheco (2015a) 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, 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),
- a general likelihood procedure (Jain, 1995).

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 (2015a) and Morais and Pacheco (2015b) 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 (2015a) for their simplicity, recursive and Markovian character. Their recursive is apparent if we note that these statistics can be rewritten as follows:

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

where

- 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) and Kendall (1953) note 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

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

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 (2015a), 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 $j \in \mathbb{N}_0$.

1.5 A few special cases

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. This certainly proves to be convenient if we want to describe in detail the run length performance of the associated control charts.

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

$$\alpha_i = \binom{k+i-1}{k-1} \left(\frac{\rho}{k+\rho}\right)^i \left(\frac{k}{k+\rho}\right)^k, \quad i \in \mathbb{N}_0 \tag{6}$$

(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 (2015a) adds that

$$\hat{\alpha}_i = \binom{k+i-1}{k-1} \left(\frac{k^{-1}}{k^{-1}+\rho}\right)^i \left(\frac{\rho}{k^{-1}+\rho}\right)^k, \quad i \in \mathbb{N}_0. \tag{7}$$

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 (2015a) wrote

$$F_{S-A}(x) = \begin{cases} \frac{\mu e^{\lambda x}}{\lambda+\mu}, & x \leq 0 \\ 1 - \frac{\lambda e^{-\mu x}}{\lambda+\mu}, & x > 0. \end{cases} \tag{8}$$

Similar calculations led Morais and Pacheco (2015a) to conclude that:

$$F_{S-A}(x) = \begin{cases} e^{\lambda x} \left(\frac{k\mu}{k\mu+\lambda}\right)^k, & x \leq 0 \\ F_{Gamma(k,k\mu)}(x) + e^{\lambda x} \left(\frac{k\mu}{k\mu+\lambda}\right)^k \bar{F}_{Gamma(k,k\mu+\lambda)}(x), & x > 0, \end{cases} \tag{9}$$

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

$$F_{S-A}(x) = \begin{cases} \bar{F}_{Gamma(k,k\lambda)}(-x) \\ \quad -e^{-\mu x}\left(\frac{k\lambda}{k\lambda+\mu}\right)^k \bar{F}_{Gamma(k,k\lambda+\mu)}(-x), & x \leq 0 \\ 1 - e^{-\mu x}\left(\frac{k\lambda}{k\lambda+\mu}\right)^k, & x > 0, \end{cases} \tag{10}$$

for the $E_k/M/1$ system. It is clear that $F_{S-A}(x)$ depends upon both λ and μ, for all these three queueing systems, therefore the entries of $\tilde{\mathbf{P}}$ will not depend exclusively on ρ like $\mathbf{P}$ and $\hat{\mathbf{P}}$.

1.6 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:

$$\sigma = \tilde{F}_A[\mu(1-\sigma)] = \int_{a=0}^{+\infty} e^{-\mu(1-\sigma)a} dF_A(t) \tag{11}$$

(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

$$E(z^X) = \frac{(1-\rho)\,\tilde{F}_S[\lambda(1-z)]\,(1-z)}{\tilde{F}_S[\lambda(1-z)] - z}, \; |z| \leq 1, \tag{12}$$

where $\tilde{F}_S(t) = \int_{s=0}^{+\infty} e^{-ts} dF_S(s)$ 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

$$E(z^X) = \frac{1-\rho}{(1-z/z_1)(1-z/z_2)}, |z| \leq 1, \tag{13}$$

$$P(X=i) = (1-\rho)\left[\frac{z_1}{z_1-z_2}\left(\frac{1}{z_2}\right)^n - \frac{z_2}{z_1-z_2}\left(\frac{1}{z_1}\right)^n\right], i \in \mathbb{N}_0, \tag{14}$$

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.

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 (2015a) 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

$$RL = \min\{n \in \mathbb{N} : Z_n > U \mid Z_0 = 0\}, \tag{15}$$

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 (2015a), 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

$$\begin{bmatrix} \mathbf{Q} & (\mathbf{I}-\mathbf{Q})\,\underline{\mathbf{1}} \\ \underline{\mathbf{0}}^{\top} & 1 \end{bmatrix}, \tag{16}$$

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

$$ARL^0=\underline{\mathbf{e}}_0^{\top}\times(\mathbf{I}-\mathbf{Q})^{-1}\times\underline{\mathbf{1}}, \tag{17}$$

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 (2016a), 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), Ramalhoto and Morais (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), Knoth and Morais (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 (2016a) to derive an ARL-unbiased $np-$chart, and by Morais (2016b) 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 $M/E_k/1$ $(k = 1, 2, 3, 4, 5, 10, 15, \infty)$, $\rho_0 = 0.1(0.1)0.9$, 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

$$\begin{bmatrix} p_{L\,L} \times (1-\gamma_L) & p_{L\,L+1} & \cdots & p_{L\,U-1} & p_{L\,U} \times (1-\gamma_U) \\ p_{L+1\,L} \times (1-\gamma_L) & p_{L+1\,L+1} & \cdots & p_{L+1\,U-1} & p_{L+1\,U} \times (1-\gamma_U) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{U-1\,L} \times (1-\gamma_L) & p_{U-1\,L+1} & \cdots & p_{U-1\,U-1} & p_{U-1\,U} \times (1-\gamma_U) \\ p_{U\,L} \times (1-\gamma_L) & p_{U\,L+1} & \cdots & p_{U\,U-1} & p_{U\,U} \times (1-\gamma_U) \end{bmatrix}. \tag{18}$$

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 paper 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

$$\tilde{\mathbf{Q}}(\gamma_L) = \begin{bmatrix} \tilde{p}_{L\,L} \times (1-\gamma_L) & \tilde{p}_{L\,L+1} & \cdots & \tilde{p}_{L\,U-1} & \tilde{p}_{L\,U} \\ \tilde{p}_{L+1\,L} \times (1-\gamma_L) & \tilde{p}_{L+1\,L+1} & \cdots & \tilde{p}_{L+1\,U-1} & \tilde{p}_{L+1\,U} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{p}_{U-1\,L} \times (1-\gamma_L) & \tilde{p}_{U-1\,L+1} & \cdots & \tilde{p}_{U-1\,U-1} & \tilde{p}_{U-1\,U} \\ \tilde{p}_{U\,L} \times (1-\gamma_L) & \tilde{p}_{U\,L+1} & \cdots & \tilde{p}_{U\,U-1} & \tilde{p}_{U\,U} \end{bmatrix}, \tag{19}$$

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).

[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.

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 1–4 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).

The corresponding ARL-profiles can be found in Figures 1–4, for $\rho_0 = 0.1, 0.5$, 0.9 and $ARL^\star = 500$. The profiles in Figures 1 and 2 (resp. Figures 3 and 4) were obtained using the Markov chain approach (resp. the collocation method, with the exception the last three in Figure 4; the Markov chain approach was used instead, with $(250 + 1)$ transient states).

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.

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.

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.

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

We believe that this last property is in part responsible for the apparent similarity of ARL profiles in Figure 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 Figure 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.

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), Paulino et al. (2016b) and Morais (2016a,b).

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

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

takes a zero value. This follows from the need to achieve a 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.

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 Figure 2.

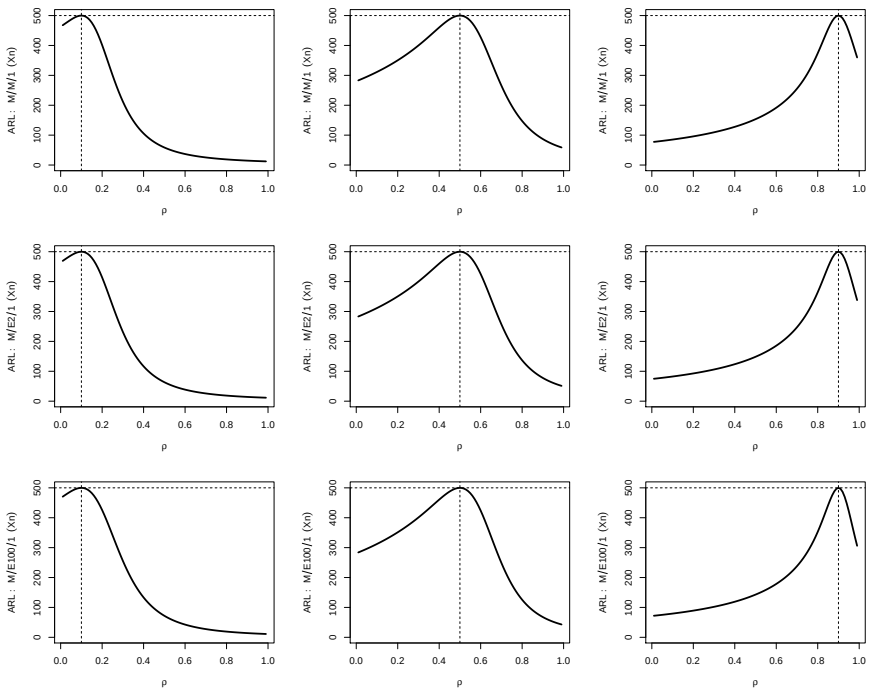


Fig. 1: ARL profiles of the ARL-unbiased X_n-chart — $M/M/1$, $M/E_2/1$ and $M/E_{100}/1$ systems with $\rho_0 = 0.1, 0.5, 0.9$.

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 Figure 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 Figures 1 and 2.

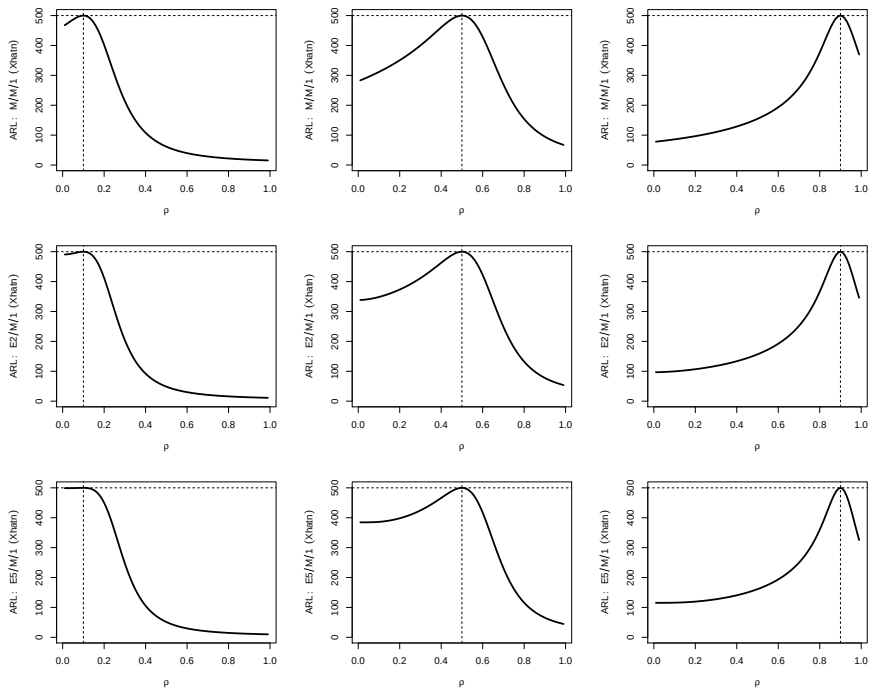


Fig. 2: ARL profiles of the ARL-unbiased $\hat{X}_n-$chart — $M/M/1$, $E_2/M/1$ and $E_5/M/1$ systems with $\rho_0 = 0.1, 0.5, 0.9$.

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 figures 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;

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	500.001	499.923
	0.2	5.563324	0.002287	499.715	500.004	499.700
	0.3	6.471000	0.002556	499.325	500.018	499.267
	0.4	7.538378	0.002921	498.656	500.057	498.489
	0.5	8.863481	0.003433	497.519	500.141	497.091
	0.6	10.604397	0.004186	495.507	500.287	494.476
	0.7	13.058999	0.005393	491.640	500.463	489.260
	0.8	16.890344	0.007621	483.295	500.541	478.126
	0.9	24.001537	0.013020	462.104	500.351	454.016
$E_2/M/1$	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	11.375281	0.003272	496.263	499.998	495.498
	0.7	13.693802	0.004144	492.552	499.981	490.670
	0.8	17.357791	0.005776	484.458	499.861	480.021
	0.9	24.234818	0.009750	463.359	499.176	455.744

- 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 Figure 3;
- when λ is fixed, the ARL profiles, in Figure 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 Figure 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;

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	499.457	500.005	499.316
	0.2	1.333646	0.002447	498.749	500.020	498.394
	0.3	2.263006	0.002760	497.807	500.060	497.118
	0.4	3.437682	0.003171	496.511	500.139	495.293
	0.5	4.957663	0.003733	494.626	500.274	492.561
	0.6	6.999476	0.004552	491.651	500.457	488.201
	0.7	9.904858	0.005855	486.438	500.627	480.699
	0.8	14.440533	0.008257	476.086	500.640	466.788
	0.9	22.823032	0.014126	451.630	500.372	440.513
$E_2/M/1$	0.1	0.807873	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	3.993489	0.002631	497.298	500.000	496.538
	0.5	5.554376	0.003021	495.473	500.000	494.093
	0.6	7.602164	0.003604	492.532	499.995	490.061
	0.7	10.471460	0.004549	487.399	499.964	482.958
	0.8	14.911878	0.006309	477.243	499.786	469.420
	0.9	23.098971	0.010632	452.652	498.926	442.630

- by comparing the ARL profiles in figures 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.

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,

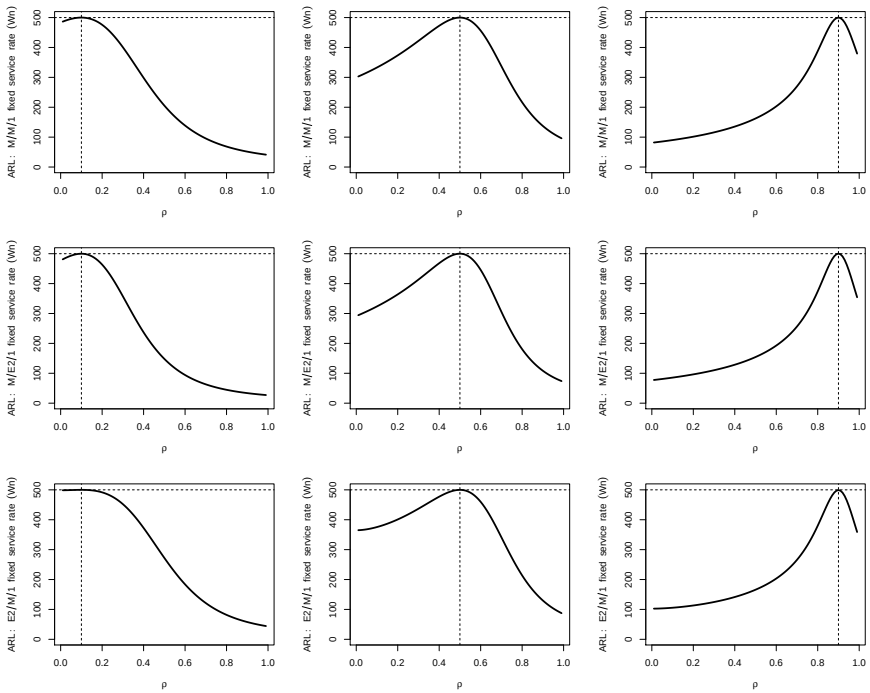


Fig. 3: ARL profiles of the ARL-unbiased W_n-chart, FIXED SERVICE RATE — $M/M/1$, $M/E_2/1$ and $E_2/M/1$ systems with $\rho_0 = 0.1, 0.5, 0.9$.

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 Figure 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.

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 (2015a), 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

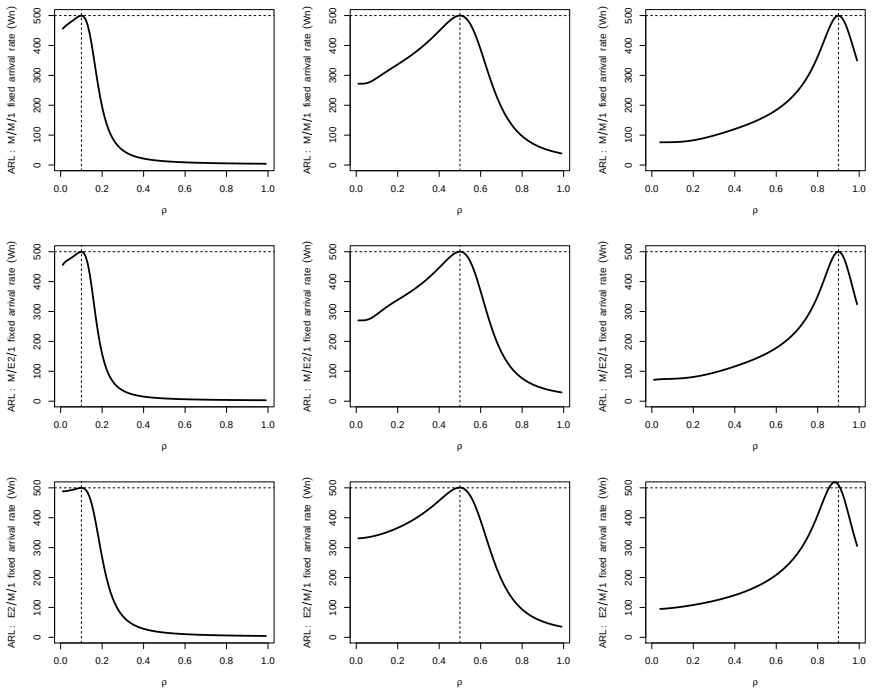


Fig. 4: ARL profiles of the ARL-unbiased W_n-chart, FIXED ARRIVAL RATE — $M/M/1$, $M/E_2/1$ and $E_2/M/1$ systems with $\rho_0 = 0.1, 0.5, 0.9$.

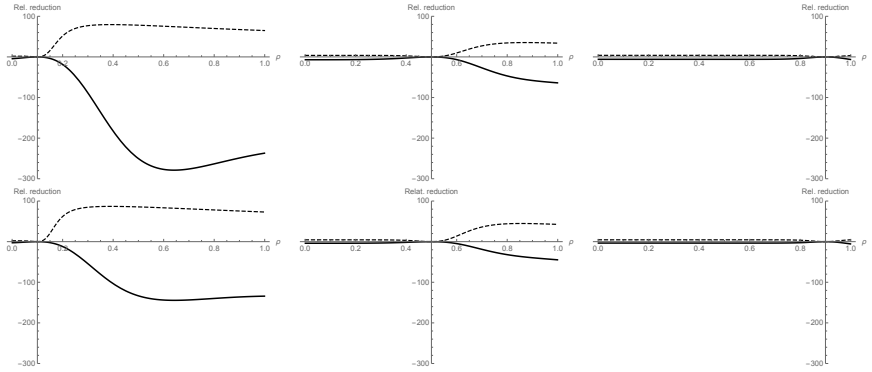


Fig. 5: Plots of the relative ARL reduction, $[ARL_{W_n}(\rho)/ARL_{X_n}(\rho) - 1] \times 100\%$ — $M/M/1$ (top) and $M/E_2/1$ (bottom) systems with $\rho_0 = 0.1, 0.5, 0.9$ and $ARL^\star = 500$; fixed service (resp. arrival) rate corresponds to the solid (resp. dashed) lines.

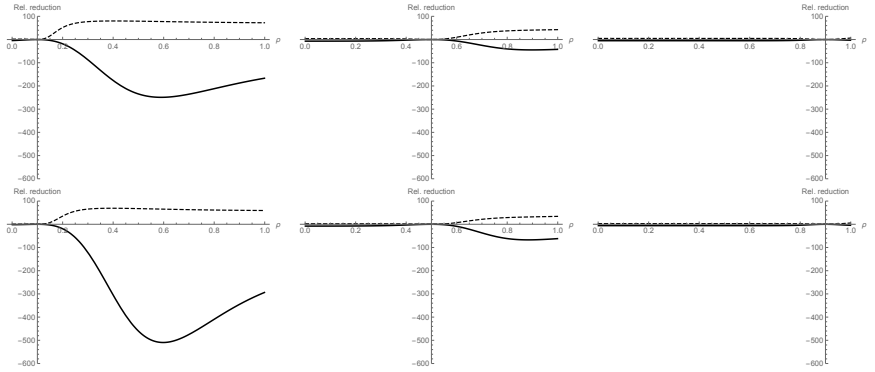


Fig. 6: Plots of the relative ARL reduction, $[ARL_{W_n}(\rho)/ARL_{\hat{X}_n}(\rho)-1]\times 100\%$ — $M/M/1$ (top) and $E_2/M/1$ (bottom) systems with $\rho_0 = 0.1, 0.5, 0.9$ and $ARL^{\star} = 500$; fixed service (resp. arrival) rate corresponds to the solid (resp. dashed) lines.

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 3 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 discrete control statistics.

4 Final thoughts

The aim of this paper is two-fold.

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 (reps. 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*, *nL* and 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 We 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 and for the fruitful phone conversations; 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).

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.

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. Accessed from http://www.win.tue.nl/~iadan/queueing.pdf on 2016-05-27.

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.

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.

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.

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 Publishing Company.

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, NC: 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. edition). 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. (1995). Estimating changes in traffic intensity for M/M/1 queueing systems. *Microelectronics Reliability, 35*, 1395–1400.

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.

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 International Publishing.

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. edition). New York: John Wiley & Sons.

Morais, M. C. (2016a). An ARL-unbiased np-chart. *Economic Quality Control, 31* 11–21.

Morais, M. C. (2016b). ARL-unbiased geometric and CCC_G control charts. Submitted for publication in *International Journal of Production Research*.

Morais, M. C., & Pacheco, A. (1998). Comparing first passage times of Markovian processes. Proceedings of the 2nd. International Symposium on Semi-Markov Models: Theory and Applications. Compiègne, France, Dec. 9–11, 1998.

Morais, M. C., & Pacheco, A. (2015a). On stochastic ordering and control charts for the traffic intensity. Submitted for publication in *Sequential Analysis*.

Morais, M. C., & Pacheco, A. (2015b). On control charts and the detection of increases in the traffic intensity. Unpublished manuscript, 44 pages.

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 (wtrnumber2012-3), Department of Mathematics, Washington State University.

Paulino, S., Morais, M. C., & Knoth, S. (2016a). An ARL-unbiased c-chart. *Quality and Reliability Engineering International* http://onlinelibrary.wiley.com/doi/10.1002/qre.1969/epdf.

Paulino, S., Morais, M. C. & Knoth, S. (2016b). On ARL-unbiased c-charts for INAR(1) Poisson counts. Submitted for publication in *Statistical Papers*.

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. In preparation.

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*. Western Electrical Corporation, Indianopolis, IN.

Wolfram Research, Inc. (2015). *Mathematica, Version 10.3*. Champaign, Illinois. Accessed from http://reference.wolfram.com/language/ on 2016-03-31.

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. Accessed from http://www.hindawi.com/journals/mpe/2013/241241/ on 2015-11-11.

Statistical process monitoring of multivariate time-between-events data: Problems and possible solutions

Chenglong Li, Amitava Mukherjee, Qin Su, and Min Xie

Abstract In the recent years, a lot of attention is paid to univariate monitoring of time-between-events. When the univariate problem is extended to a multivariate situation, this study indicates that there is an emerging issue about the presence of asynchronous observations. This brings the new statistical challenge due to the way of data acquisition changes. This study also tries to give some possible solving ideas for the monitoring and analysis of multivariate TBE data stream.

Key words: asynchronous observation; multivariate exponential distribution; statistical process monitoring; time-between-events

1 Introduction

Monitoring the time between certain consecutive events in a system is an important research problem in the field of statistical process monitoring. In the recent years, host of researchers have proposed a variety of Time-Between-Events (TBE) monitoring schemes. This has become an increasingly popular research area in the dawning years of the twenty-first century. Indeed, the term 'system' here could refer to various processes, equipment or any entity of interest in a particular context, and as a result, 'event' may have different meanings in different contexts. For example, in the context

Chenglong Li · Qin Su
School of Management, Xi'an Jiaotong University, Xi'an, China,

Amitava Mukherjee
Production Operations and Decision Sciences Area, XLRI-Xavier School of Management, Jamshedpur, India, e-mail: amitmukh2@xlri.ac.in

Min Xie · Chenglong Li
Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China, e-mail: minxie@um.cityu.edu.hk, e-mail: lchenglon2-c@my.cityu.edu.hk

of manufacturing, an event may refer to the occurrence of a nonconforming unit. In the context of reliability testing, an event may indicate the failure of a component. The occurrence of various events often results in a negative way and some have hazardous consequences.

For univariate monitoring of time-between-events, after each failure (or each occurrence of a nonconforming unit), the time interval, i.e., $T_1, T_2, \ldots$ is recorded and can be plotted on the chart. If the plotted point falls outside the calculated control limits, it indicates an out-of-control situation. Then, the out-of-control action plan should be implemented. Figure 1 displays three time intervals T_1, T_2, and T_3 between the failures of a simple single-unit system that occurred at t_1, t_2, and t_3, respectively. To monitor a complex system, we often need to take into account multiple aspects

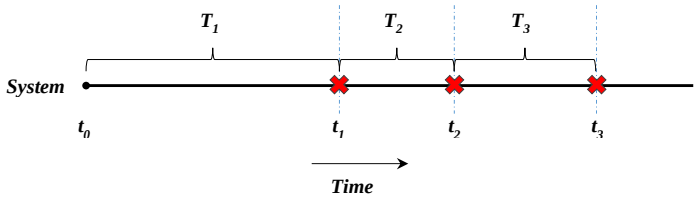


Fig. 1: The occurrences of the failures of a single-unit system.

of the system for an overall evaluation of its underlying operation condition. Depending upon different situations, these aspects could refer to concrete objects (e.g., interrelated subsystems, units, or components) and also could be immaterial such as operating characteristics, features or behaviors. When a system is in the normal state of operation, failures (of each subsystem or component) or abnormities (in one or more operating characteristics or behaviors) are random event caused by, for example, occasional sudden increase of stress and human error (Xie *et al*. 2002). The occurrences of various discrete events of concern are often modeled by a multivariate Poisson process. Then, the time interval $\boldsymbol{T} = (x_1, x_2, \ldots, x_p)$ between these events follows a multivariate exponential distribution (abbreviated MVE), writing $\boldsymbol{T} \sim \text{MVE}(\boldsymbol{\lambda})$ as a p-dimensional random variable. The most common form was given by Marshall and Olkin (1967) as follows,

$$F(x_1, x_2, \ldots, x_p) = \exp\left[-\sum_{i=1}^{p} \lambda_i x_i - \sum_{i<j} \lambda_{ij} \max(x_i, x_j) \right.$$
$$\left. - \sum_{i<j<k} \lambda_{ijk} \max(x_i, x_j, x_k) - \ldots - \lambda_{12\ldots p} \max(x_1, x_2, \ldots, x_p) \right].$$

Or we have a more compact notation expressed as

$$F(x_1, x_2, \ldots, x_p) = \exp\left[-\sum_{s \in S} \lambda_s \max(x_i s_i)\right]$$

where S denotes the set of vectors $(s_1, \ldots, s_p)$ and each $s_j = 0$ or 1 but $(s_1, \ldots, s_p) \neq (0, \ldots, 0)$. For any vector $s \in S$, $\max(x_i s_i)$ is the maximum of the x_i's for which $s_i = 1$.

Analogous to the univariate TBE monitoring problem taken into account in Xie *et al*. (2002), Zhang *et al*. (2005), Khoo and Xie (2009), Qu *et al*. (2014), among others, in this context of monitoring a complex system with multivariate TBE data, the primary aim is to detect any change of the parameter $\boldsymbol{\lambda}$. Other than those traditional studies on multivariate statistical process control (see for example, Zou and Tsung (2011), Aparisi *et al*. (2012), Wang and Reynolds (2013), Li *et al*. (2014)), when monitoring multivariate TBE data, an emerging issue is about the presence of asynchronous observations (i.e., the time intervals) with regard to various dimensions. That is, the observations of various dimensions will not be generated at the same time, because the occurrences of various events (e.g., the signals of failure of various units) are not synchronous. This issue will be addressed in more detail later in this study. Accordingly, the challenge of statistical process monitoring of multivariate TBE data not only comes from the difficulty of studying a multivariate non-normal distribution (rather than the commonly studied multivariate normal distribution), but also the complexity due to the way of data acquisition changes. Although people still can apply the traditional approach of combining several univariate TBE control charts to each individual dimension, it is known that this will lead to erroneous conclusions, even be misleading and inefficient. Therefore, how to realize the monitoring and analysis of such multivariate data streams is a real-life problem of significance, since such data are common in all walks of life but have been rarely considered in the control chart literature. Motivated by the actuality, this study will also attempt to provide some solving ideas.

2 Problem description

Consider a complex system that has p interrelated key units. For example, one can imagine a running computer server as the system and the p key units are such as CPU, GPU, RAM, and HD etc. Failures may happen to any unit and when a signal of failure occurs, we can promptly locate and know which unit is disabled, then fix or replace it. The respective time interval $\boldsymbol{T}$ between the failures to each unit follows MVE($\boldsymbol{\lambda}$). When the system operates in a perfect condition, the distribution of $\boldsymbol{T}$ remains stable with the in-control rate parameter $\boldsymbol{\lambda}_0$. In this context, it can be assumed that the in-control $\boldsymbol{\lambda}_0$ is known a priori. However, the system could be disturbed by the unobservable occurrence of some assignable cause at some random time, causing variations in the rate parameter (especially the increasing shifts, as one is usually more interested in detecting a decrease in the time interval).

Then, the statistical monitoring problem becomes similar to testing the hypothesis,

$$H_0 : \lambda_i = \lambda_{0,i} \text{ for all } i \quad \text{against} \quad H_1 : \lambda_i \geq \lambda_{0,i} \text{ with at least one strict inequality.}$$

To implement a statistical process monitoring procedure, we need to collect the TBE observations over time. As stated earlier, the desired time-to-failure observations with regard to various units are not generated simultaneously, such that each time we probably could observe only a single failure on one of those units, as shown in Figure 2. When we observe a failure of the ith unit, we have an exact observation (i.e., exact surviving time) for this unit. At the same time, however, we equivalently have additional $(p-1)$ censored observations for other rest $(p-1)$ units, because the rest $(p-1)$ units are still running. For example, for the three-unit system displayed in Figure 2 when we observe a failure of Unit 2 at t_3, we immediately have the latest time-to-failure observation T_{22} for Unit 2. But Unit 1 and Unit 3 do not fail at t_3, such that at this moment we do not have the matched time-to-failure observations for them. We just have the most recent time-to-failure observation T_{11} for Unit 1 which was, however, generated $(t_3 - t_1)$ ago. In addition, the two time intervals, $(t_3 - t_1)$ and $(t_3 - t_0)$ can be treated as the right-censored observations for Unit 1 and Unit 3, respectively. If the system has more than three units, the analysis procedure is compatible but more cumbersome.

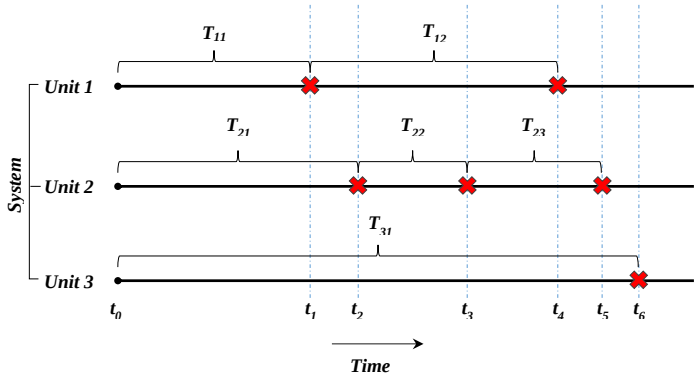


Fig. 2: The occurrences of the failures of the three units of a system ($p = 3$).

Note that, if the context is changed to the manufacturing process, the object of study turns into a high-yield process that produces a kind of products with p key quality characteristics. Nonconformities may occur in any quality characteristic of the products in unpredictable ways and the respective time interval $\boldsymbol{T}$ required to observe the nonconformity of various characteristics approximatively follows MVE($\boldsymbol{\lambda}$).

3 Possible solving ideas

General guidelines are offered below for the design of a feasible solution to realize the statistical monitoring and analysis of multivariate TBE data.

Step-1: Record the time interval for the failed unit each time when one failure is observed, and estimate the rest $(p-1)$ *observations.*

Step-2: Fuse the exact and estimated observations into a single value in a manner for statistical monitoring purpose.

Step-3: Plot this data point against the calculated control limit and the horizontal axis of the chart is the accumulated failure number in total.

Step-4: If the point falls outside the limit, the system is declared out of control. The operation of the system is immediately interrupted and an action plan is implemented. Then, the system is restored, and go back to Step-1.

Step-5: If the point falls within the limit, the system is thought to be in control, and the monitoring continues to the next failure. Then, go back to Step-1.

Remark:

As the acquiring of these time-to-failure observations of various units is asynchronous, each time when one failure is observed, we can get this fresh time-to-failure observation, but do not have other $(p-1)$ matched time-to-failure observations with regard to the rest $(p-1)$ units that are still running at this moment. It is necessary to estimate the rest $(p-1)$ observations based on the knowledge we have, that is, the historical time-to-failure observations of these units as well as the $(p-1)$ censored observations. Probably the most straightforward but inefficient way is to use the most recent time-to-failure observations of each unit. Of course, using weighted time intervals will improve the estimation and a much better choice is to adopt the Bayesian inference. However, in this way, the corresponding formulation and calculation will be much more complicated.

Once the rest $(p-1)$ observations are estimated by using some of the tips above, we need to properly fuse these 'observations' from different units in a manner for evaluating the underlying system state. At this point, some classical multivariate statistical process monitoring techniques can be adapted to the monitoring of multivariate TBE data. These observations can also be fused through Bayesian theory to give a posterior probabilistic estimate of the underlying system state.

After Step-1 and Step-2, all that remains is the common operational manipulation of a control chart. There is one difference to note here: the horizontal axis of the chart represents the total failure number, not the sample number.

4 Conclusions

This study indicates the statistical challenge of monitoring multivariate TBE data. Some general guidelines for possible solving ideas are outlined as well.

References

Aparisi, F., Epprecht, E. K., Ruiz, O. (2012). T^2 control charts with variable dimension. *Journal of Quality Technology* **44**(4), 375-393.

Khoo, M. B. C., Xie, M. (2009). A study of time-between-events control chart for the monitoring of regularly maintained systems. *Quality and Reliability Engineering International* **25**(7), 805-819.

Li, J., Tsung, F., Zou, C. (2014). Multivariate binomial/multinomial control chart. *IIE Transactions* **46**(5), 526-542.

Marshall, A. W., Olkin, I. (1967). A multivariate exponential distribution. *Journal of the American Statistical Association* **62**(317), 30-44.

Qu, L., Wu, Z., Khoo, M. B. C., Rahim, A. (2014). Time-between-event control charts for sampling inspection. *Technometrics* **56**(3), 336-346.

Wang, S., Reynolds, M. R., Jr. (2013). A GLR control chart for monitoring the mean vector of a multivariate normal process. *Journal of Quality Technology* **45**(1), 18-33.

Xie, M., Goh, T. N., Ranjan, P. (2002). Some effective control chart procedures for reliability monitoring. *Reliability Engineering & System Safety* **77**(2), 143-150.

Zhang, C. W., Xie, M., Goh, T. N. (2005). Economic design of exponential charts for time between events monitoring. *International Journal of Production Research* **43**(23), 5019-5032.

Zou, C., Tsung, F. (2011). A multivariate sign EWMA control chart. *Technometrics* **53**(1), 84-97.

Integrating Statistical and Machine Learning Approaches in Improving Inspection Process

Tomomichi Suzuki, Tatsuya Iwasawa, Kenta Yoshida, Natsuki Sano, Mirai Tanaka

Abstract Nowadays, many products are manufactured in larger quantities and at higher speed. Inspection processes also need to be operated at higher speed, without loss of accuracy at detecting nonconforming products. Regarding inspection of external appearances of the products, visual inspections have often been used in many processes, which many of them are now replaced by automatic inspections using sensors such as cameras. In this study, statistical tools and machine learning methods are applied to improve accuracy of an actual automatic inspection process.

1 Introduction

Nowadays, many products are manufactured in larger quantities and at higher speed. Inspection processes also need to be operated at higher speed, without loss of accuracy at detecting nonconforming products. Regarding inspection of external appear-

Tomomichi Suzuki
Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: szk@rs.tus.ac.jp

Tatsuya Iwasawa
Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: 7416607@ed.tus.ac.jp

Kenta Yoshida
Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: 7416644@ed.tus.ac.jp

Natsuki Sano
Faculty of Economic, Management and Information Science, Onomichi City University, Onomichi, Hiroshima, 722-8506, Japan, e-mail: sano@onomichi-u.ac.jp

Mirai Tanaka
Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: mirai@rs.tus.ac.jp

ances of the products, visual inspections have often been used in many processes, which many of them are now replaced by automatic inspections using sensors such as cameras. In this study, statistical tools and machine learning methods are applied to improve accuracy of an actual automatic inspection process.

2 Product and Data

2.1 Data

The product taken up in this study is a cylindrical metal product. The external appearance is inspected automatically using the images taken by cameras installed in the process. The characteristics (RGB values) of the original images are converted into polar coordinates because the images are taken from above the product and the nonconforming defects tend to appear concentrically. This concept is shown in Figure 1.

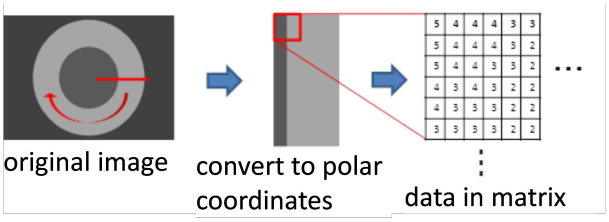


Fig. 1: Data used in analysis

Regarding accuracy of the actual inspection process, the probability of type II error was satisfactory but the probability of type I error needed improvement. In other words, sensitivity was satisfactory but specificity was not. Since there will be a need to detect nonconforming products which is more difficult to detect, the objective of the study is not limited to reduce type I errors but also reduce type II errors, in other words propose algorithms that improves overall accuracy of the inspection process.

2.2 Defects

Detection of defects are not performed by directly analyzing rows and columns (i, j) of the matrix data but by analyzing the waveform data. To detect defects, we also

examine the characteristic waveforms. Figure 2 shows the waveform for an average product height (i.e., the average of a row of matrix data).

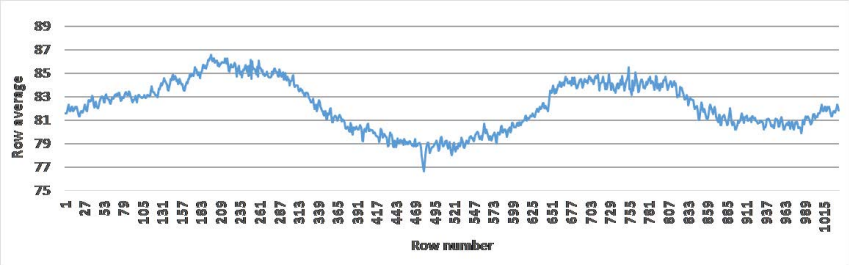


Fig. 2: Waveform data from sample product

Figure 2 shows a relatively large sample of a given defect, which contains high and low points. Near row 470, the waveform suddenly dips because of defects. Defects are betrayed not by the mean maximum or minimum but by sudden changes in the numerical values. Thus, detecting small defects is difficult because they cause small changes in these numerical values.

We define a non-defective product as a product that has very few or no defects such as dents or blots. To investigate the accuracy of the inspection system, we prepared artificial defective products by attaching a very small colored seal which resembles the defects. The size of the seal is in three levels and there are four colors.

2.3 Extracting features and creating variables

One of the most important parts of this study is to extract features from the original images and to create variables which will be used in later analyses. Many aspects of the images need to be accommodated such as; the first data of each row is next to the last data of the same row because the data are in polar coordinates, the position of the product changes slightly among them each time they are being imaged, etc.

Some image processing techniques such as Gaussian filters and Laplacian filters are used. The data for each column are considered as waveform data so that techniques for time series analysis can be applied to extract the features. Quite a number of variables were created as candidates for future analyses.

3 Selection of variables by statistical approach

3.1 Application of design of experiments

As mentioned earlier, detecting small defects is difficult. Facilitating their detection requires some type of preprocessing after polar conversion. However, we neither know to what extent each method affects accuracy nor how to combine the different methods to improve accuracy. We therefore introduce orthogonal arrays to test the effect of many factors via a few experiments.

We use an L32 orthogonal array, run 32 experiments for each area of the product, and calculate three accuracies from each experiment. The overall accuracy of each experiment is defined to be the average of the three accuracies. We convert these results by using the logit function and use them as characteristic values. After analyzing the variance, we determine the optimum level and the optimum combination of methods for each product area.

The main effects and interactions considered are shown in Table 1. Previous studies have shown that "feature quantity extraction with window" (FQEW) and number of lags are especially effective in improving accuracy, so we insert four-level FQEWs and lags into the orthogonal array. The first-level factor is "Used" and second level is "Not used." If the first level is selected, then we use this method to calculate the accuracy. If the second level is selected, we do not use this method to calculate the accuracy.

Table 1: Factors investigated.

Symbol	Factor	Levels
A	Overlap	Use / Not Use
B	Nonlinear density conversion	Use / Not Use
C	Gaussian filter	Use / Not Use
D	Laplacian filter	Use / Not Use
H	Maximum value	Use / Not Use
P	Minimum value	Use / Not Use
J	Variance	Use / Not Use
K	Kurtosis	Use / Not Use
L	Skewness	Use / Not Use
M	Range	Use / Not Use
N	EHOG feature quantity	Use / Not Use
O	Texture characteristic quantity	Use / Not Use
F	Statistic from a row	Variance / Maximum
E	FQEW (size of window)	10x1 / 20x1 / 10x4 / 20x4
G	Number of lags	0 / 1 / 3 / 5

In this study, we calculate the accuracy by using SVM (support vector machine). To analyze how to combine preprocessing and feature-extraction methods, we use a distinction method, which dispenses with variable selection. We therefore introduce

the SVM. The preprocessing methods, feature-extraction methods, and SVM are explained below.

3.2 Results of DOE

The factors significant in the analysis of the data in Area 1were B, C, E, F, G, H, K, M, and O. The estimate of the accuracy was 85.7% with the confidence interval {83.2%, 87.9%}. The factors significant in the analysis of the data in Area 2 were E, F, G, H, K, M, and N. The estimate of the accuracy was 85.6% with the confidence interval {81.4%, 89.0%}. The factors significant in the analysis of the data in Area 3 were D, E, F, G, P, K, and M. The estimate of the accuracy was 90.8% with the confidence interval {88.5%, 92.7%}. Note that the results of the optimum combination of methods differs among areas.

Area1 gives a noisy and dark image because it is near the bottom of the product and far from the camera. To reduce noise and create a brighter image, we apply a Gaussian filter and nonlinear density conversion. The matrix of data of Area1 has 21 columns, whereas that of Area2 has 14 columns. Defects are sufficiently detectable from waveform data created by using small windows in the FQEW method. Conversely, the matrix of data for Area3 contains over twice as many columns as do the matrices of areas1 and 2. Thus, defects are not sufficiently detectable when waveform data are created using a 10x1 window in the FQEW method. However, if the window size is too large, the noise increases in Area1. Thus, a 20x1 window is optimal for Area3. The Laplacian filter is used in the optimum combination of methods for Area3 because defects are not sufficiently detectable within the waveform data when using only the FQEW method.

These results showed considerable improvement over the current system.

4 Defect detection by machine learning approach

This section concerns the classifiers applied in this study: Bagging, AdaBoost, and Random Forest. Regarding Bagging and AdaBoost, we applied decision trees (DTs) and neural networks (NNs) as weak classifiers. The details of these classifiers are described below.

4.1 Bagging

Bagging, along with AdaBoost and Random Forest described below, is one of the methods of so-called Ensemble classification, algorithms that provide classifiers that have essentially high generalization ability by combining multiple classifiers that lack

such ability (Hirai (2012)). Bagging is an ensemble classification method in which the outputs of test data of a class are determined by majority vote of multiple weak classifiers, after letting these weak classifiers learn by using bootstrap samples of training data. The two classes in this study are non-defective and defective products. The structure of Bagging is shown in Figure 3. To provide further details of ensemble classification, DT is generally applied as a weak classifier.

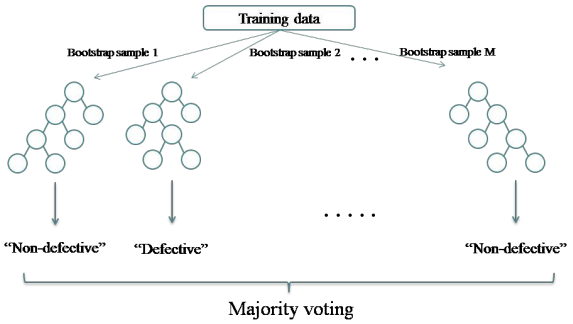


Fig. 3: Structure of bagging

4.2 AdaBoost

AdaBoost is one of the Boosting methods, whose algorithms train multiple serially cascaded weak classifier's, with each weak classifier trained individually. In AdaBoost, heaviness is assigned to training data according to the training result of a weak classifier, and as a result of minimizing the heaviness of missed training data, the later the weak classifier trains, the more it concentrates on training data that are missed many times.

4.3 Random forest

The Random Forest method improves Bagging by applying DTs as weak classifiers and can create a large variety of DTs in which each correlation coefficient is not high, by selecting a definite number of features at random used in classification at non-terminal nodes. The DTs in Random Forest increase at the dividing point, and the reduction of the Gini coefficient in selected features is maximized. We designed preprocessing and feature extracting processes based on variable importance from Random Forest, which can be evaluated by calculating Average for the reduction of the Gini coefficient of each feature in each node of all the weak classifiers.

4.4 Discussion of weak classifiers

DTs are generally applied as weak classifiers in Bagging and AdaBoost. In this study, we aimed to construct classifiers that have high generalization ability by using classifiers other than DTs as weak classifiers. When using the new weak classifiers, we selected logistic regression and neural networks in which the number of hidden units is only one as its candidates so that products can be inspected as quickly as possible. From the results of discriminating products exclusively of DTs, logistic regression, or neural networks, we decided to construct Bagging- and AdaBoost-applied neural networks as weak classifiers. Similarly, Sano (2003) constructed AdaBoost-applied neural networks as weak classifiers, and verified its precision.

4.5 Results of defect detection

In this section, we compare and discuss the discrimination result of each area, applying five classifiers: Random Forest, Bagging and AdaBoost-applied DTs as weak classifiers, Bagging and AdaBoost-applied neural networks in which the number of hidden units is only one, and using the features extracted in preprocessing and feature extraction. In addition, we conducted three-fold cross-validation to evaluate their generalization ability. The discrimination results of each area are shown in Figures 4, 5 and 6. The values with each point refer to the correct answer rate.

From the figures, the correct answer rate of AdaBoost (NN) is the highest in Areas 1 and 2. In Area 3, the correct answer rates of Random Forest and Bagging (NN) are equivalent, but the beta error of Bagging (NN) is slightly smaller than that of Random Forest. Therefore, we evaluated Bagging (NN) as the classifier that can discriminate products most precisely. Regardless of the kind of weak classifier, we found that AdaBoost can discriminate products more precisely than Bagging in Areas 1 and 2; on the other hand, we could not find whether Bagging or AdaBoost is a better classifier for discriminating products, because the correct answer rate of AdaBoost for the case of applying DT as a weak classifier is better than that of Bagging. However, the correct answer rate of Bagging for the case of applying NN as a weak classifier is better than that of AdaBoost. Here are some reasons to consider. In Area 3, there are fewer data than in Areas 1 and 2, because of the narrowness of the analysis range. In addition, it is difficult to see the defective parts, as a result of the distance from the analyzed part to the camera, with the result that the defective parts are photographed smaller in Area 3. We could make the most of AdaBoost's algorithm by reasoning that the training errors achieve relatively large value because of the difficulty of detecting defective parts in Areas 1 and 2; however, the training errors in Area 3 area are very low, and hence, there is no basis for choosing between Bagging and AdaBoost. In addition, the discrimination precision in Area 2 is the worst in this study, because on the manufacturing line, the products are discriminated one by one at high speed and are irradiated by green light to verify the location of

the analysis part that appears just in the product interior in Area 2, which is caused by misclassification.

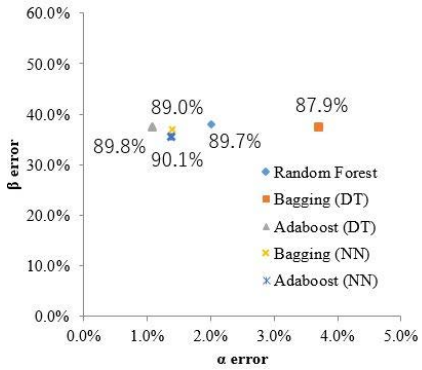


Fig. 4: Results of Area 1

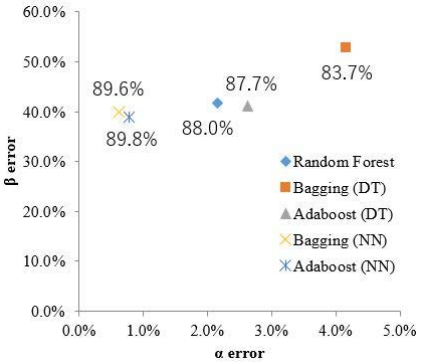


Fig. 5: Results of Area 2

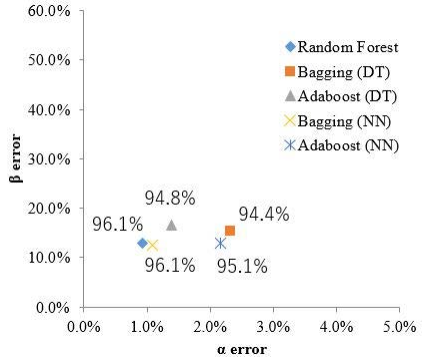


Fig. 6: Results of Area 3

5 Summary

Combination of statistical and machine learning approches are applied in order to improve an insepection process. In variable selection stage, orthogonal arrays which is one of the popular tools in design of experiments are used. In defect detection stage, classification methods of machine learning techniques are used. Results showed considerable improvement over the current system.

References

Boser, B.E., Guyon, I.M., Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Hirai, Y. (2012). *Introduction to Pattern Recognition*, Morikita Press.

Sano, N. (2003). Implementation of AdaBoost by Splus *https://www.msi.co.jp/splus/events/student/2003pdf/sano.pdf*. Last visted 2016-06-18. (in Japanese).

Takagi, M., and Shimoda, Y. (2004). *Handbook of Image Analysis*, University of Tokyo Press.

A MGF Based Approximation to Cumulative Exposure Models

Watalu Yamamoto and Lu Jin

Abstract Online monitoring data contains various measurements of the activity of the system. The amounts of works are also measured in various ways. When we model the reliability of a system, the intensity or the 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 be able to provide evidence for them. There are many uncontrollable factors in the field. Many variables are monotone increasing and highly correlated with each other within a system. Yet they also represent the differences among systems. This article tries to build a bridge between two useful approaches, alternative time scale (Kordonsky and Gertsbakh (1997), Duchesne and Lawless (2002)) and cumulative exposure model (Hong and Meeker (2013)), by assuming the stationarity of the increments of these measurements within a system.

Key words: Time-scale, cumulative exposure model, accelerated lifetime model, approximation, moment generating function.

1 Time Scales

When there are more than one variables, $T_0, T_1, T_2, \ldots, T_p$, to measure the lifetime of a system or a product, the problem of time scale identification arises. 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. Oakes (1995) defined the notion of collapsibility of the time scale and proposed a parametric

Watalu Yamamoto,
University of Electro-Communications, Japan, e-mail: watalu@inf.uec.ac.jp

Lu Jin,
University of Electro-Communications, Japan, e-mail: jinlu@inf.uec.ac.jp

inference for choosing time scales and failure distributions. This problem was also investigated by Kordonsky and Gertsbakh (1993, 1995a,b, 1997). They consider so called the linear time scale model,

$$U_L = \beta_0 T_0 + \beta_1 T_1 + \cdots + \beta_p T_p,$$

and investigate properties by estimating parameters with minimum coefficient of variation. The choice of the estimating criterion was made because it is scale invariant. In their studies, the parameter space is restricted as

$$\Theta_L = \left\{ \boldsymbol{\beta}; \beta_k \leq 0, k = 0, \ldots, p \text{ and } \sum_k \beta_k = 1 \right\}. \tag{1}$$

There are also another class of time scale,

$$U_M = T_0^{\beta_0} T_1^{\beta_1} \cdots T_p^{\beta_p}.$$

This is called the multiplicative time scale model. This is actually a log linear time scale model in that

$$T_0^{\beta_0} T_1^{\beta_1} \cdots T_p^{\beta_p} = \exp\left\{ \beta_0 \tilde{T}_0 + \beta_1 \tilde{T}_1 + \cdots + \beta_p \tilde{T}_p \right\} \tag{2}$$

where $\tilde{T}_k = \log T_k$, $k = 0, 1, \ldots, p$. Duchesne and Lawless (2000) called these models as alternative time scales. Duchesne and Lawless (2002) proposed a semiparametric approach to estimate the parameters of time scale models under the assumption of collapsibility proposed by Oakes (1995). These models are well served for selecting the best time scale model when the measurements are done only when products are failed or censored.

Recently there is literature on the assessing the reliability of products under continuous on-line surveillance. Hong and Meeker (2013) propose to hire Nelson's cumulative damage model to model the effect of use rate variation onto the lifetime of a product and predict the lifetime distribution by estimating the use rate process. Hong, Duan, Meeker, Stanley, and Gu (2015) model a physical degradation process using the dynamic measurements of the environmental conditions. They apply a smoothing regression technique to estimate the trends of degradation paths. We believe that the cumulative damage model is also useful for the problem of time scales.

2 Cumulative Exposure Time Scales

A general cumulative damage model is specified by a pair of formulas, the cumulative damage,

$$U(T)\,|\mathcal{H}_\infty \sim \int_0^T \mathcal{D}(s;\mathcal{H}_s)\,ds \tag{3}$$

and the distribution on the cumulative damage scale,

$$U(T) \sim F(u). \tag{4}$$

$\mathcal{H}_t$ is the history up to time t,

$$\mathcal{H}_t = \{N(s), 0 \le s < t; \boldsymbol{x}(s), 0 \le s \le t\} \tag{5}$$

where $N(t)$ is the counting process and $\boldsymbol{x}(s)$ is the covariate process which are the measurements on the conditions of the product and/or around the product. Generally the damage at time t, $\mathcal{D}(t;\mathcal{H}_t)$ could depend on the history up to time t, $\mathcal{H}_t$. However it is difficult to model in such a flexible way. So we restrict ourselves to model lifetime data with continuous monitoring as

$$\mathcal{D}(s;\mathcal{H}_s) \approx \mathcal{D}(s;\boldsymbol{x}(s)). \tag{6}$$

We believe this is a version of collapsibility for time scale modeling under continuous surveillance.

For the simplicity of the argument, we focus on "no-covariate-situation." All covariates are assumed to be the increments of individual time scale varaibles. It is also assumed that $X_0(t) \equiv 1$ for all products and for any t. So $\int_0^T X_0(s)\,ds$ is the lifetime on the chronological time scale. Since the problem of interest here is the identification of the best time scale, all $X_k(t)$ have marginal distributions with

$$\mathrm{E}[X_k(s)] = 1. \tag{7}$$

Under these assumtions, $\mathcal{D}(s;\boldsymbol{x}(s))$ can assess whether the variations of any $x_k(t)$ makes the lifetime longer or shorter.

There are two primary choices of parameteric models. One is the linear model,

$$\mathcal{D}_L(t;\boldsymbol{x}(t)) = \beta_0 x_0(t) + \beta_1 x_1(t) + \cdots + \beta_p x_p(t). \tag{8}$$

This model is derived from a general model by approximating with Taylor expansion around $\boldsymbol{x} = \mathbf{1}$,

$$\mathcal{D}_L(t;\boldsymbol{x}(t)) \approx \mathcal{D}_L(t_0;\mathbf{1}) + \sum_k \left.\frac{\partial \mathcal{D}}{\partial x_k}\right|_{x_k=1} (x_k - 1). \tag{9}$$

This model coincides with the linear time scale model. But there is one difference. The parameter space need not be positive.

Another is the multiplicative model,

$$\mathcal{D}_M(t;\boldsymbol{x}(t)) = \exp\left(\beta_0 \tilde{x}_0(t) + \beta_1 \tilde{x}_1(t) + \cdots + \beta_p \tilde{x}_p(t)\right), \tag{10}$$

where $\tilde{x}_k(t) = \log x_k(t)$. This model is derived from a general model by approximating the logarithm with Taylor expansion around $\boldsymbol{x} = \mathbf{1}$,

$$\log \mathcal{D}_M(t; \boldsymbol{x}(t)) \approx \log \mathcal{D}_L(t_0; \mathbf{1}) + \sum_k \left.\frac{\partial \mathcal{D}}{\partial \tilde{x}_k}\right|_{\tilde{x}_k=0} (\tilde{x}_k). \tag{11}$$

Unlike the first model, this s not the same as the multiplicative time scale model.

3 Parameter Estimation

We assume that the online monitoring system collects the sample path of its covariate process $\mathcal{X}_{i,\infty}$, the time of event t_i and the type of event δ_i, from each system to be monitored. $\delta_i = 1$ indicates that the system is failed and $\delta_i = 0$ for censored. The contribution of this system to the log-likelihood is

$$\begin{aligned}\log L_i = {} & \delta_i \boldsymbol{\beta}' \boldsymbol{x}_i(t_i) \\ & + \delta_i \log g\left(U\left(t_i; \boldsymbol{\beta} | \mathcal{X}_{i,\infty}\right); \boldsymbol{\theta}\right) \\ & + (1-\delta_i) \log\left\{1 - G\left(U\left(t_i; \boldsymbol{\beta} | \mathcal{X}_{i,\infty}\right); \boldsymbol{\theta}\right)\right\}.\end{aligned}$$

where $\boldsymbol{\beta}' \boldsymbol{x}_i(t_i)$ is $\log \partial U\left(t; \boldsymbol{\beta} | \mathcal{X}_{i,\infty}\right) / \partial t$ evaluated at $t = t_i$. Hereafter we abbreviate $U_C\left(t_i; \boldsymbol{\beta} | \mathcal{X}_{i,\infty}\right)$ as U_i.

The score vector consists of

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log L_i = {} & \delta_i \boldsymbol{x}_i(t_i) + \delta_i \frac{\partial U\left(t_i; \boldsymbol{\beta} | \mathcal{X}_{i,\infty}\right)}{\partial \boldsymbol{\beta}} \left.\frac{\partial}{\partial u} \log g(u; \boldsymbol{\theta})\right|_{u=U_i} \\ & + (1-\delta_i) \frac{\partial U\left(t_i; \boldsymbol{\beta} | \mathcal{X}_{i,\infty}\right)}{\partial \boldsymbol{\beta}} \left.\frac{\partial}{\partial u} \log\{1 - G(u; \boldsymbol{\theta})\}\right|_{u=U_i}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \log L_i = {} & \delta_i \frac{\partial}{\partial \boldsymbol{\theta}} \log g(U_i; \boldsymbol{\theta}) \\ & + (1-\delta_i) \frac{\partial}{\partial \boldsymbol{\theta}} \log\{1 - G(U_i; \boldsymbol{\theta})\}.\end{aligned}$$

The observed Fisher information matrix consists of

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\log L_i = \delta_i \frac{\partial^2 U\left(t_i;\boldsymbol{\beta}\,|\mathcal{X}_{i,\infty}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\,\frac{\partial}{\partial u}\log g\left(u;\boldsymbol{\theta}\right)\Big|_{u=U_i}$$
$$+(1-\delta_i)\,\frac{\partial^2 U\left(t_i;\boldsymbol{\beta}\,|\mathcal{X}_{i,\infty}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\,\frac{\partial}{\partial u}\log\left\{1-G\left(u;\boldsymbol{\theta}\right)\right\}\Big|_{u=U_i}.$$
$$+\delta_i \frac{\partial U\left(t_i;\boldsymbol{\beta}\,|\mathcal{X}_{i,\infty}\right)}{\partial\boldsymbol{\beta}}\,\frac{\partial}{\partial\boldsymbol{\beta}^t}\left[\frac{\partial}{\partial u}\log g\left(u;\boldsymbol{\theta}\right)\Big|_{u=U_i}\right]$$
$$+\delta_i \frac{\partial U\left(t_i;\boldsymbol{\beta}\,|\mathcal{X}_{i,\infty}\right)}{\partial\boldsymbol{\beta}}\,\frac{\partial}{\partial\boldsymbol{\beta}^t}\left[\frac{\partial}{\partial u}\frac{\partial}{\partial\boldsymbol{\theta}}\log\left\{1-G\left(u;\boldsymbol{\theta}\right)\right\}\Big|_{u=U_i}\right]$$
$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L_i = \delta_i \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log g\left(U_i;\boldsymbol{\theta}\right)$$
$$+(1-\delta_i)\,\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log\left\{1-G\left(U_i;\boldsymbol{\theta}\right)\right\}$$

and also off-diagonal components

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\beta}^T}\log L_i = \delta_i \frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{\partial}{\partial u}\log g\left(u;\boldsymbol{\theta}\right)\Big|_{u=U_i}\right]\frac{\partial U\left(t_i;\boldsymbol{\beta}\,|\mathcal{X}_{i,\infty}\right)}{\partial\boldsymbol{\beta}^T}$$
$$+(1-\delta_i)\,\frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{\partial}{\partial u}\log\left\{1-G\left(u;\boldsymbol{\theta}\right)\right\}\Big|_{u=U_i}\right]\frac{\partial U\left(t_i;\boldsymbol{\beta}\,|\mathcal{X}_{i,\infty}\right)}{\partial\boldsymbol{\beta}^T} \tag{12}$$

The first derivatives and second derivatives with respect to $\boldsymbol{\theta}$ are readily available on many packages or software which help us in fitting parametric lifetime distributions to the failure data with censoring. However the derivatives with respect to the components of parameter $\boldsymbol{\beta}$ requires numerical integration for each system, every time we need to evaluate.

4 Cumulative Exposure Model and Empirical Moment Generating Function

We would like to allow cumulative exposure model by Hong and Meeker (2013) to be more useful for applications. To achieve this goal, we restrict our attention to cases with covariates related to the works of the systems and also the time variables only.

We focus on the integral processes of work amounts among many types of covariates. If the covariate process $\mathcal{X}_{i,\infty}$ is stationary,

$$U\left(t;\boldsymbol{\beta}\,|\mathcal{X}_{i,\infty}\right)/t = \frac{1}{t}\int_0^t \exp\left(\boldsymbol{\beta}^T\boldsymbol{x}\left(s\right)\right)ds \tag{13}$$

is a nonparametric estimate of the joint moment generating function of the marginal distributions of $\boldsymbol{X}(t)$,

$$M_{\boldsymbol{X}}(\boldsymbol{\beta}) = E\left[\exp\left(\boldsymbol{\beta}^T \boldsymbol{X}(t)\right)\right]. \tag{14}$$

Under certain regularity conditions for the existence of the moment generating function, this estimate, also called as *the empirical moment generating function*, is the consistent estimate of the underlying moment generating function.

Once the empirical moment generating function of the covariate process is estimated as $\hat{M}_{\boldsymbol{X}}(\boldsymbol{\beta})$, we could approximate the cumulative exposure as

$$U(t;\boldsymbol{\beta}|\mathcal{X}_{i,\infty}) \simeq \hat{M}_{\boldsymbol{X}}(\beta)\, t. \tag{15}$$

This approximation also establishes the relationship between the cumulative exposure model and the accelerated lifetime 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 simultaneous distributions. 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. Then the estimates of the first two moments, $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ are plugged into the joint moment generating function and have

$$\hat{M}_{\boldsymbol{X}}(\beta) = \exp\left(\hat{\boldsymbol{\mu}}^T \boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^T \hat{\Sigma}\boldsymbol{\beta}\right). \tag{16}$$

We 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 covariate processes and the integral processes of which need no be positive or monotone for their purposes.

5 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 same as that for the evaluation of $U(t;\boldsymbol{\beta}|\mathcal{X}_{i,\infty})$. The estimation of the cumulative exposure model needs the evaluation of this function for each individual product within the online monitoring data. If we want to monitor the changes in fitting of the model regularly, the total amount of computation for this model grows at every moment we receive a new record. So it is very useful to invent the decreasing in the amount of computation.

The simplest way is Taylor series approximation of the moment generating function. If the moment generating function exists, it has the Taylor series expansion

$$M_X(\boldsymbol{\beta}) = 1 + \boldsymbol{\beta}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\beta}^T \left(\boldsymbol{\mu}\boldsymbol{\mu}^T + \Sigma\right)\boldsymbol{\beta} + \cdots \tag{17}$$

around the origin of the space of $\boldsymbol{\beta}$. The first order approximation of the moment generating function is

$$\tilde{M}_1(\boldsymbol{\beta}) = 1 + \boldsymbol{\beta}^T \boldsymbol{\mu}. \tag{18}$$

A moment estimator of $\boldsymbol{\mu}$ is the vector of sample means of $x_{ik}(t)$'s. This approximation holds under the first order stationarity where the expected values of covariates do not depend on time, i.e. $\mathrm{e}\boldsymbol{X}_i(t) = \boldsymbol{\mu}_i$.

The second order approximation gives another formula

$$\tilde{M}_2(\boldsymbol{\beta}) = 1 + \boldsymbol{\beta}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\beta}^T \left(\boldsymbol{\mu}\boldsymbol{\mu}^T + \Sigma\right)\boldsymbol{\beta}. \tag{19}$$

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 normal distribution could be considered.

$$\tilde{M}_G(\boldsymbol{\beta}) = \exp\left(\boldsymbol{\beta}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}\right) \tag{20}$$

If the covariates are conditionally independent with 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. So we have another identification by

$$\tilde{M}_{\boldsymbol{X}}(\boldsymbol{\beta}) = \prod_j \tilde{M}_{X_j}\left(\beta_j\right). \tag{21}$$

There are also other ways of approximations. We state 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 pre-specified points $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, the multilinear interpolation is obtained as

$$\tilde{M}_L(\boldsymbol{\beta}) = \sum_k N_k \hat{M}_{\boldsymbol{X}}(\boldsymbol{\beta}_k), \tag{22}$$

where N_k is the normalizing quantity which depends on both $\boldsymbol{\beta}$ and the set of points $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$.

Another way is to have the random set of points $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ and construct multidimensional spline interpolation by multiple adaptive regression splines (Friedman (1991)) or generalized additive models (Hastie and Tibshirani (2004)). We note that though there are many flexible and useful interpolation techniques, they tend to re-increase the amount of computation.

Acknowledgments

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.

References

Duchesne, T., and Lawless, J. F. (2000). Alternative time scales and failure time models. *Lifetime Data Analysis*, 6, 157-179.

Duchesne, T., and Lawless, J. F. (2002). Semiparametric inference method for general time scale models. *Lifetime Data Analysis*, 8, 263-276.

Farewell, V. T., and Cox, D. R. (1975). A note on multiple time scales in life testing. *Applied Statistics*, 28, 115-124.

Finkelstein, M. S. (1999). Wearing-out of components in a variable environment. *Reliability Engineering and System Safety*, 66, 235-242.

Finkelstein, M. S. (2004). Minimum repair in heterogeneous populations. *Journal of Applied Probability*, 41, 281-286.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19, 1-67.

Hastie, T. and Tibshirani, R. (1986) Generalized Additive Models. *Statistical Science*, 1, 297-310.

Hong, Y., and Meeker, W. Q. (2013). Field-failure predictions based on failure-time data with dynamic covariate information. *Technometrics*, 55, 135–149.

Hong, Y., Duan, Y., Meeker, W. Q., Stanley, D. L. , and Gu, X. (2015). Statistical Methods for Degradation Data With Dynamic Covariates Information and an Application to Outdoor Weathering Data. *Technometrics*, 57, 180–193.

Kordonsky, K. B. and Gertsbakh, I. (1993). Choice of the best time scale for system reliability analysis. *European Journal of Operational Research*, 65, 235-246.

Kordonsky, K. B. and Gertsbakh, I. (1995a). System state monitoring and lifetime scales - I. *Reliability Engineering and System Safety*, 47, 1-14.

Kordonsky, K. B. and Gertsbakh, I. (1995b). System state monitoring and lifetime scales - II. *Reliability Engineering and System Safety*, 49, 145-154.

Kordonsky, K. B. and Gertsbakh, I. (1997). Multiple time scales and the lifetime coefficient of variation: engineering applications. *Lifetime Data Analysis*, 2, 139-156.

Lawless, J. F., Crowder, M. J., and Lee, K.-A. (2009). Analysis of reliability and warranty claims in products with age and usage scales. *Technometrics*, 51, 14–24.

Oakes, D. (1995). Multiple time scales in survival analysis. *Lifetime Data Analysis*, 1, 7-18.

New results for two-sided CUSUM-Shewhart control charts

Sven Knoth

Abstract Already Yashchin (1985b), and of course Lucas (1982) three years earlier, studied CUSUM chart supplemented by Shewhart limits. Interestingly, Yashchin (1985b) 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 are slightly complicated numerical tasks. Similarly to Capizzi and Masarotto (2010) we deploy less common numerical techniques (Clenshaw-Curtis quadrature, collocation) to determine the ARL and other RL based measures. 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.

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 signal 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 et al. (1977) introduced it into statistical process control (SPC) literature. For an application in clinical chemistry, they proposed Shewhart-CUSUM combinations. However, their two-sided CUSUM chart is not the well-known pair of two one-sided schemes. It resembles a CUSUM pheno-

Sven Knoth
Institute of Mathematics and Statistics, Department of Economics and Social Sciences, Helmut Schmidt University Hamburg, Postfach 700822, 22008 Hamburg, Germany, e-mail: Sven.Knoth@hsu-hh.de

type 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 provide, 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 charts. Both authors discussed as well one-sided designs. 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 utilized 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 calculation of the ARL was not questioned after its first treatment in Lucas (1982) and Lucas (1982). 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 analyzed. Examples are provided, technical details moved into the Appendix, and some conclusions complete the paper.

2 One-sided CUSUM-Shewhart chart for mean

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.

$$\begin{aligned}
\text{Shewhart rule} \quad \ell_S &= \inf\{i \geq 1 : X_i > c_S\}\,. \\
Z_0 = z_0 = 0,\ Z_i &= \max\{0, Z_{i-1} + X_i - k\}, \\
\text{CUSUM rule} \quad \ell_c &= \inf\{i \geq 1 : Z_i > h\}\,. \\
\text{combo rule} \quad \ell &= \min\{\ell_S, \ell_c\}\,. \\
\text{ARL} \quad &= E_\mu(\ell)\,. \\
\text{ARL function} \quad \mathcal{L}(z) &= E_\mu(\ell \mid z_0 = z)\,.
\end{aligned}$$

The terms ARL and ARL function label the well-known Average Run Length both universally 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 is 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

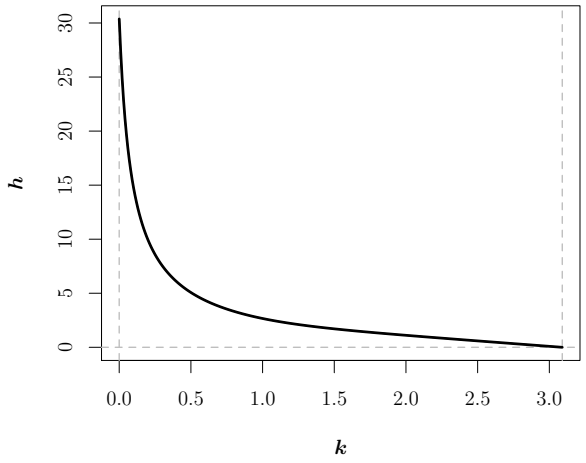


Fig. 1: CUSUM setup: Relationship between reference value k and threshold h for an in-control ARL 1000. Admissible k values belong to the interval $(0, \Phi^{-1}(1 - 1/1000) = 3.09)$.

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 resembles a single CUSUM chart. Namely, any 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, Figure 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,

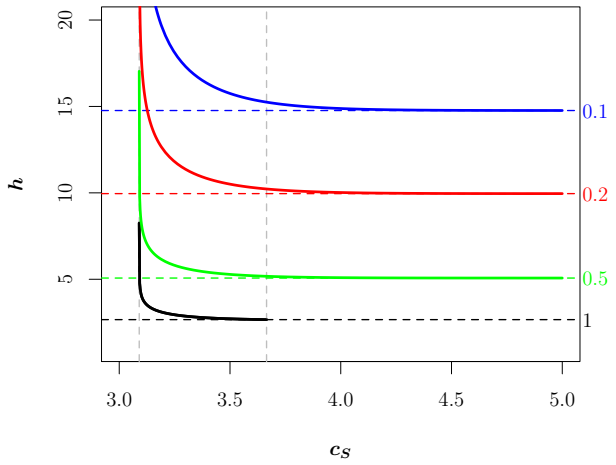


Fig. 2: Combinations of Shewhart threshold c_S and CUSUM's h ($k \in \{1, 0.5, 0.2, 0.1\}$) for an overall in-control ARL 1000.

the normal quantile $\Phi^{-1}(1-1/A)$, the upper one by the threshold h of a standalone CUSUM chart increased by k:

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

In the sequel we assume that (1) is fulfilled. From Figure 2 we see that for small $k < 1$, the interval could be even more tightened, 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:

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

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), 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. Here, we want to exercise collocation with piecewise defined Chebyshev polynomials — see Knoth (2006) for their successful usage 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.

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

The integer r is determined from $r = \lceil h/\varepsilon \rceil = \lceil h/(c_S - k) \rceil$. From Figure 3 one

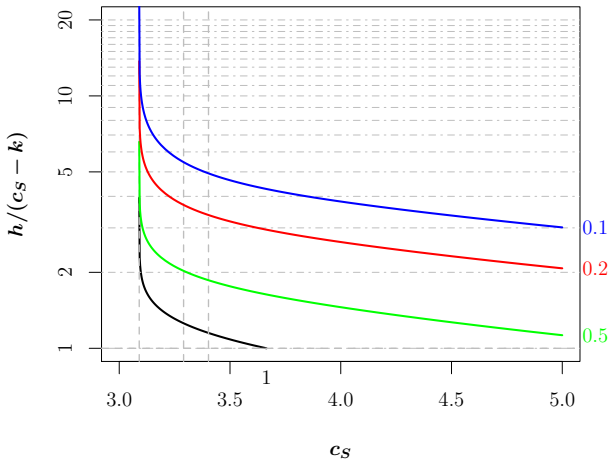


Fig. 3: The number of intervals, $r = \lceil h/(c_S - k) \rceil$, for combinations of Shewhart threshold c_S, CUSUM's h ($k \in \{1, 0.5, 0.2, 0.1\}$), and in-control ARL 1000.

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 given in the Appendix. Essentially, we distinguish for $\mathcal{L}(s)$ the intervals $0 < s \le h - \varepsilon$ or $h - \varepsilon < s \le h$. 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

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

For $\mathcal{L}(0)$ we could use

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

For the two intervals we obtain

$0 < s \le h - \varepsilon$:

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

$h - \varepsilon < s \le h$:

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

As already mentioned, we derive

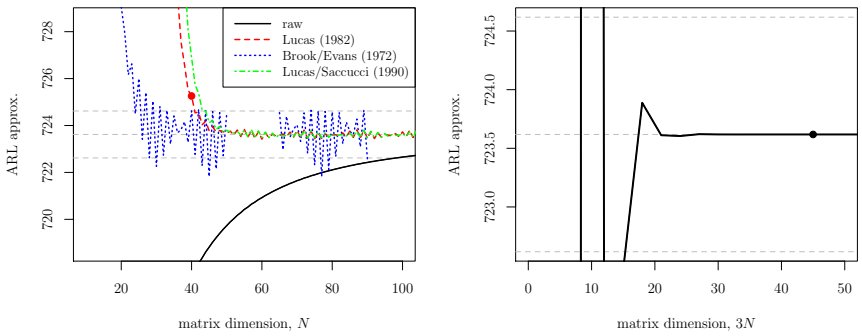
$$\mathcal{L}(0) = \sum_{j=1}^{N} c_{1j} T_{1j}(0) = \sum_{j=1}^{N} c_{1j} (-1)^{j+1}$$

so that, eventually, a linear equation system with dimension $2N$ has to be solved.

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 resulted in 1509.94 with s.e. 0.048), which differs considerably from the value from Yashchin (1985b) in the table printed as Figure 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 $\varepsilon = 3.75$ and $r = 3$ intervals. In Figure 4 the related ARL approximations are plotted versus matrix dimension. In Figure 4(a) we display besides the "raw" Markov chain



(a) Markov chain (b) Collocation

Fig. 4: (In-control) ARL approximation vs. matrix dimension; $k = 0.25, h = 8, c_S = 4$.

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 combine them by the same linear model. For the sake of visibility, we omit some segments for the highly varying profile following Brook and Evans (1972). From Figure 4(a) 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 Figure 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 in the results into a simple regression model, he obtained the final results which surprisingly well match the collocation based numbers.

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 percent 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 Figure 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

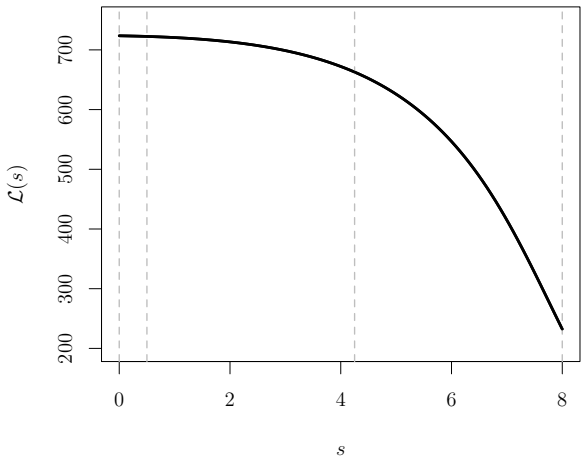
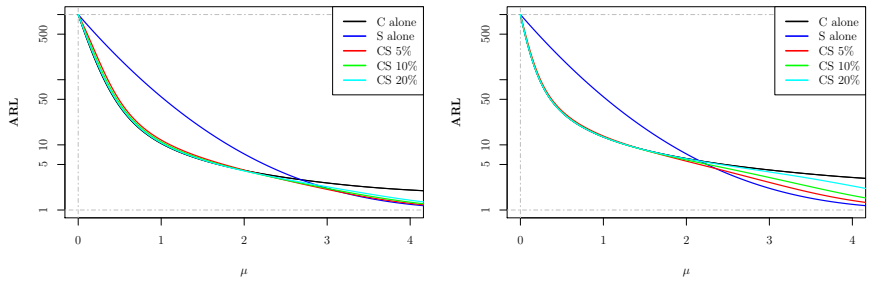


Fig. 5: Complete function $\mathcal{L}(s)$ for $k = 0.25, h = 8, c_S = 4, \varepsilon = 3.75, r = 3$ intervals.



(a) $k = 0.5$ (b) $k = 0.2$

Fig. 6: ARL performance of different Shewhart/CUSUM combos and standalone charts.

detection performance for changes larger than 2.5. In summary, it looks like a handy improvement of the prim CUSUM procedure.

3 Two-sided case

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

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
			202.1	48.19	20.16	11.94	8.344	5.062	3.461	2.471	1.820
6	0.25	3	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
			241.7	50.79	20.77	12.29	8.640	5.384	3.852	2.910	2.239
6	0.25	3.5	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
			249.7	51.27	20.89	12.36	8.713	5.487	4.013	3.143	2.525
6	0.25	4	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
			395.8	73.98	27.04	15.42	10.58	6.194	4.045	2.733	1.912
8	0.25	3	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
			646.4	82.12	28.44	16.17	11.20	6.835	4.760	3.451	2.511
8	0.25	3.5	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
			725.2	83.75	28.72	16.34	11.36	7.051	5.082	3.888	3.010
8	0.25	4	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
			571.3	101.7	33.68	18.75	12.68	7.202	4.509	2.903	1.955
10	0.25	3	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
			1441	119.9	36.09	20.00	13.71	8.222	5.584	3.897	2.706
10	0.25	3.5	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
			1974	124.0	36.62	20.31	13.99	8.596	6.119	4.586	3.445
10	0.25	4	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

$$
\begin{aligned}
\text{Shewhart rule}\quad & \ell_S^{(2)} = \inf\{i \geq 1 : |X_i| > c_S\}\,. \\
& Z_0^+ = z_0^+ = 0,\ Z_n^+ = \max\{0, Z_{n-1}^+ + X_n - k\}, \\
\text{upper CUSUM rule}\quad & \ell_c^+ = \inf\{n \geq 1 : Z_n^+ > h\}\,. \\
& Z_0^- = z_0^- = 0,\ Z_n^- = \max\{0, Z_{n-1}^- - X_n - k\}, \\
\text{lower CUSUM rule}\quad & \ell_c^- = \inf\{n \geq 1 : Z_n^- > h\}\,. \\
\text{2-sided CUSUM rule}\quad & \ell_c^{(2)} = \min\{\ell_c^+, \ell_c^-\}\,. \\
\text{combo rule}\quad & \ell^{(2)} = \min\{\ell_S^{(2)}, \ell_c^{(2)}\}\,.
\end{aligned}
$$

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.

Now, 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 common, the values of the CUSUM statistic:

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

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

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

Hence, for $s^+ + s^- \le 2k$, 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

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

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

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

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 $s^+ + s^- \le h + 2k$, 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 $s^+ + s^- \le h + 2k$ — hence (i) and (ii), but not (iii) — to link $\mathcal{L}(s^+, s^-)$ to the ARL function of the simpler one-sided CUSUM chart

$$\mathcal{L}(s^+, s^-) = \frac{\mathcal{L}^+(s^+)\mathcal{L}^-(0) + \mathcal{L}^+(0)\mathcal{L}^-(s^-) - \mathcal{L}^+(0)\mathcal{L}^-(0)}{\mathcal{L}^+(0) + \mathcal{L}^-(0)} . \tag{3}$$

It turns out that it solves the integral equation (i)+(ii). 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, $\max\{-c_S, lower\}$ and $\min\{c_S, upper\}$ 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 (3) would work? Starting with (i) and re-writing the corresponding integral equation results in:

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

From (3) we derive:

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

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

Impact to first line:

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

If we substitute $x = z + k - s$ in (2), then we obtain

$$\mathcal{L}(s) = 1+\Phi(k-s)\mathcal{L}(0)+\int_{k-s}^{\min\{h+k-s,c_S\}}\varphi(x)\mathcal{L}(s+x-k)\,dx, \tag{4}$$

so that the line under analysis simplifies heavily to

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

and accordingly the second line to

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

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)

$$\frac{\mathcal{L}^+(s^+)\mathcal{L}^-(0)+\mathcal{L}^-(s^-)\mathcal{L}^+(0)-\mathcal{L}^+(0)\mathcal{L}^-(0)}{\mathcal{L}^+(0)+\mathcal{L}^-(0)} \quad \text{which confirms (3).}$$

Similar ideas are used for case (ii) which is described in the appendix. Essentially, (3) remains valid.

Hence, the zero-state ARL of a two-sided Shewhart-CUSUM 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 (3).

3.1 Examples for two-sided designs

Again we start with a result from Yashchin (1985b). We re-collect some numbers from Yashchin's Figure 6 and new results in Table 2. Both, the results by Yashchin

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

(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. 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 (3) differ to the Monte Carlo derived results. All significant (5 % level) differences are marked with bold letters. Of course, the differences are really small. But it was claimed that our new approach is highly accurate. It is less surprising that the accuracy problems vanish for increasing Shewhart limit c_S, because the combo

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
			99.05	45.51	19.86	11.81	8.244	4.974	3.382	2.419	1.789
6	0.25	3	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
			121.6	49.78	20.79	12.31	8.667	5.419	3.901	2.977	2.319
6	0.25	3.5	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
			124.8	50.22	20.86	12.35	8.704	5.474	3.990	3.105	2.473
6	0.25	4	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
			188.9	68.01	26.08	14.96	10.24	5.933	3.855	2.640	1.893
8	0.25	3	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
			325.4	81.65	28.51	16.23	11.25	6.894	4.829	3.520	2.567
8	0.25	3.5	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
			361.4	83.27	28.69	16.32	11.34	7.021	5.034	3.820	2.942
8	0.25	4	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
			301.5	101.9	34.92	19.49	13.25	7.628	4.797	3.032	1.987
10	0.25	3	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
			704.1	117.2	35.72	19.78	13.49	7.958	5.263	3.585	2.511
10	0.25	3.5	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
			975.5	124.4	36.73	20.36	14.03	8.651	6.224	4.759	3.682
10	0.25	4	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

becomes more similar to a pure CUSUM scheme, where the validity of (3) is well

established. 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

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

presented method provides quite good approximations, but they do not attain the traditional accuracy of ARL integral equation related methods. Some "root cause analysis" should be done to identify the actual reason for the deviations and to develop some guide lines, when the simple (3) is really valid.

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. After some necessary hardening of the implementation, we will apply these new algorithms to identify useful CUSUM-Shewhart setups for control charting practice.

5 Appendix

5.1 Collocation design for more than $r = 2$ intervals

$s = 0$:

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

$0 < s \le h - \varepsilon$:

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

$h - \varepsilon < s \le h$:

$$\sum_{j=1}^{N} c_{2j} T_{2j}(s) = 1 + \Phi(k-s)\,\mathcal{L}(0) + \sum_{j=1}^{N} c_{1j} \int_0^{h-(r-1)\varepsilon} \varphi(z+k-s) T_{1j}(z)\,dz + \sum_{j=1}^{N} c_{2j} \int_{h-(r-1)\varepsilon}^{h-(i-2)\varepsilon} \varphi(z+k-s) T_{2j}(z)\,dz + \sum_{j=1}^{N} c_{3j} \int_{h-(r-2)\varepsilon}^{\varepsilon+s} \varphi(z+k-s) T_{3j}(z)\,dz$$

$h - (r-m+1)\varepsilon < s \le h - (r-m)\varepsilon,\ m = 1, 2, \ldots, r-1$:

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

$h-2\varepsilon < s \le h-\varepsilon$:

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

$h-\varepsilon < s \le h$:

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

$h-\varepsilon < s \le h$:

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

5.2 Two-sided CUSUM-Shewhart, case (ii)

More or less the same arithmetics is utilized as for case (i). Recall the shape of the integral equation for case (ii), that is for $2k \le s^+ + s^- \le h + 2k$:

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

First we plug (3) into and transform the second line

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

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 $\mathcal{L}^-(0)/(\mathcal{L}^+(0) + \mathcal{L}^-(0))$,

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

to get

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

by applying again (4). 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

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

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

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

References

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

Abujiya, M. R.; Riaz, M.; and Lee, M. H. (2013). "Improving the Performance of Combined Shewhart-Cumulative Sum Control Charts". *Quality and Reliability Engineering International*, 29(8), pp. 1193–1206.

Blacksell, S. D.; Gleeson, L. J.; Lunt, R. A.; and 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), pp. 687–699.

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

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

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

Gibbons, R. D. (1999). "Use of Combined Shewhart-CUSUM Control Charts for Ground Water Monitoring Applications". *Ground Water*, 37(5), pp. 682–691.

Henning, E.; Konrath, A. C.; da Cunha Alves, C.; Walter, O. M. F. C.; and 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), pp. 235–243.

Knoth, S. (2006). "Computation of the ARL for CUSUM-S^2 schemes". *Comput. Stat. Data Anal.*, 51(2), pp. 499–512.

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

Lucas, J. M. (1982). "Combined Shewhart-CUSUM Quality Control Schemes". *Journal of Quality Technology*, 14(2), pp. 51–59.

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

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

Montgomery, D. C. (2009). *Statistical quality control: a modern introduction*. Wiley, Hoboken, NJ, 6. ed., internat. student version edition.

Morais, M. C. and Pacheco, A. (2006). "Combined CUSUM-Shewhart Schemes for Binomial Data". *Econ. Qual. Control*, 21(1), pp. 43–57.

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

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

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

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

Shewhart, W. A. (1926). "Quality control charts". *Bell System Technical Journal*, 5(4), pp. 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), pp. 189–193.

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

Yashchin, E. (1985a). "On a unified approach to the analysis of two-sided cumulative sum control schemes with headstarts". *Adv. Appl. Prob.*, 17, pp. 562–593.

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

An Empirical Bayes Approach for Detecting Changes in the Basal Body Temperature

Giovanna Capizzi and Guido Masarotto

Abstract During a normal menstrual cycle, the basal body temperature (BBT) rises around the day of the ovulation. The shift is small; it can be abrupt or gradual and it typically persists almost until the end of the cycle. Detecting the beginning of the BBT rise is an important problem in infertility management, natural family planning and, also, in those medical studies using the day of the BBT rise as a proxy of the day of the ovulation. Traditionally, either some simple run-rules or a CUSUM control chart have been used for detecting such a shift. However, both these approaches do not take into account the information available from the previous cycles of the same woman, or from cycles of other women. Further, the existing procedures do not provide any information on the uncertainty of the detection, and are not robust with respect to some phenomena, such as a fever attack producing outliers in the temperature measurements. In order to overcome these drawbacks, in the paper, we investigate a solution based on the empirical Bayesian paradigm. Real data will be used to illustrate the performance of the suggested approach.

1 Introduction

The basal body temperature (BBT) is the lowest body temperature in a 24-hour period. In clinical practice, it is usually estimated by a temperature measurement immediately after awakening and before any physical activity.

Giovanna Capizzi
Department of Statistical Sciences, University of Padua, Italy, e-mail: giovanna.capizzi@unipd.it

Guido Masarotto
Department of Statistical Sciences, University of Padua, Italy, e-mail: guido.masarotto@unipd.it

In women, the BBT exhibits a typical biphasic behaviour. In particular, during normal menstrual cycles, the BBT slightly increases around the middle of the cycle, and remains to the new level almost until the end. Many studies have correlated the changes in the BBT with other events in the menstrual cycle and established that the rise in the temperature level is due to ovulation (see Marshall, 1963, Moghissi, 1976). Hence, detection of BBT shifts is important in infertile management and natural family planning (e.g. Keck et al, 2007, McVeigh et al, 2013, Furuya et al, 2013, Ecochard et al, 2015). Further, since neglecting the menstrual cycle phase has been shown to cause misinterpretation of some laboratory tests, including many important cardio-metabolic biomarkers (e.g. Schisterman et al, 2014), BBT has been used to guide the correct interpretation of the test results. The time of the BBT rise has also been used in many scientific investigations requiring a *proxy* of the time of ovulation (e.g. Dunson et al, 1999, Colombo and Masarotto, 2000, Bigelow et al, 2004, Bortot et al, 2010, Tenan et al, 2014, 2016, Faustmann et al, 2016)).

Visual detection of the BBT shift is challenging for many menstrual cycles. Indeed, (i) the signal to detect can be as low as two times the precision of a standard thermometer; (ii) the initial level, size and pattern (e.g. abrupt vs gradual) of the temperature shift and the variability within a particular cycle, vary from cycle to cycle and from woman to woman; (iii) missing values and anomalous observations are often present. For this reason, the use of run-rules, like the "three over six rule" that signals the BBT rise the first time three consecutive temperatures are above the level of the immediately preceding six recordings, or CUSUM control charts have been suggested (see Marshall, 1968, Royston and Abrams, 1980, Royston, 1991). However, these traditional approaches do not use the information on the previous cycles of the same woman and their performance is heavily affected by missing or anomalous data. Further, they do not provide any information on the uncertainty of the detection.

In this paper, we propose and investigate an empirical Bayes approach which combines the information provided for a particular cycle by the BBT with the information on the possible time of the BBT shift provided by the previous cycles of the same woman, and also by the cycles of other women. The proposed method is very low-demanding from a computational point of view, and, hence, it can also be implemented on handheld electronic devices. Observe that, while we only focus on the BBT, we believe that our approach can also be useful to detect changes of other menstrual parameters, like the concentration of ovulation-related hormones in the urine.

2 Two databases

To investigate the properties of our method, we use the following two databases.

LONDON: The database, which includes more than 36000 temperature charts provided by about 1800 women, is fully described by Miolo et al (1994). The data

were collected in England and Wales by the Catholic Marriage Advisory Council under the supervision of a scientific committee.

FERTILI: The database, which includes information on more than 7000 menstrual cycles provided by more than 800 women, is described by Colombo and Masarotto (2000). The data have been collected, in six European countries, during a multicentre study on daily fecundability. It contains not only information on the BBT but also on the timing of the sexual intercourses during the menstrual cycle.

From the two databases, we selected menstrual cycles with the following characteristics: (i) length longer than 9 days; (ii) percentage of missing temperatures inferior to 25%. The size of the subsets selected according to (i)-(ii) are

	Women	Cycles
LONDON	1769	26295
FERTILI	769	6865

We use LONDON as a *training* dataset to estimate some characteristics of menstrual cycles and BBTs. FERTILI is used for *testing* the method, and in particular studying the relationship between the BBT rise, as detected by our procedure, and the time of the ovulation.

Observe that both databases only include healthy women. Hence, our results can only be generalized to the corresponding population.

3 An empirical Bayes approach

3.1 Notations

Denote the length (in days) of the jth menstrual cycle for the ith woman by $l_{i,j}$, and the number of days *preceding* the BBT shift by $\tau_{i,j}$. Since long preovulatory phases are very rare, if not physiologically impossible, we suppose that $\tau_{i,j} \leq \tau_{max}$ for some suitable integer τ_{max}. In the following, in order to simplify the notation, we also write $\tau_{i,j} = \tau_{max} + 1$ to indicate that cycle j is monophasic. Further, the BBT behaviour during the final part of long cycles can be anomalous (for example, apparently long cycles can be due to a conception followed by an early miscarriage). Hence, we only consider the temperatures observed at or before day

$$l^{\star}_{i,j} = \min(l_{max}, l_{i,j})$$

with $\tau_{max} < l_{max}$. The choice of τ_{max} and l_{max} is discussed in Subsection 3.5.

Furthermore, denote with

- $y_{i,j,d}$ the BBT measurement on the dth day of the jth cycle for the ith woman;
- $\mathbf{Y}_{i,j,d} = (y_{i,j,1}, \ldots, y_{i,j,d})'$ the vector containing all the BBTs collected up to day d of the same cycle;

– $\mathbf{H}_{i,j} = (l_{i,1}, y_{i,1,1}, \ldots, y_{i,1,l_{i,j,l^{\star}_{i,j}}}, \ldots, l_{i,j-1}, y_{i,j-1,1}, \ldots, y_{i,j-1,l_{i,j-1,l^{\star}_{i,j-1}}})'$ the vector containing all information on the the ith woman gathered *before* the jth cycle.

Note that, when describing our proposal, we assume that no temperature is missing. However, our method can be easily adapted to take into account the presence of missing data. Indeed, missing temperatures are very common in the databases used to empirically test the procedure (see Sections 2 and 4).

To avoid the introduction of many symbols, we use $p(\cdot)$ to indicate the density or probability function of its arguments. In particular, the density function of $\mathbf{Y}_{i,j,d}$ is written as $p(\mathbf{Y}_{i,j,d}|\tau_{i,j}, \boldsymbol{\vartheta}_{i,j})$ where $\boldsymbol{\vartheta}_{i,j}$ denotes all the parameters, different from $\tau_{i,j}$, needed to specify the BBT distribution, e.g., the mean and variance of the temperatures before and after the raise, the proportion of anovulatory cycles, etc. Since $\boldsymbol{\vartheta}_{i,j}$ is completely arbitrary and can also be infinite-dimensional, our notation does not impose any restriction on the BBT distribution. Observe, that $\boldsymbol{\vartheta}_{i,j}$ could have a hierarchical structure of the type

$$\boldsymbol{\vartheta}_{i,j} = (\boldsymbol{\vartheta}^{\star}_{i}, \boldsymbol{\vartheta}^{\star\star}_{i,j}),$$

i.e., $\boldsymbol{\vartheta}_{i,j}$ could comprise parameters, $\boldsymbol{\vartheta}^{\star}_{i}$, which are common to the different cycles of a woman, and parameters, $\boldsymbol{\vartheta}^{\star\star}_{i,j}$, which are specific of the jth cycle. However, in the following, this hierarchical structure will not be exploited. Since $\tau_{i,j}$ and $\vartheta_{i,j}$ vary from cycle to cycle and from woman to woman, we assume that they are random variables.

3.2 The ideal solution

3.2.1 On-line detection

Consider the problem of sequentially monitoring the temperatures of the jth menstrual cycle for the ith women with the aim of detecting as fast as possible the BBT shift. In particular, suppose that on the dth day, $2 \le d \le l^{\star}_{i,j}$, no shift has been signaled yet.

On day d, all the information available on the beginning of the BBT rise are summarized by the conditional probabilities

$$\pi_{i,j}(r|d) = \Pr(\tau_{i,j} = r|\mathbf{Y}_{i,j,d}, \mathbf{H}_{i,j}) \quad r = 1, \ldots, \tau_{max} + 1. \tag{1}$$

In particular, we can signal that the shift has already occured using a rule of the type

$$\Pr(\tau_{i,j} < d|\mathbf{Y}_{i,j,d}, \mathbf{H}_{i,j}) = \sum_{r=1}^{d-1} \pi_{i,j}(r|d) > 1 - \alpha \tag{2}$$

where $0 < \alpha < 1$ is a small risk of false detection which can be chosen on the basis of the particular application. Further, the mode, mean, or median of this conditional

distribution can be used as a point estimate of $\tau_{i,j}$, and the entire distribution (or a summary measure of its dispersion) provides information about the uncertainty on $\tau_{i,j}$.

It is possible to show that

$$\pi_{i,j}(r|d) = \frac{\pi_{i,j}(r)\mathrm{BF}_{i,j}(d|r)}{1+\sum_{s=1}^{d}\pi_{i,j}(s)[\mathrm{BF}_{i,j}(d|s)-1]} \tag{3}$$

where

$$\pi_{i,j}(r) = \Pr(\tau_{i,j} = r|\mathbf{H}_{i,j}) \tag{4}$$

and

$$\begin{aligned}\mathrm{BF}_{i,j}(d|r) &= \frac{p(\mathbf{Y}_{i,j,d}|\tau_{i,j}=r,\mathbf{H}_{i,j})}{p(\mathbf{Y}_{i,j,d}|\tau_{i,j}>d,\mathbf{H}_{i,j})} \\ &= \frac{\int p(\mathbf{Y}_{i,j,d}|\tau_{i,j}=r,\boldsymbol{\vartheta}_{i,j})p(\boldsymbol{\vartheta}_{i,j}|\tau_{i,j}=r,\mathbf{H}_{i,j})d\boldsymbol{\vartheta}_{i,j}}{\int p(\mathbf{Y}_{i,j,d}|\tau_{i,j}>d,\boldsymbol{\vartheta}_{i,j})p(\boldsymbol{\vartheta}_{i,j}|\tau_{i,j}>d,\mathbf{H}_{i,j})d\boldsymbol{\vartheta}_{i,j}}.\end{aligned} \tag{5}$$

Indeed,

$$\begin{aligned}\pi_{i,j}(r|d) &= \frac{p(\tau_{i,j}=r,\mathbf{Y}_{i,j,d}|\mathbf{H}_{i,j})}{p(\mathbf{Y}_{i,j,d}|\mathbf{H}_{i,j})} \\ &= \frac{\pi_{i,j}(r)p(\mathbf{Y}_{i,j,d}|\tau_{i,j}=r,\mathbf{H}_{i,j})}{\sum_{s=1}^{d}\pi_{i,j}(s)p(\mathbf{Y}_{i,j,d}|\tau_{i,j}=s,\mathbf{H}_{i,j})+\Pr(\tau_{i,j}>d|\mathbf{H}_{i,j})p(\mathbf{Y}_{i,j,d}|\tau_{i,j}>d,\mathbf{H}_{i,j})}.\end{aligned}$$

Equation (3) can be obtained dividing the numerator and denominator by $p(\mathbf{Y}_{i,j,d}|\tau_{i,j} > d,\mathbf{H}_{i,j})$ and substituting $\Pr(\tau_{i,j} > d|\mathbf{H}_{i,j})$ with $1-\sum_{s=1}^{d}\pi_{i,j}(s)$.

3.2.2 Off-line detection

Consider now the problem of detecting the BBT shift using all the information on the jth menstrual cycle for the ith woman available on day $l^{\star}_{i,j}+1$. Observe that on this day, we know all the temperatures of the jth cycle we plan to use, and we also know whether the $(l^{\star}_{i,j})$th was or not the last day of the jth cycle, i.e., if the bleeding marking the starting of the next menstrual cycle occured at $l^{\star}_{i,j}+1$.

Regarding the inference on $\tau_{i,j}$, the quantities of interest are

$$\pi_{i,j}(r|l^{\star}_{i,j}+1) = \begin{cases}\Pr(\tau_{i,j}=r|\tau_{i,j}<l_{i,j} \text{ or } \tau_{i,j}=\tau_{max}+1,\mathbf{Y}_{i,j,l_{i,j}},\mathbf{H}_{i,j}) & \text{if } l^{\star}_{i,j}=l_{i,j} \\ \Pr(\tau_{i,j}=r|\mathbf{Y}_{i,j,l^{\star}_{i,j}},\mathbf{H}_{i,j}) & \text{if } l^{\star}_{i,j}\neq l_{i,j}\end{cases}.$$

Indeed, if $l^{\star}_{i,j}$ is equal to the length of the cycle $l_{i,j}$, we know that either the number of days preceding the BBT shift is less than $l_{i,j}$, or the cycle is monophasic.

These conditional probabilities can be computed using the formula

$$\pi_{i,j}(r|l^{\star}_{i,j}+1) = \frac{\pi^{\star}_{i,j}(r)\mathrm{BF}_{i,j}(l^{\star}_{i,j}|r)}{1+\sum_{s=1}^{l^{\star}_{i,j}}\pi^{\star}_{i,j}(s)[\mathrm{BF}_{i,j}(d|s)-1]} \tag{6}$$

where, when $l^{\star}_{i,j} = l_{i,j}$,

$$\pi^{\star}_{i,j}(r) = \begin{cases} \dfrac{\pi_{i,j}(r)}{\sum_{s=1}^{l_{i,j}-1}\pi_{i,j}(s)+\pi_{i,j}(\tau_{max}+1)} & \text{if } 1 \le r < l_{i,j} \text{ or } r = \tau_{max}+1 \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

while, when $l^{\star}_{i,j} \neq l_{i,j}$,

$$\pi^{\star}_{i,j}(r) = \pi_{i,j}(r) \quad (r = 1,\ldots,\tau_{max}+1). \tag{8}$$

3.2.3 Comment

Equations (3)-(5) and (6)-(8) show that the only *ingredients* needed for computing the conditional probabilities $\pi_{i,j}(r|d)$ are:

1. the *prior* probabilities $\pi_{i,j}(r)$, i.e., the distribution of the number of days before the BBT rise $\tau_{i,j}$, given the BBT of the previous cycles, when no observation on the jth cycle has been gathered.
2. The *Bayes factors* $\mathrm{BF}_{i,j}(d|r)$, i.e., a measure of how much the data $\mathbf{Y}_{i,j,d}$ support the hypothesis
$$H_0 = \{\tau_{i,j} = r\} \tag{9}$$
with respect to the hypothesis
$$H_1 = \{\tau_{i,j} > d\}. \tag{10}$$
Observe that $\mathrm{BF}_{i,j}(d|r)$ is similar to the classical *likelihood-ratio* test statistic. The main difference consists in the way the nuisance parameters $\boldsymbol{\vartheta}_{i,j}$ are eliminated, i.e., by integrating (not maximizing) over the parameter space.

In the following subsections, we suggest reasonable approximations of these two "ingredients".

3.3 Distribution of the number of days preceding the BBT rise

Regarding the distribution of $\tau_{i,j}$, we obtained encouraging results assuming that:

(i) For the ith woman, the number of days preceding the BBT rise, $\tau_{i,j}$, $j = 1,2,\ldots$, are independently and identically distributed (i.i.d.) such that

$$p(\tau_{i,j} = r) = \beta_{i,r}, \quad (r = 1, \ldots, \tau_{max} + 1),$$

where $\beta_{i,r}$ are parameters, depending on the characteristics of the ith woman such that

$$\beta_{i,r} \geq 0 \quad \text{and} \quad \sum_{r=1}^{\tau_{max}+1} \beta_{i,r} = 1.$$

(ii) $\boldsymbol{\beta}_i = (\beta_{i,1}, \ldots, \beta_{i,\tau_{max}+1})'$ has a Dirichlet distribution , i.e.,

$$p(\boldsymbol{\beta}_i) = \frac{\Gamma(\lambda_1 + \cdots + \lambda_{\tau_{max}+1})}{\Gamma(\lambda_1) \cdots \lambda(\lambda_{\tau_{max}+1})} \prod_{r=1}^{\tau_{max}+1} \beta_{i,r}^{\lambda_r - 1}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{\tau_{max}+1})'$ is a vector of non-negative parameters. Hence, for $r, s = 1, \ldots, \tau_{max} + 1$, $r \neq s$,

$$E(\beta_{i,r}) = \mu_r, \quad \text{var}(\beta_{i,r}) = \frac{\mu_r(1 - \mu_r)}{1 + \eta}, \quad \text{cov}(\beta_{i,r}, \beta_{i,s}) = -\frac{\mu_r \mu_s}{1 + \eta} \tag{11}$$

where $\mu_r = \lambda_r / \eta$ and $\eta = \lambda_1 + \cdots + \lambda_{\tau_{max}}$. The estimation of $\boldsymbol{\lambda}$ is discussed in Subsection 3.5.

Under assumptions (i) and (ii),

$$\pi_{i,j}(r) = E(\beta_{i,r} | \mathbf{H}_{i,j}) \quad r = 1, \ldots, \tau_{max} + 1.$$

The computation is immediate for $j = 1$ and gives

$$\pi_{i,1}(r) = \mu_r \quad r = 1, \ldots, \tau_{max} + 1.$$

Since the Dirichlet distribution is the conjugate prior of a multinomial distribution, it is possible to show that $p(\boldsymbol{\beta}_i | \mathbf{H}_{i,2})$ is a mixture of $\tau_{max} + 1$ Dirichlet distributions, and, in particular, that

$$p(\boldsymbol{\beta}_i | \mathbf{H}_{i,2}) = \sum_{r=1}^{\tau_{max}+1} \pi^{\star}_{i,1}(r | l^{\star}_{i,1} + 1) p(\boldsymbol{\beta}_i | \tau_{i,1}), \tag{12}$$

where $p(\boldsymbol{\beta}_i | \tau_{i,1})$ is the density of a Dirichlet random variable with parameters $\boldsymbol{\lambda} + \mathbf{e}_{i,1}$. Here, $\mathbf{e}_{i,j}$ denotes vector of length $\tau_{max} + 1$ with the $(\tau_{i,j})$th element equal to one and all the others equal to zero. In the same way, it is possible to show that $p(\boldsymbol{\beta}_i | \mathbf{H}_{i,3})$ is a mixture of $(\tau_{max} + 1)^2$ Dirichlet distributions, $p(\boldsymbol{\beta}_i | \mathbf{H}_{i,4})$ of $(\tau_{max} + 1)^3$ Dirichlet distributions, etc..

As time increases, the computational burden also increases due to the related explosion of the number of mixture components. For this reason, following an idea often found in the literature on non-linear time-series filtering (see Prado and West, 2010, Barber, 2012), we suggest to replace mixtures (12) with a single Dirichlet distribution determined minimizing the Kullback-Leibler divergence, i.e., to recur-

sively approximate the conditional density $p(\boldsymbol{\beta}_i|\mathbf{H}_{i,j})$, and its expected value $\pi_{i,j}(\cdot)$, using the following algorithm:

Assumption. For each i and j, *pretend* that $\boldsymbol{\beta}_i$ given $\mathbf{H}_{i,j}$ has a Dirichlet distribution with parameters $\boldsymbol{\lambda}_{i,j} = (\lambda_{i,j,1}, \ldots, \lambda_{i,j,\tau_{max}+1})'$.

Initialization. Set $\boldsymbol{\lambda}_{i,1} = \boldsymbol{\lambda}$.

Updating. On the first day of the jth cycle, $j = 2, 3, \ldots$, compute the "new" parameters $\boldsymbol{\lambda}_{i,j}$ minimizing

$$\sum_{r=1}^{\tau_{max}+1} \left[\log\Gamma(\lambda_{i,j,r}) - \lambda_{i,j,r}u_{i,j,r}\right] - \log\Gamma(\lambda_{i,j,1} + \cdots + \lambda_{i,j,\tau_{max}+1}) \qquad (13)$$

where

$$u_{i,j,r} = \psi(\lambda_{i,j-1,r}) - \psi(1 + \lambda_{i,j-1,1} + \cdots + \lambda_{i,j-1,\tau_{max}+1}) - \frac{\pi_{i,j-1}(r|l^{\star}_{i,j-1})}{\lambda_{i,j-1,r}}$$

with $\psi(\cdot)$ denoting the digamma function, i.e., the derivative of the logarithm of the gamma function. Indeed, it is possible to show that minimizing (13) is equivalent to minimizing the Kullback-Leibler divergence between

- the density of a Dirichlet distribution of parameters $\boldsymbol{\lambda}_{i,j}$, and
- the density of the mixture (12), with $\boldsymbol{\lambda}$ replaced by $\boldsymbol{\lambda}_{i,j-1}$ and $\mathbf{e}_{i,1}$ by $\mathbf{e}_{i,j-1}$.

For minimizing (13), we implemented a Newton-Raphson algorithm similar to that proposed by Sklar (2014) for computing the maximum likelihood estimate of the parameters of a Dirichlet distribution.

Approximation of the prior probabilities. Set

$$\pi_{i,j}(r) = \frac{\lambda_{i,j,r}}{\lambda_{i,j,1} + \cdots + \lambda_{i,j,\tau_{max}+1}}.$$

3.4 Test-based Bayes factors

Two approaches can be followed for computing the Bayes factors $BF_{i,j}(d|r)$:

1. *Full model-based Bayes factors:* this approach requires the specification of the "extra" parameters $\boldsymbol{\vartheta}_{i,j}$, including their woman-to-woman and within-woman variation, and the density function $p(Y_{i,j,d}|\tau_{i,j}, \boldsymbol{\vartheta}_{i,j})$. It also requires the computation of the conditional distributions $p(\tau_{i,j}, \boldsymbol{\vartheta}_{i,j}|\mathbf{H}_{i,j})$ and of the integrals in (5).
2. *Test-based Bayes factors:* using this approach, we must choose a suitable test statistic for the hypothesis system (9)-(10); then, the Bayes factors are compute summarizing the evidence provided by the test statistic (see Johnson, 2005, 2008, Yuan and Johnson, 2008, Held et al, 2015).

Table 1: Kass and Raftery (1995) scale of interpretation of a Bayes factor and corresponding values of $W_{i,j}(d|r)$ when $\omega_0 = 0$.

Strength of the evidence in favour of the alternative hypothesis	Bayes factor	Standardized Wilcoxon test statistic
negative	≤ 1	≤ 0
not worth more than a bare mention	1 to 3	0 to 0.67
positive	3 to 20	0.67 to 1.67
strong	20 to 150	1.67 to 2.48
very strong	> 150	> 2.48

In principle, the first approach may offer a better efficiency. However, in the present context, it would be inevitably based on arbitrary and simplistic assumptions. For these reasons, and because we believe that the efficiency loss would be modest, given the high BBT variability, we have decided to follow the second approach. In particular, we obtained good results using Bayes factors based on the Wilcoxon rank-sum test statistics.

Let $v_{i,j,1}, \ldots, v_{i,j,d}$ be the ranks of the observations $y_{i,j,1}, \ldots, y_{i,j,d}$. If there are tied observations, give to tied observations the average of the ranks for which those observations are competing. Further, define for $1 \leq r < d$,

$$W_{i,j}(d|r) = \frac{\sum_{s=r+1}^{d} v_{i,j,s} - \mu(d,r)}{\sigma(d,r)}$$

where $\mu(r,d)$ and $\sigma(d,r)$ are such that

$$E(W_{i,j}(d|r)) = 0 \quad \text{and} \quad \text{var}(W_{i,j}(d|r)) = 1$$

when $y_{i,j,1}, \ldots, y_{i,j,d}$ are i.i.d. (see Hollander et al (2014), p. 118). Let $\omega = E(W_{i,j}(d|r))$. The hypothesis system (9)-(10) roughly corresponds to the system

$$H_0 : \omega \leq \omega_0 \quad vs. \quad H_1 : \omega > \omega_0$$

where $\omega_0 \geq 0$ is a constant whose choice will be discussed in Subsection 3.5. Assuming a non-informative (improper) prior distribution for ω,

$$p(\omega) \propto \text{constant} \quad \text{for } -\infty < \omega < \infty,$$

and pretending that $W_{i,j}(d|r) \sim N(\omega, 1)$, we obtain the Bayes factor

$$BF_{i,j}(d|r) = \frac{\Pr(\omega > \omega_0 | W_{i,j}(d|r))}{\Pr(\omega \leq \omega_0 | W_{i,j}(d|r))} = \frac{1 - \Phi(W_{i,j}(d|r) - \omega_0)}{\Phi(W_{i,j}(d|r) - \omega_0)} \quad (1 \leq r < d) \quad (14)$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Further, we define $BF_{i,j}(d|r) = 1$ for $r \geq d$. As shown by Table 1, formula (14) essentially "translates" the evidence provided by the Wilcoxon test statistic into that often used for the Bayes factors.

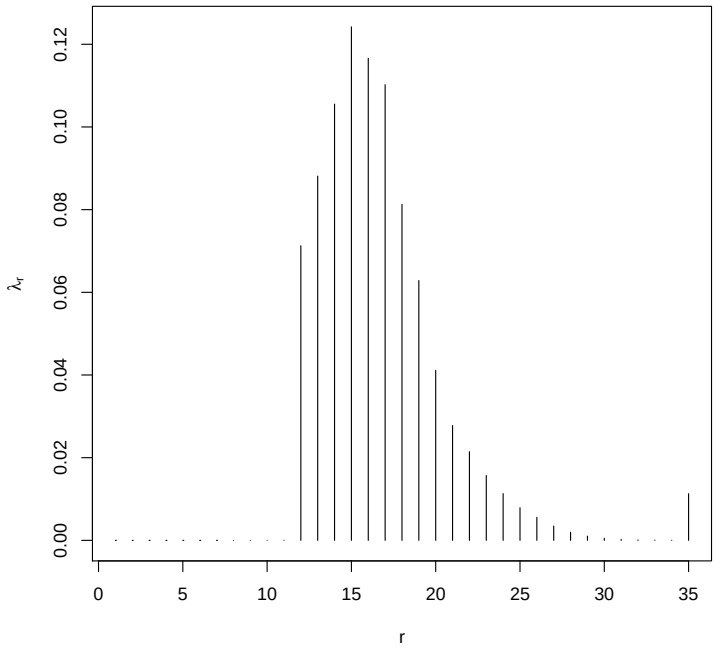


Fig. 1: Maximum likelihood estimates of $\boldsymbol{\lambda}$ (database: LONDON, $l_{max} = 35$, $\tau_{max} = 34$).

3.5 Implementation

The practical implementation of the suggested approach requires the selection of:

- α: the risk of giving false alarms when the detection rule (2) is used.
- l_{max} and τ_{max}: the maximum number of BBT measurements to consider and the maximum number of days preceding the BBT shift (see Subsection 3.1).
- $\boldsymbol{\lambda}$: the parameters of the Dirichlet distribution describing the heterogeneity between women of the distribution of $\tau_{i,j}$ (see Subsection 3.3).
- ω_0: the threshold used in the definition of the Bayes factors (14).

We believe that the choice of α should strongly depend on the particular application. For example, in natural family planning, the detection of the BBT raise is used to mark the end of the so-called "fertile window", i.e., the end of the days in which the woman is fertile. When a false detection occurs, some fertile days are declared as infertile resulting in an increase risk of an undesired pregnancy. It is not difficult to understand that different women/couples can have a different attitude with respect to this risk. In any case, in all the examples shown in the next Section, we use $\alpha = 0.05$.

On the basis of the study by Guo et al (2006), we suggest to set $l_{max} = 35$. Observe that according to Guo et al (2006), this value approximately corresponds to

$$\text{mean} + 2 \times \text{standard deviation}$$

of the distribution of the lengths of "normal menstrual cycles".

In our opinion, the choice of τ_{max} and $\boldsymbol{\lambda}$ should be handled simultaneously. In particular, using the training database LONDON, we estimated $\boldsymbol{\lambda}$ maximizing

$$\ell = \sum_{i,j} \log \sum_{r=1}^{\tau_{max}+1} \pi^{\star}_{i,j}(r) BF_{i,j}(l^{\star}_{i,j}|r).$$

Under the reasonable hypothesis that, when the cycle is monophasic, the BBT distribution does not depend on $\boldsymbol{\lambda}$, maximizing ℓ is equivalent to maximizing the log-likelihood function. Indeed, the latter is equal to

$$\begin{aligned}
&\sum_{i,j} \log p(\mathbf{Y}_{i,j,l^{\star}_{i,j}}|H_{i,j}) \\
&\quad = \sum_{i,j} \log \sum_{r=1}^{\tau_{max}+1} \pi^{\star}_{i,j}(r) p(\mathbf{Y}_{i,j,l^{\star}_{i,j}}|\tau_{i,j} = r, H_{i,j}) \\
&\quad = \sum_{i,j} \log \sum_{r=1}^{\tau_{max}+1} \pi^{\star}_{i,j}(r) \frac{p(\mathbf{Y}_{i,j,l^{\star}_{i,j}}|\tau_{i,j} = r, H_{i,j})}{p(\mathbf{Y}_{i,j,l^{\star}_{i,j}}|\tau_{i,j} = \tau_{max}+1, H_{i,j})} + \\
&\quad\quad \sum_{i,j} \log p(\mathbf{Y}_{i,j,l^{\star}_{i,j}}|\tau_{i,j} = \tau_{max}+1, H_{i,j}) \\
&\quad = \ell + \sum_{i,j} \log p(\mathbf{Y}_{i,j,l^{\star}_{i,j}}|\tau_{i,j} = \tau_{max}+1, H_{i,j}).
\end{aligned}$$

The estimates, computed assuming $\tau_{max} = 34$, are shown in Figure 1. Observe that λ_r is practically zero when $30 < r \leq 34$. Hence, the choice of τ_{max} seems to be reasonable. The estimated value of $\eta = \lambda_1 + \cdots + \lambda_{\tau_{max}}$ is 0.901. Such a small value points to a high heterogeneity of $\boldsymbol{\beta}_i$ among women (see equation (11)).

Finally, ω_0 should be selected to balance the probability to detect real BBT changes with that of signaling false changes. Indeed, ω_0 can be viewed as a penalization applied to the evidence against the null hypothesis provided by the Wilcoxon test statistic $W_{i,j}(d|r)$. A large value of ω_0 decreases the Bayes factors $BF_{i,j}(d|r)$, and, hence, reduces the number of false detections, but also the ability to detect real shifts. For this reason, we suggest to select ω_0 on the basis of the data. In particular, as documented in the next Section, we obtained good results using

$$\omega_0 = z_1 \max\left(0, \min\left(1, \frac{W^{max}_{i,j}(d) - z_1}{z_2 - z_1}\right)\right)$$

with

$$W^{max}_{i,j}(d) = \max_r W_{i,j}(d,r),\ z_1 = \Phi^{-1}\left(1 - \frac{1}{2d}\right) \text{ and } z_2 = \Phi^{-1}\left(1 - \frac{1}{4d}\right).$$

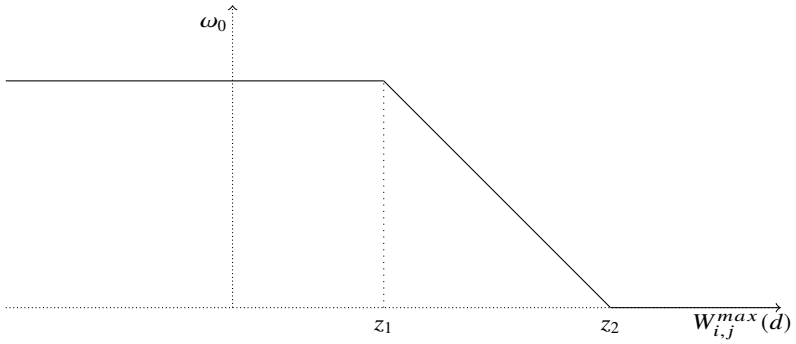


Fig. 2: An adaptive choice of ω_0

Note that (i) the suggested value of ω_0 decreases with the maximum of the test statistics used on the dth day (see Figure 2); (ii) z_1 and z_2 can be viewed as a (very rough) Bonferroni-based approximation to the median and the 75% percentile of the distribution of the maximum when there is no change.

4 Some Empirical Results

4.1 Examples

Figure 3 illustrates the prospective application of the proposed method. In particular,

- The first panel displays the BBTs observed during a menstrual cycle. The cycle lasted 28 days ($l_{i,j} = 28$). Note the missing temperature on day 2. The cycle is clearly biphasic, with a shift occuring around day 15.
- In the second panel, the conditional probabilities $\Pr(\tau_{i,j} < d|Y_{i,j,d}, H_{i,j})$ are plotted as a function of the day of the cycle d. Note that this probabilities only depends on the temperatures up to the dth day, i.e., they are appropriate for on-line monitoring (see Subsection (3.2)). These probabilities have been computed using a prior distribution $\pi_{i,j}(\cdot)$ proportional to the estimated λs shown in Figure 1.

  For this cycle, $\Pr(\tau_{i,j} < d|Y_{i,j,d}, H_{i,j})$ is very small for $d \leq 15$. Then, it abruptly increases. In particular, it exceeds the threshold 0.95, shown by the dashed line, from day 18 onwards. Hence, if a rule of type (2) is used, with $\alpha = 0.05$, the method signals the BBT shift on day 18.
- In the third panel, the conditional probabilities $\pi_{i,j}(r|d = 18)$ are displayed as a function of the day r. These probabilities, which only depend on the temperatures available on the day of the alarm, strongly suggest that the BBT shift occured from day 12 to 17 and that the 15th day is the most probable.

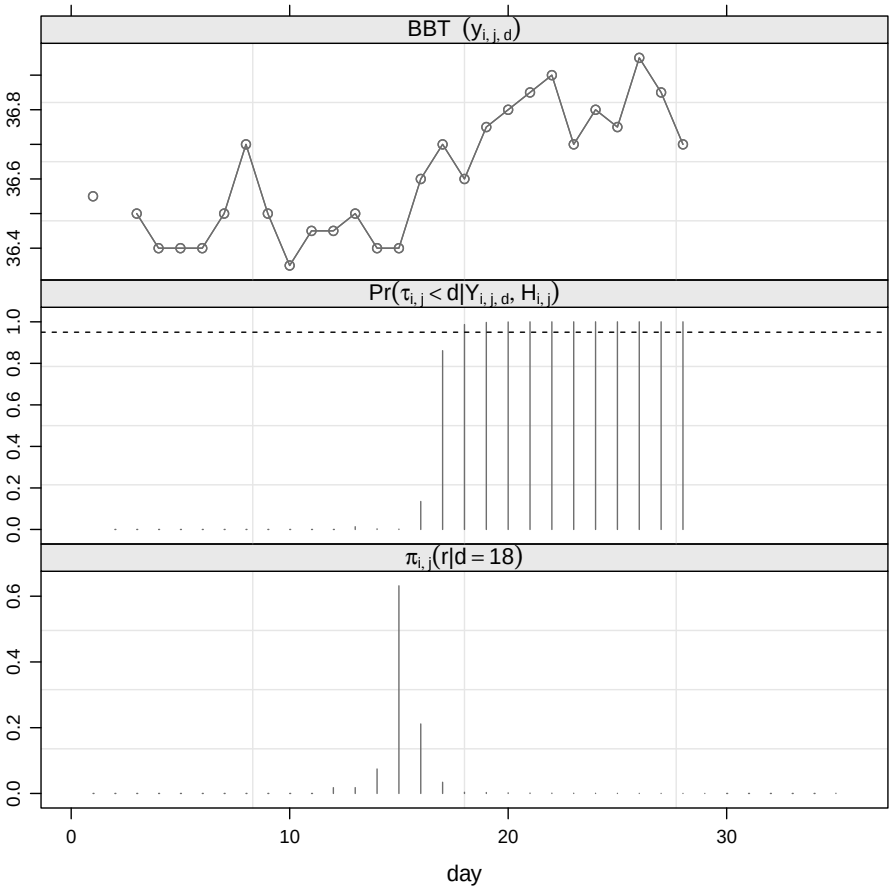


Fig. 3: Sequential detection of the BBT shift in a particular menstrual cycle.

Figures 4-6 illustrate the learning of the woman characteristics proposed in Subsection 3.3. In particular,

- Figure 4 shows the temperatures collected during five consecutive menstrual cycles.
- Figure 5 displays the corresponding Bayes factors $BF(l^{\star}_{i,j}|r)$ computed using all the cycle temperatures.
  Figures 4 and 5 suggest that, for the considered woman, the BBT typically starts to increasing around days $19-20$.
- Figure 6 shows the prior probabilities $\pi_{i,j}(r)$, $j = 1,\ldots,5$. For the first cycle ($j = 1$), no prior information on the woman is available. Hence, the prior probabilities are proportional to the λs displayed by Figure 1. However, for the successive cycles ($j > 1$), the prior probabilities take into account the information gathered during the previous cycles. Observe that, coherently with the evidence shown by

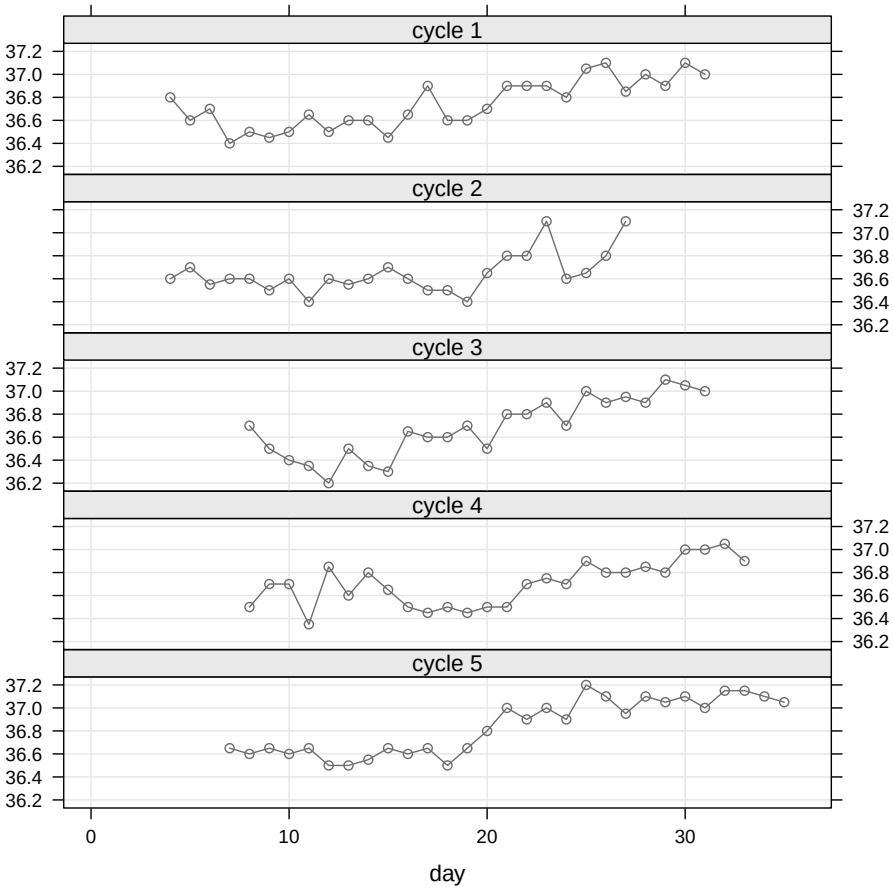


Fig. 4: BBT measurements in five consecutive menstrual cycles of the same woman.

Figures 4 and 5, the mode of the prior distribution shifts from 15 to 19, and that, in general, the $\pi_{i,j}(r)$ probabilities progressively increase for r close to $r = 19$, and decrease when r is close to 15.

4.2 What are we detecting?

In this Subsection, using the FERTILI database, we investigate the relationship between the day of the BBT shift, determined using the proposed approach, and the day of the ovulation. Results suggest that the conditional probabilities $\pi_{i,j}(r|d^{\star}_{i,j})$ describe not only the uncertainty about the time of the BBT shift but also about the time of the ovulation. Here, $d^{\star}_{i,j}$ denotes the day of the signal, i.e.,

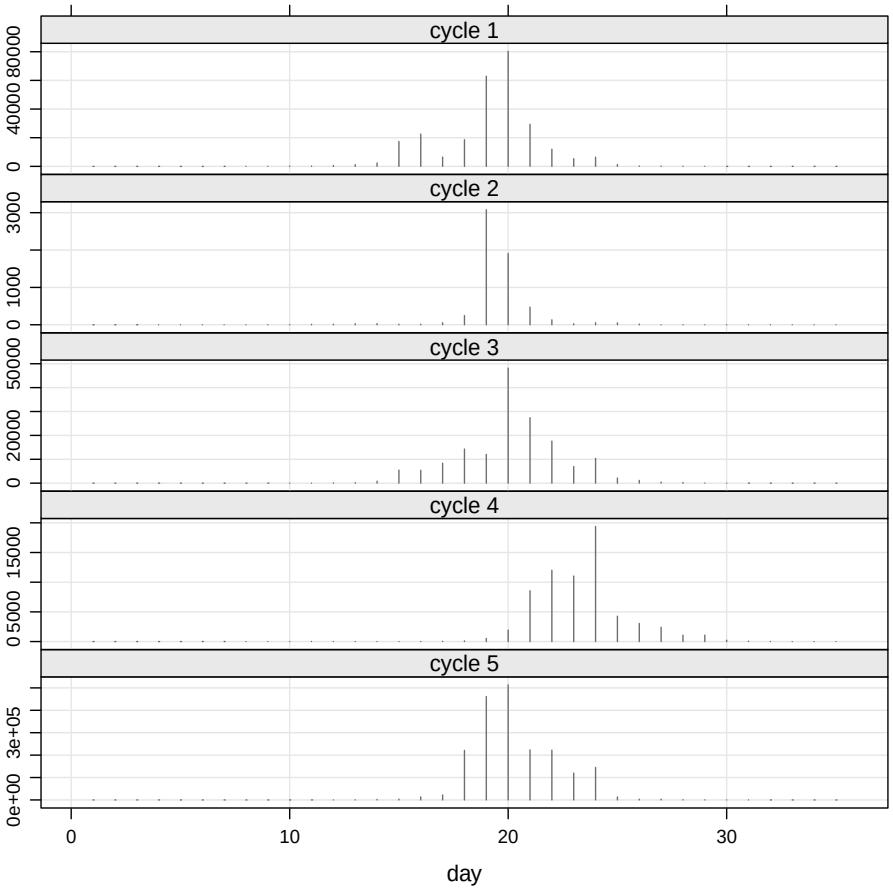


Fig. 5: Bayes factors for the five menstrual cycles shown in Figure 4.

$$d^{\star}_{i,j} = \inf\{2 < d \leq l^{\star}_{i,j} : \Pr(\tau_{i,j} < d | \mathbf{Y}_{i,j,d}, \mathbf{H}_{i,j}) > 1 - \alpha\}.$$

Let

- $C_{i,j} = 1$ if conception occurs in cycle j for couple i, and $C_{i,j} = 0$ otherwise;
- $x_{i,j,d} = 1$ if there was intercourse on day d, $d = 1, \ldots, l_{i,j}$ of the same cycle, and $x_{i,j,d} = 0$ otherwise. In order to simplify the notation, we also define $x_{i,j,d} = 0$ if $d \leq 0$ or $d > l_{i,j}$.

The Barrett-Marshall-Schwartz model is a biologically plausible model used to relate the sexual intercourse pattern $\mathbf{X}_{i,j} = (x_{i,j,1}, \ldots, x_{i,j,l_{i,j}})'$ and an ovulation marker, like $\tau_{i,j}$, to the probability of conception (see Barrett and Marshall, 1969, Schwartz et al, 1980, Dunson et al, 1999, Colombo and Masarotto, 2000, Dunson and Weinberg, 2000, Dunson, 2001, Dunson et al, 2001, Bigelow et al, 2004). This model assumes that sperm introduced into the reproductive tract on different days commingle and

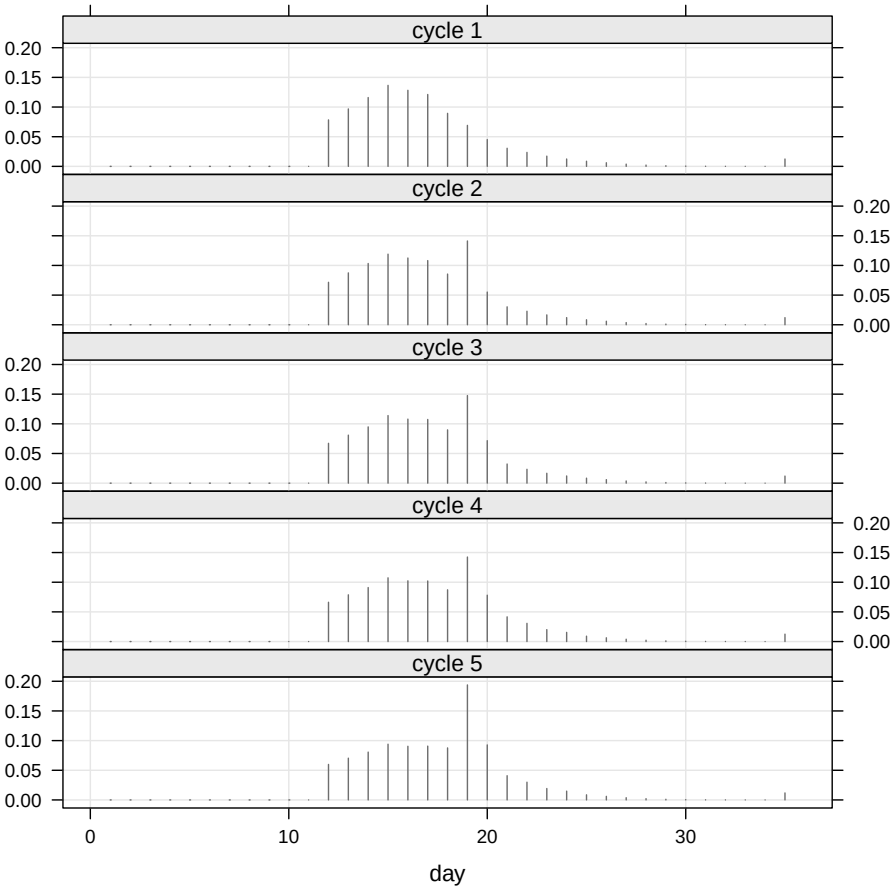


Fig. 6: Prior probabilities $\pi_{i,j}(r)$ for the five menstrual cycles shown in Figure 4.

then compete independently to fertilize the ovum. According to the model,

$$\Pr(C_{i,j} = 1 | \xi_i, \tau_{i,j}, \mathbf{X}_{i,j}) = \xi_i \left\{ 1 - \prod_{r=-u}^{v} (1 - \rho_r)^{x_{i,j,\tau_{i,j}+r}} \right\} \tag{15}$$

where

- $0 \le \xi_i \le 1$ represents the probability that factors unrelated to the timing of intercourse are favorable for conception. It is often referred to the probability of cycle viability, and used to capture the fertility heterogeneity among couples. In particular, we assume that ξ_i is distributed as a beta random variable with parameters ζ_1 and ζ_2.
- $u \ge 0$ and $v \ge 0$ describe the width of the fertile window around $\tau_{i,j}$; days outside the interval $[\tau_{i,j} - u; \tau_{i,j} + v]$ are assumed infertile.

– $0 \leq \rho_r \leq 1, r = -u, \ldots, v$, represent the probabilities that conception would have resulted from intercourse on the $(\tau_{i,j} + r)$th day, if the cycle were viable. Observe that $\xi_i \rho_r$ is the probability that conception would result from intercourse only on the $(\tau_{i,j} + r)$th day.

In practice, we know the temperatures $y_{i,j,d}$, not $\tau_{i,j}$. However, combining equations (1) and (15), we obtain

$$\Pr(C_{i,j} = 1 | \xi_i, \mathbf{Y}_{i,j,d^\star_{i,j}}, \mathbf{H}_{i,j}, \mathbf{X}_{i,j}) = \sum_{r=1}^{\tau_{max}} \frac{\pi_{i,j}(r|d^\star_{i,j})}{1 - \pi_{i,j}(\tau_{max} + 1|d^\star_{i,j})} \Pr(C_{i,j} = 1 | \xi_i, \tau_{i,j} = r, \mathbf{X}_{i,j}).$$

Last expression is similar to the mixture model for fecundability introduced by Dunson and Weinberg (2000) and Dunson et al (2001) for taking into account the measurement errors of one or more ovulation markers. The main difference is that using our approach the error distribution is known and cycle-specific.

Assuming, as it is usually done, that the conception indicators $C_{i,j}$ are conditionally independent given the viability probabilities ξ_i, parameters $\zeta_1, \zeta_2, \rho_{-u}, \ldots, \rho_v$ can be estimated maximizing the likelihood function. Figure 7 displays the estimates computed from the FERTILI database. It is interesting to observe that the day-specific probabilities ρ_r are very close to zero when $r < -4$ and $r > 0$. Indeed, if we estimate u and v using the information criterion BIC, we obtain $\hat{u} = -4$ and $\hat{v} = 0$. Hence, the fertile window is essentially given by the intervals $[\tau_{i,j} - 4; \tau_{i,j}]$. But this finding is what we expect if $\tau_{i,j}$ is the day of the ovulation. In addition, the level and pattern of the day-specific probabilities of conception, around $\tau_{i,j}$, are very similar to those around the ovulation day estimated by Dunson et al (2001) using the data from the North Carolina Early Pregnancy Study. In that study, the ovulation was estimated for each menstrual cycle using measures of estrogen and progesterone metabolites and lutenizing hormone in the urine. Hence, our tentative conclusion is that, the conditional probabilities $\pi_{i,j}(r|d^\star_{i,j})$, available on the day of the signal, describe not only the uncertainty about the BBT shift but, at least approximately, also the uncertainty about the day of the ovulation.

5 Conclusions

We have suggested a new method, applicable both prospectively and retrospectively, for detecting the BBT rise, during a menstrual cycle. The approach combines prior information provided by the previous cycles with the information on the BBT of the current cycle. A distinctive aspect of our proposal consists in providing not only an estimate of the time of the shift but also an estimate of the distribution of the error committed in determining the BBT shift.

Future research will include the comparison of our methods with other suggested in the literature and the inclusion of possible covariates, e.g., the age of the woman.

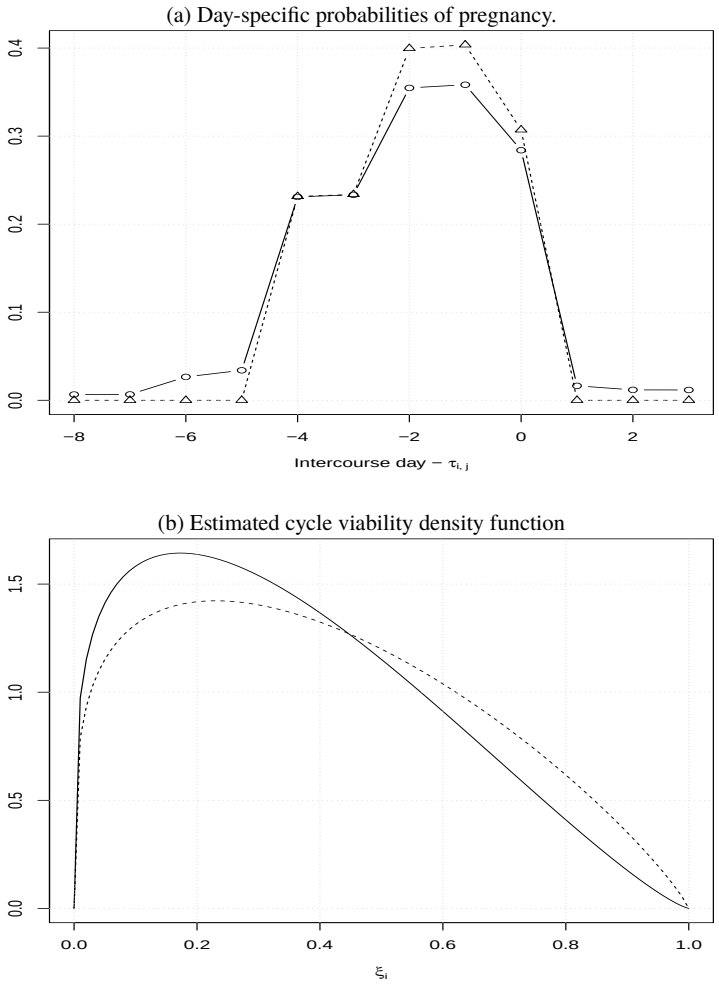


Fig. 7: Barrett-Marshall-Schwartz fecundability model. Panel (a) shows, for a couple with average viability, the maximum-likelihood estimates of the probabilities that conception would result from intercourse only on the $(\tau_{i,j} + r)$ day. Panel (b) shows the maximum-likelihood estimate of the density function of the cycle viability ξ_i. Solid lines: estimates obtained using $u = -8$ and $v = 3$. Dashed lines: estimates obtained selecting the fertile window using the information criterion BIC ($\hat{u} = -4$ and $\hat{v} = 0$).

References

Barber D (2012) Bayesian Reasoning and Machine Learning. Cambridge University Press, Cambridge, UK

Barrett JC, Marshall J (1969) The risk of conception on different days of the menstrual cycle. Population studies 23:455–461

Bigelow JL, Dunson DB, Stanford JB, Ecochard R, Gnoth C, Colombo B (2004) Mucus observations in the fertile window: A better predictor of conception than timing of intercourse. Human Reproduction 19:889–892

Bortot P, Masarotto G, Scarpa B (2010) Sequential predictions of menstrual cycle lengths. Biostatistics 11:741–55

Colombo B, Masarotto G (2000) Daily fecundability: First results from a new data base. Demographic Research 3:1–39

Dunson D, Baird D, Wilcox AJ, Weinberg CR (1999) Day-specific probabilities of clinical pregnancy based on two studies with imperfect measures of ovulation. Human Reproduction 14:1835–1839

Dunson DB (2001) Bayesian modeling of the level and duration of fertility in the menstrual cycle. Biometrics 57:1067–1073

Dunson DB, Weinberg CR (2000) Modeling human fertility in the presence of measurement error. Biometrics 56:288–292

Dunson DB, Weinberg CR, Baird DD, Kesner JS, Wilcox AJ (2001) Assessing human fertility using several markers of ovulation. Statistics in Medicine 20:965–978

Ecochard R, Duterque O, Leiva R, Bouchard T, Vigil P (2015) Self-identification of the clinical fertile window and the ovulation period. Fertility and Sterility 103:1319–132,525

Faustmann G, Tiran B, Maimari T, Kieslinger P, Obermayer-Pietsch B, Gruber HJ, Roob JM, Winklhofer-Roob BM (2016) Circulating leptin and nf-κb activation in peripheral blood mononuclear cells across the menstrual cycle. BioFactors Early view

Furuya S, Kikuchi F, Kagawa T (2013) Revisiting basal body temperature (BBT) measurement in non-ART infertility treatment – Do couples seeking pregnancy really need it? The second report. Fertility and Sterility 100:S319

Guo Y, Manatunga AK, Chen S, Marcus M (2006) Modeling menstrual cycle length using a mixture distribution. Biostatistics 7:100–114

Held L, Sabanés Bové D, Gravestock I (2015) Approximate bayesian model selection with the deviance statistic. Statistical Science 30:242–257

Hollander M, Wolfe DA, Chicken E (2014) Nonparametric Statistical Methods, 3rd edn. Wiley

Johnson VE (2005) Bayes factors based on test statistics. Journal of the Royal Statistical Society Series B (Statistical Methodology) 67:689–701

Johnson VE (2008) Properties of Bayes factors based on test statistics. Scandinavian Journal of Statistics 35:354–368

Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90:773–795

Keck C, Tempfer CB, Hugues JN (2007) Conservative Infertility Management. CRC Press, Boca Raton, FL

Marshall J (1963) Thermal changes in the normal menstrual cycle. British Medical Journal 1:102–104

Marshall J (1968) A field-trial of the basal-body temperature method for regulating births. Lancet 2:8–10

McVeigh E, Guillebaud J, Homburg R (2013) Oxford Handbook of Reproductive Medicine and Family Planning, 2nd edn. Oxford University Press, Oxford, UK

Miolo L, Colombo B, Marshall J (1994) A data base for biometric research on changes in basal body temperature in the menstrual cycle. Statistica LIII:563–572

Moghissi KS (1976) Accuracy of basal body temperature for ovulation detection. Fertility and Sterility 27:1415–1421

Prado R, West M (2010) Time Series: Modeling, Computation, and Inference. CRC Press, Boca Raton, FL

Royston JP, Abrams RM (1980) An objective method for detecting the shift in basal body temperature in women. Biometrics 36:217–224

Royston P (1991) Identifying the fertile phase of the human menstrual cycle. Statistics in Medicine 10:221–240

Schisterman EF, Mumford SL, Sjaarda LA (2014) Failure to consider the menstrual cycle phase may cause misinterpretation of clinical and research findings of cardiometabolic biomarkers in premenopausal women. Epidemiologic Reviews 36:71–82

Schwartz D, MacDonald P, Heuchel V (1980) Fecundability, coital frequency and the viability of ova. Population studies 34:397–400

Sklar M (2014) Fast MLE computation for the Dirichlet multinomial. Tech. rep., arXiv:1405.0099

Tenan MS, Brothers RM, Tweedell AJ, Hackney AC, Griffin L (2014) Changes in resting heart rate variability across the menstrual cycle. Psychophysiology 51:996–1004

Tenan MS, Hackney AC, Griffin L (2016) Maximal force and tremor changes across the menstrual cycle. European Journal of Applied Physiology 116:153–160

Yuan Y, Johnson VE (2008) Bayesian hypothesis tests using nonparametric statistics. Statistical Sinica 18:1185–1200

A Generalized Likelihood Ratio Test for Monitoring Profile Data

Yang Liu, JunJia Zhu and Dennis K. J. Lin

Abstract Profile data emerges when the quality of a product or process is characterized by a functional relationship among (input and output) variables. In this paper, it is assumed that each profile has one response variable Y, one explanatory variable x, and the functional relationship between these two variables can be rather arbitrary. We propose a general method based on the Generalized Likelihood Ratio Test (GLRT) to perform Phase II monitoring of profile data. Unlike existing methods in profile monitoring area, the proposed method uses nonparametric regression to estimate the on-line profiles and thus does not require any functional form for the profiles. Both Shewhart-type and EWMA-type control charts are considered. The average run length (ARL) performance of the proposed method is studied by using a nonlinear profile dataset. It is shown that the proposed GLRT-based control chart can efficiently detect both location and dispersion shifts of the on-line profiles from the baseline profile. An upper control limit (UCL) corresponding to a desired in-control ARL value is constructed.

Key words: Average Run Length; Generalized Likelihood Ratio Test; Nonparametric Regression; Profile Monitoring; Statistical Process Control

1 Introduction

In statistical process control (SPC) applications, the quality of a process can often be adequately described by a univariate quality characteristic. Sometimes it needs to be characterized by a relationship between two or more variables, however. Specifically,

Yang Liu
Shanxi University of Finance and Economics, Taiyuan 030006, China, e-mail: ly8941@126.com

JunJia Zhu and K. J. Lin
The Pennsylvania State University, University Park, PA 16803, USA, e-mail: jxz203@psu.edu and e-mail: DKL5@psu.edu

the system engineer has one main variable of interest (the response variable Y) and one or more environmental variables or control variables (the explanatory variable x's). In such situations, to monitor the quality of the process is to monitor the relationship or "profile" between Y and x's. The profile data is also referred as "waveform signal" or "signature" in some literatures. The application of statistical process control techniques on the profile data is called profile monitoring. Profile monitoring is inevitable when the variability of the response variable Y cannot be adequately explained by the values of Y themselves.

Here, we focus on monitoring relationship between Y and x. Both variables are assumed to be continuous. In a profile n pairs of (Y, x) are observed. The observed Y_i $(i = 1, 2, ..., n)$ is believed to be the function $f(x_i)$ plus a random error ϵ_i, which is assumed as a random variable with mean 0 and a constant variance (σ^2). Under the settings above, monitoring the profiles is equivalent to monitoring both the change of the function $f(\cdot)$ and the change of the distribution of ϵ_i.

Profile monitoring methods and applications, like many applications in control charts, can also be divided into two phases: Phase I and Phase II. In Phase I applications, one analyzes a set of historical profile data. The main goals in Phase I applications are (1) to understand the variation in a process over time; (2) to evaluate the process stability; and (3) to model the in-control process performance. The evaluation of Phase I methods is mainly focused on assessing the probability of signal (POS) studies, i.e. the probability of giving at least one out-of-control signal when applying the control chart to the historical profile dataset.

In Phase II applications, one is interested in monitoring the process using on-line data. The goal is to detect shifts in the process from the baseline profile obtained in Phase I as quickly as possible. The evaluation of Phase II methods is mainly focused on the run-length distribution, where the run length is the number of samples taken before an out-of-control signal is given. The average run length (ARL) is often used to compare the performance of competing control charts in Phase II.

The main focus here is on Phase II monitoring of profiles. A general method based on Generalized Likelihood Ratio Test (GLRT) is proposed to detect the shift of the on-line profiles from the baseline profile. The existing methods in profile monitoring describe the profile function via parametric models. Therefore monitoring the profiles becomes monitoring the shift of parameters in the model. However, fixing the profile data to a specific parametric model is sometimes unrealistic and could be burdensome. The proposed control charts use nonparametric regression to estimate the on-line profiles and thus does not require any functional form for the profiles. It is flexible for many practical situations.

The remainder of this paper is organized as follows: in Section 2, the preliminary statistical method is discussed. The Shewhart-type and EWMA-type control charts are then considered and the details are discussed in Section 3. In Section 4, the application of the proposed control chart is illustrated by a vertical density profile (VDP) dataset of nonlinear profiles. In Section 5, the ARL performance of the proposed control chart is demonstrated using the VDP data again. This paper is finished with conclusions and discussion in Section 6.

2 Preliminary

2.1 Modeling for Phase II Monitoring of Profiles

The profile (the relationship between the response variable Y and the explanatory variable x) can usually be described as the following theoretical relationship:

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, ..., n, \tag{1}$$

where $f(x_i)$ is the mean function, ϵ is the random error, and n is the number of points within a profile.

The simplest profile is the so-called linear profile, where the mean function is a simple linear regression function: $f(x_i) = \beta_0 + \beta_1 X_i$ and the random error $\epsilon_i \sim N(0, \sigma^2)$, for $i = 1, 2, ..., n$. The existing literatures focus on monitoring the linear profiles, see, for example, Kang and Albin (2000), Kim et al (2003), Mahmoud and Woodall (2004),Zhang et al(2009) and Xu et al(2012).

Figure 1 shows a typical example of a non-linear on-line profile. For the nonlinear profiles, one can either use parametric model to describe the relationship, or more naturally, model the profile in a nonparametric way.

Building a statistical model to appropriately describe the mean function of non-linear profile is a main task in Phase I application of profile monitoring. The system engineer needs to carefully distinguish the in-control profiles from the out-of-control profiles and use only the group of in-control profiles to build up the statistical model. Sometimes the mean function $f(x)$ already has a built-in parametric form due to the inherent nature of the production process of the profile. Only the in-control profiles are needed to estimate the parameter for Phase II use. However, in many cases the mean function are too complicated to be modeled by any parametric models. In this situation, if the sample size n within each profile is sufficiently large, then one solution is to use nonparametric regression methods.

In Phase II application, one assume that a "baseline" profile has been built and the goal is to compare the on-line profile data with the baseline profile, i.e. one is testing:

$$\begin{aligned} H_0: &\quad Y = f_0(x) + \epsilon_0, \quad \epsilon_0 \sim N(0, \sigma_0^2), \\ H_1: &\quad Y = f_1(x) + \epsilon_1, \quad \epsilon_1 \sim N(0, \sigma_1^2). \end{aligned} \tag{2}$$

where $f_0(x)$ is the baseline profile function, which is known (in either parametric or nonparametric form), and $f_1(x)$ is the on-line profile function, which needs to be estimated.

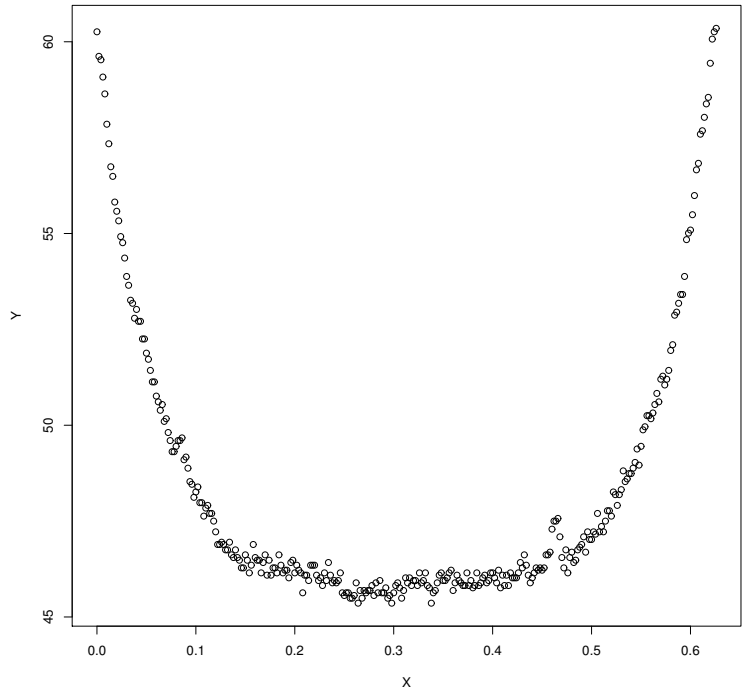


Fig. 1: An Example of Non-Linear Profile

2.2 Likelihood Ratio Test (LRT)

If a parametric function is used to model both the baseline profile $f_0(x)$ and the on-line profile $f_1(x)$, then likelihood ratio test (LRT) is a straightforward approach to perform Phase II test for the departure of on-line profiles from the baseline. Likelihood ratio test is a commonly used statistical hypothesis test for testing the departure of a vector of parameter values from their hypothetical values. Define Θ as a vector of parameters that are used in the parametric model. To test the null hypothesis : $H_0 : \theta \in \Theta_0$ against the alternative hypothesis $H_1 : \theta \in \Theta$ where $\Theta \neq \Theta_0$, the LRT statistic is defined as:

$$\lambda(Y|x) = \frac{\sup\{L(\theta|Y|x : \theta \in \Theta_0)\}}{\sup\{L(\theta|Y|x : \theta \in \Theta)\}}.$$

A likelihood ratio test has any rejection region of the form $\{\lambda(x) < c\}$, where c is any number satisfying $0 \leq c \leq 1$ (Casella and Berger 2002). Among all tests with a

given probability of Type-I error, the likelihood ratio test is shown to minimize the probability of a Type-II error (Rice 1995).

The Likelihood Ratio Test can also be used if nonparametric regression model is used for on-line profiles, as long as the maximum likelihood estimator (MLE) for the nonparametric regression is available. Fan et al. (2001) gave one example of using LRT to test if the model function is linear. Suppose under model (1) with ϵ_i being a sequence of independent and identically distributed (iid) random variables from $N(0, \sigma^2)$, one wants to test $H_0 : f(x) = \beta_0 + \beta_1 x$ against $H_1 : f(x) \neq \beta_0 + \beta_1 x$. Let $(\hat{\beta}_0, \hat{\beta}_1)$ be the maximum likelihood estimator (MLE) under H_0, and $\hat{f}_{MLE}(.)$ be the nonparametric MLE function under H_1 obtained by minimizing $\sum_{i=0}^{n}(Y_i - f(x_i))^2$ (subject to the support of the parameter space for the nonparametric regression). The logarithm of the conditional maximum likelihood ratio statistic is then:

$$\lambda_n = \ell_n(\hat{m}_{MLE}) - \ell_n(H_0) = \frac{n}{2} \log \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{n}(Y_i - \hat{f}_{MLE}(x_i))^2},$$

The asymptotic distribution of $2\lambda_n$ under null hypothesis is a χ^2 distribution with appropriate degrees of freedom. This properties can be used to set up the upper control limit (UCL) for the control chart to be used for profile monitoring.

2.3 Generalized Likelihood Ratio Test (GLRT)

However, likelihood ratio test is not always applicable when nonparametric regression is used. Fan et al. (2001) pointed out that, in general, MLEs under nonparametric regression models can be difficult to obtain. They proposed a new method to replace the maximum likelihood estimator under the alternative nonparametric model, leading to the generalized likelihood ratio

$$\lambda_n = \ell_n(H_1) - \ell_n(H_0), \tag{3}$$

where $\ell_n(H_1)$ is the log-likelihood with unknown regression function replaced by a reasonable nonparametric regression function. The λ_n is called the generalized likelihood ratio test (in short GLRT) statistic.

Fan et al. (2001) showed that similar to the LRT statistics, the GLRT statistics also have the so-called "Wilks-type phenomenon", i.e., their asymptotic null distribution are independent of nuisance parameters. Because of this property, the null distribution of the GLRT statistic does not have to be derived theoretically. Instead, one can simply simulate the null distributions. This makes the generalized likelihood ratio test powerful and suitable for many practical situations. It will be discussed (later in this paper) that parametric bootstrap Monte-Carlo simulation can be used to obtain the null distribution of the GLRT statistic, which is shown to be nearly χ^2.

3 Proposed Control Charts

In Phase II monitoring of profile data, it is assumed that the baseline profile and the in-control variance have been appropriately estimated. A sequence of tests is conducted to check whether the on-line profile is different from the baseline profile. Both Shewhart-type control charts and EWMA-type control charts based on GLRT will be discussed.

3.1 Derivation of the GLRT statistic for Profile Monitoring

The null and alternative hypotheses is shown in (2). To distinguish H_0 and H_1, it is assumed that either $f_0(x) \neq f_1(x)$, or $\sigma_0 \neq \sigma_1$, or both. In Phase II, $f_0(x)$ and σ_0 are known, while $f_1(x)$ and σ_1 need to be estimated.

The likelihood functions under H_0 and H_1 are:

$$Under\ H_0 :\ _{H_0} = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (Y_i - f_0(x_i))^2\right),$$
$$Under\ H_1 :\ _{H_1} = \left(\frac{1}{\sqrt{2\pi}\hat{\sigma_1}}\right)^n \exp\left(-\frac{1}{2\hat{\sigma_1}^2}\sum_{i=1}^n (Y_i - \hat{f}_1(x_i))^2\right),$$

where $f_0(x_i)$ is the value of Y_i given x_i based on the baseline profile; $\hat{f}_1(x_i)$ is the estimated value of Y_i given x_i based on the on-line profile and $\hat{\sigma_1}^2$ is the estimated variance for the on-line profile (this is obtained by a nonparametric regression method, such as spline or local linear regression method).

Define $RSS_0 = \sum_{i=1}^n (Y_i - f_0(x_i))^2$ and $RSS_1 = \sum_{i=1}^n (Y_i - \hat{f}_1(x_i))^2$, the log-likelihoods will be:

$$\text{Under } H_0 : \ell_{H_0} = -n\log\sqrt{2\pi} - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}RSS_0, \text{ and}$$
$$\text{Under } H_1 : \ell_{H_1} = -n\log\sqrt{2\pi} - \frac{n}{2}\log\hat{\sigma_1}^2 - \frac{1}{2\hat{\sigma_1}^2}RSS_1$$
$$= -n\log\sqrt{2\pi} - \frac{n}{2}\log\frac{RSS_1}{n} - \frac{n}{2}.$$

The generalized likelihood ratio test (GLRT) statistic to be used in our proposed control chart is then:

$$\lambda = \ell_{H_1} - \ell_{H_0} = \frac{n}{2}\log\sigma_0^2 + \frac{1}{2\sigma_0^2}RSS_0 - \frac{n}{2}\log\frac{RSS_1}{n} - \frac{n}{2}. \tag{4}$$

3.2 Shewhart-type Control Chart

Shewhart-type control chart treats each individual profile as independent entry and thus is ideal for detecting isolated spike shift (outliers) and large sustained step shift of profile parameters. Once an on-line (linear) profile is established, one can calculate the GLRT statistic λ and compare it to an upper control limit (UCL), which is associated with a pre-defined type-I error rate (α). The lower control limit (LCL) for Shewhart-type GLRT control chart is usually set as 0.

The UCL is obtained via the empirical null distribution of the GLRT statistic. This can be done by a parametric bootstrap Monte-Carlo simulation, as described below:

Step 1: Generate an on-line profile under the null hypothesis, and obtain the estimate of the GLRT statistic $\hat{\lambda}$ for this profile;
Step 2: Repeat step 1 many times to obtain the empirical distribution of λ;
Step 3: Use the $100(1-\alpha)$th percentile of the empirical distribution of λ as the UCL.

3.3 EWMA-type Control Chart

EWMA-type control charts make use of both current profile and all previous profiles to generate the test statistic, and thus are more efficient in detecting small sustained shift (change point). An EWMA-type control chart based on GLRT can be built by defining:

$$EWMA_i = \theta\lambda_i + (1-\theta)EWMA_{i-1}, \tag{5}$$

where $i \geq 1$ and $0 < \theta \leq 1$ is a smoothing constant. The initial value $EWMA_0$ is set as the mean value of λ under H_0, which can be estimated by a bootstrap Monte-Carlo simulation introduced above.

The LCL for the EWMA-type control chart is set as 0 and the UCL can be determined by Monte-Carlo simulations (Yeh et al, 2004) according to a pre-specified α. Compared with the Shewhart chart, the EWMA chart has a complicated settings. Thus its UCL is difficult to obtain. The UCL of EWMA control chart for monitoring χ^2 random variables has been thoroughly studied in the literatures, however (Knoth 2005). The theoretical UCL for a pre-specified α, when the random variable being monitored is a χ^2 random variable, can be evaluated by an `R` package called `spc`. The desired UCL for EWMA-type GLRT chat can be estimated either by using bi-sectional search method or by approximation method, as discussed below.

Bi-sectional Search Method:

The bi-sectional algorithm is a commonly used computational algorithm in searching for solutions in non-linear equations. The Bi-sectional search for desired UCL for EWMA-type GLRT chart can be done in the following steps:

Step 1: Select a temporary value for the lower limit of UCL (UCL_l) and a temporary value for the upper limit of UCL (UCL_u), then calculate the corresponding ARL estimate for these two limits: ARL_l and ARL_u respectively. It is desirable that $ARL_l < ARL_{IC} < ARL_u$, where ARL_{IC} stands for the in-control ARL.

Step 2: Let UCL_{tmp} = average(UCL_l, UCL_u), then calculate the corresponding ARL estimate ARL_{tmp}. If $ARL_{tmp} > ARL_{IC}$, then assign $UCL_u = UCL_{tmp}$ and $ARL_u = ARL_{tmp}$. Otherwise, assign $UCL_l = UCL_{tmp}$ and $ARL_l = ARL_{tmp}$.

Step 3: If the absolute value of $|ARL_u - ARL_l|$ is less than a pre-defined threshold, the convergence criteria met and the desired UCL = average(UCL_l, UCL_u), otherwise go back to Step 2.

Note that since the calculation of the ARL for EWMA-type chart can be time consuming, the initial value of UCL_u in Step 1 should not be large.

Our empirical experience indicates that the bi-sectional search method works well in finding the desired UCL for EWMA-type GLRT charts when the simulation size to calculate the ARL values is large. In our simulation we use 5000 Monte-Carlo runs to estimate the ARL values when the ARL.IC is set as 20 (correspond to $\alpha = 0.05$). The simulation size, however, should increase accordingly when the desired ARL_{IC} value increases. The larger the simulation size, the more accurate the desired UCL. Note that the computational time will also increase exponentially. So the bi-sectional search could be slow in convergence, especially when the ARL_{IC} value and the simulation size are large.We thus propose an approximation method to overcome the computational problem by taking advantage of the simulated empirical LRT statistics. The approximation method is based on the estimated mean of the GLRT statistic under the null hypothesis. The `R` codes for estimation of the UCL are given in the Appendix. The algorithm can be outlined in the following steps.

Step 1. Simulate the empirical distribution of λ.

Step 2. Let df_{approx} equal to the sample mean of the simulated λ's.

Step 3. Use the `R` package `spc` to calculate the UCL corresponding to a desired in-control ARL for χ^2 random variable with degrees of freedom equals to df_{approx}.

The approximation method can be regard as a quick-and-dirty way to find the UCL for a given in-control ARL. If the α value is relatively large (such as 0.05), then the approximation works fairly well. However, for smaller α values (such as 0.001 which correspond to larger in-control ARLs) the use of the approximation method should be cautious since it tends to overestimate the UCL values.

To study the distribution of the GLRT statistic under the null hypothesis, a quantile-quantile plot (QQ plot) of the GLRT statistic under H_0 against χ^2 Random variables with df = mean(GLRT) is given as Figure 2. The empirical $100(1-\alpha)$%th percentile of GLRT statistic are also plotted as reference lines. It can been seen that for $\alpha \geq 0.05$ the quantiles of the GLRT statistic under H_0 are almost identical to their counterparts of the χ^2 random variables. But the GLRT statistic tend to have larger values for the tail of its distribution.

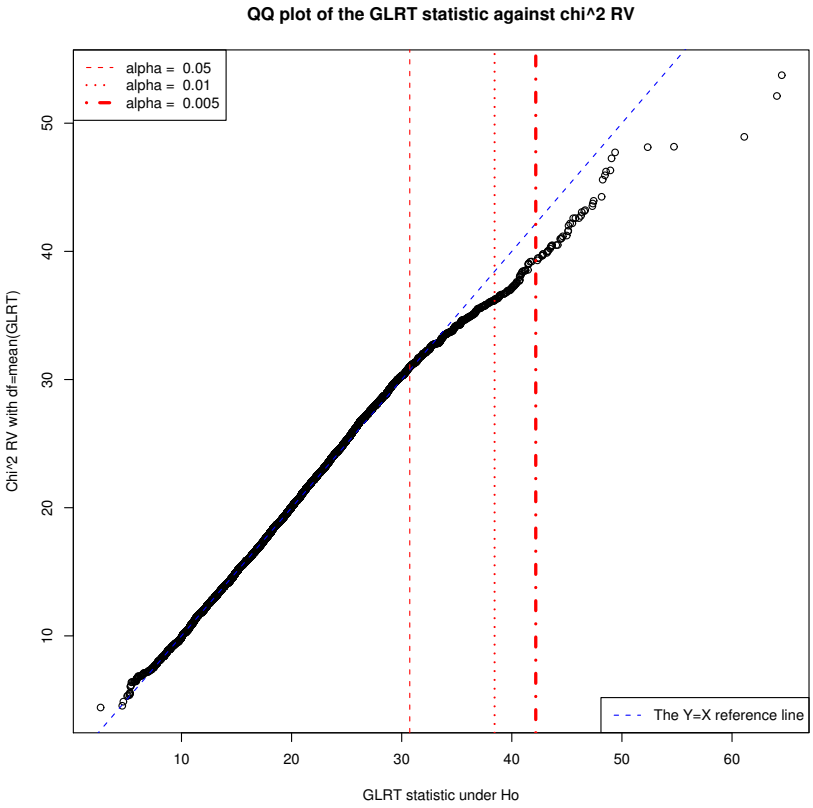


Fig. 2: QQ Plot of the GLRT Statistic against χ^2 Random variables with df = mean(GLRT)

4 An Illustrative Example

The vertical density profile (VDP) dataset reported in Walker and Wright (2002) (available at the website http://bus.utk.edu/stat/walker/VDP/Allstack.TXT) is discussed here for illustration of the proposed method. For the VDP data, the density of the wood board (Y) is measured by using a profilometer that uses a laser device to take measurements at fixed depths (x) across the depth of the thickness of the board. The first VDP profile, for example, is displayed in Figure 1.

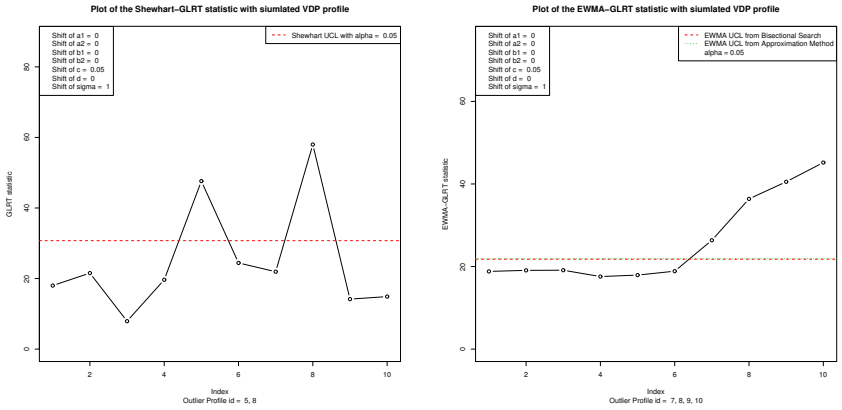
Williams et al. (2003) used a parametric nonlinear "bathtub-shape" function to model the VDP data:

$$f(x_i, \underline{\beta}) = \begin{cases} a_1(x_i - d)^{b_1} + c & x_i > d \\ a_2(d - x_i)^{b_2} + c & x_i \leq d \end{cases}$$

where $\underline{\beta} = (a_1, a_2, b_1, b_2, c, d)'$. The parameters a_1 and a_2 determine the width of the "bathtub"; b_1 and b_2 determine the flatness of the "bathtub"; c is the bottom of the "bathtub"; and d is the center of the "bathtub". Furthermore, it is assumed that $\epsilon \sim Normal(0, \sigma^2)$. Note that if nonparametric regression model is used to build the baseline profile, the proposed GLRT chart still works. However, the parametric baseline profile model does give us more flexibility in manipulating the shifts of out-of-control profiles.

The parameter estimates based on the first VDP profile from Williams et al. (2003) are $\underline{\beta} = (a_1, a_2, b_1, b_2, c, d)' = (5708, 3921, 5.14, 4.87, 46.0, 0.313)'$. These values are used to build the baseline profile. In addition, the in-control variance is set as $\sigma^2 = 0.1^2$.

We first demonstrate the situation of random shifts (random outliers). Two outlier profiles are randomly generated out of ten on-line profiles. For each on-line profile the GLRT statistic is calculated against the baseline profile. The Shewhart-type GLRT control chart is plotted in Figure 3(a). The UCLs correspond to $\alpha = 0.05$ is ploted, which corresponds to in-control ARL value of 20. The UCL is estimated based on 1,000,000 runs of parametric bootstrap Monte-Carlo simulations.



(a) Shewhart-type GLRT Chart for Isolated Shifts

(b) EWMA-type GLRT Chart for Sustained Step Shifts

Fig. 3: GLRT-based Control Chart on Simulated VDP Data

For parameter $\underline{\beta}$ the shifted parameter is defined as $\underline{\beta_{acutal}} = \Delta\underline{\beta} + \underline{\beta_{IC}}$, while for the variance parameter σ it is defined as $\sigma_{actual} = \Delta\sigma * \sigma_{IC}$. In Figure 3(a) the two outliers (#5 and #8) have the same shift in parameter c which shifts from the in-control value 46 to 46.05. It is clear that the Shewhart-type GLRT control chart can efficiently detect these two outliers.

Next, we demonstrate the situation of substantial shifts. Four out-of-control on-line profiles (profile # 7 through #10) are generated after six in-control profiles

(profile #1 through #6). The EWMA-type GLRT control chart is plotted in Figure 3(b). Again, the α is set at 0.05, and the four sustained step shift profiles have the same shift in parameter c from the in-control value 46 to 46.05. The UCLs obtained using both bisectional search and approximation method are used. It can be seen that the EWMA-type easily detect those four out-of-control profiles, and their corresponding EWMA-GLRT statistic are increasing over time.

5 ARL Performance of the Proposed Chart

In Phase II applications of control charts, the performance of proposed control chart is typically evaluated by the study of its run length distribution. Especially, the average run length (ARL) under in-control and various out-of-control situations has been a major criteria in evaluating the effectiveness of control charts (see also, Wang and Lin, 2016).

Following the settings in the previous section, the "bathtub-shape" function is used to describe the baseline VDP profile and test the ARL performance of the proposed control chart under various out-of-control situations. The in-control ARL is 20. The ARLs of Shewhart-type GLRT chart are estimated by 1,000,000 runs of Monte Carlo simulations, while the ARLs of EWMA-type GLRT chart are estimated by 10,000 runs of simulations. Again, for EWMA-type control chart, the UCLs obtained from both bisectional search method (referred as UCL1) and approximation method (referred as UCL2) are used. The following shifts will be studied in details: $a_1 \to a_1 + \Delta_{a_1}$, $b_1 \to b_1 + \Delta_{b_1}$, $c \to c + \Delta_c$, $d \to d + \Delta_d$, and $\sigma \to \Delta_\sigma \sigma$.

Shift of Parameter a_1:

For VDP data, the parameter a_1 and a_2 determine the width of the "bathtub". Since a_1 and a_2 are symmetric in the bathtub formula, only a_1 is used to test the ARL performance of the proposed chart. With all other parameters unchanged, we shift $a_1 \to a_1 + \Delta_{a_1}$, with Δ_{a_1} = 5, 10, ..., 100.

The estimated ARL of the proposed GLRT chart for shift of parameter a_1 is listed in Table 1 and plotted in Figure 4. As one can see, the EWMA-type GLRT chart outperform (with smaller out-of-control ARL) the Shewhart-type GLRT chart, when the shift is small. With the increasing of the level of the shift, the performances of these two types of charts become very close. It is noted that for the EWMA-type GLRT control chart using UCL obtained by approximation method (UCL2) tend to over estimate ARL a little bit than that using UCL obtained by bisectional search method (UCL1), in both in-control and out-of-control case. However the difference between these two methods become almost unnoticeable when shift gets larger.

Shift of Parameter b_1, c, d, and σ:

- The parameter b_1 and b_2 determine the flatness of the "bathtub". Since b_1 and b_2 are symmetric in the formula, only b_1 is used to test the ARL performance of the

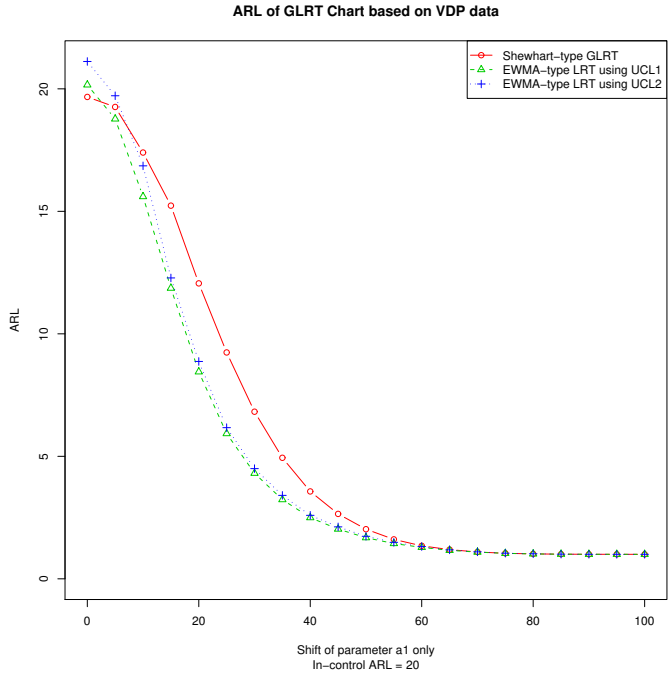


Fig. 4: Plot of the ARL Estimate of the Proposed GLRT Control Chart (For parameter a_1 shift of VDP profile)

proposed chart. With all other parameters unchanged, we shift $b_1 \to b_1 + \Delta_{b_1}$, with $\Delta_{b_1} = 0.001, 0.002, ..., 0.01$.

- The parameter c determine the bottom of the "bathtub". So a shift of the parameter c yields a vertical shift of the whole "bathtub" curve. With all other parameters unchanged, we shift $c \to c + \Delta_c$, with $\Delta_c = 0.005, 0.01, ..., 0.05$.
- The parameter d determine the center of the "bathtub". So a shift of the parameter d yields a horizontal shift of the whole "bathtub" curve. With all other parameters unchanged, we shift $d \to d + \Delta_d$, with $\Delta_d = 0.0001, 0.0002, ..., 0.0005$.
- While the other parameters determine the mean value (location) of the in-control VDP profile, the parameter σ determine the dispersion of the in-control profile.
- The proposed control chart is designed to monitor the shift of dispersion parameter also. With all other parameters unchanged, we shift $\sigma \to \Delta_\sigma \sigma$, with $\Delta_\sigma = 1.05, 1.1, ..., 1.4$.

The estimated ARL of the proposed GLRT charts in monitoring the shifts of those four parameters (b_1, c, d, and σ) are displayed in Figure 5. The general pattern of these ARL performances is similar to Figure 4 as discussed above.

Table 1: ARL Estimate of the Proposed GLRT Control Chart (For parameter a_1 shift of VDP profile)

Δ_{a_1}	Shewhart-type	EWMA-type (UCL2)	EWMA-type (UCL1)
0	19.67	21.12	20.16
5	19.26	19.72	18.77
10	17.40	16.86	15.60
15	15.23	12.28	11.86
20	12.06	8.87	8.45
25	9.24	6.17	5.93
30	6.82	4.50	4.31
35	4.94	3.41	3.23
40	3.57	2.59	2.50
45	2.65	2.12	2.03
50	2.03	1.74	1.68
55	1.61	1.49	1.44
60	1.35	1.32	1.28
65	1.19	1.18	1.16
70	1.10	1.10	1.09
75	1.04	1.05	1.04
80	1.02	1.02	1.02
85	1.01	1.01	1.01
90	1.00	1.00	1.00
95	1.00	1.00	1.00
100	1.00	1.00	1.00

In summary, the proposed GLRT-based control charts can efficiently monitor various types of shift (both location and dispersion) of the on-line profiles from the baseline profile. The cases in which the system has multiple parameter shifts simultaneously are also studied in our simulation. The results are similar to those of the single parameter shift cases so they are omitted in this paper. Within the scope of our simulation, the ARL performance of the EWMA-type GLRT chart is (slightly) better than that of the Shewhart-type GLRT chat.

6 Conclusions and Discussions

In this paper, we develop the control chart based on generalized likelihood ratio test (GLRT) statistic to monitor profile data. The proposed method uses nonparametric regression and thus there is no restriction on the functional form of the profiles. The proposed control chart is mainly used in Phase II applications, i.e. in detecting the shift of on-line profiles from the baseline profile.

Both Shewhart-type and EWMA-type control charts are discussed. Finding the UCL associated with the desired in-control ARL for the proposed GLRT-based chart can sometimes be computationally challenge, especially for EWMA-type LRT-based

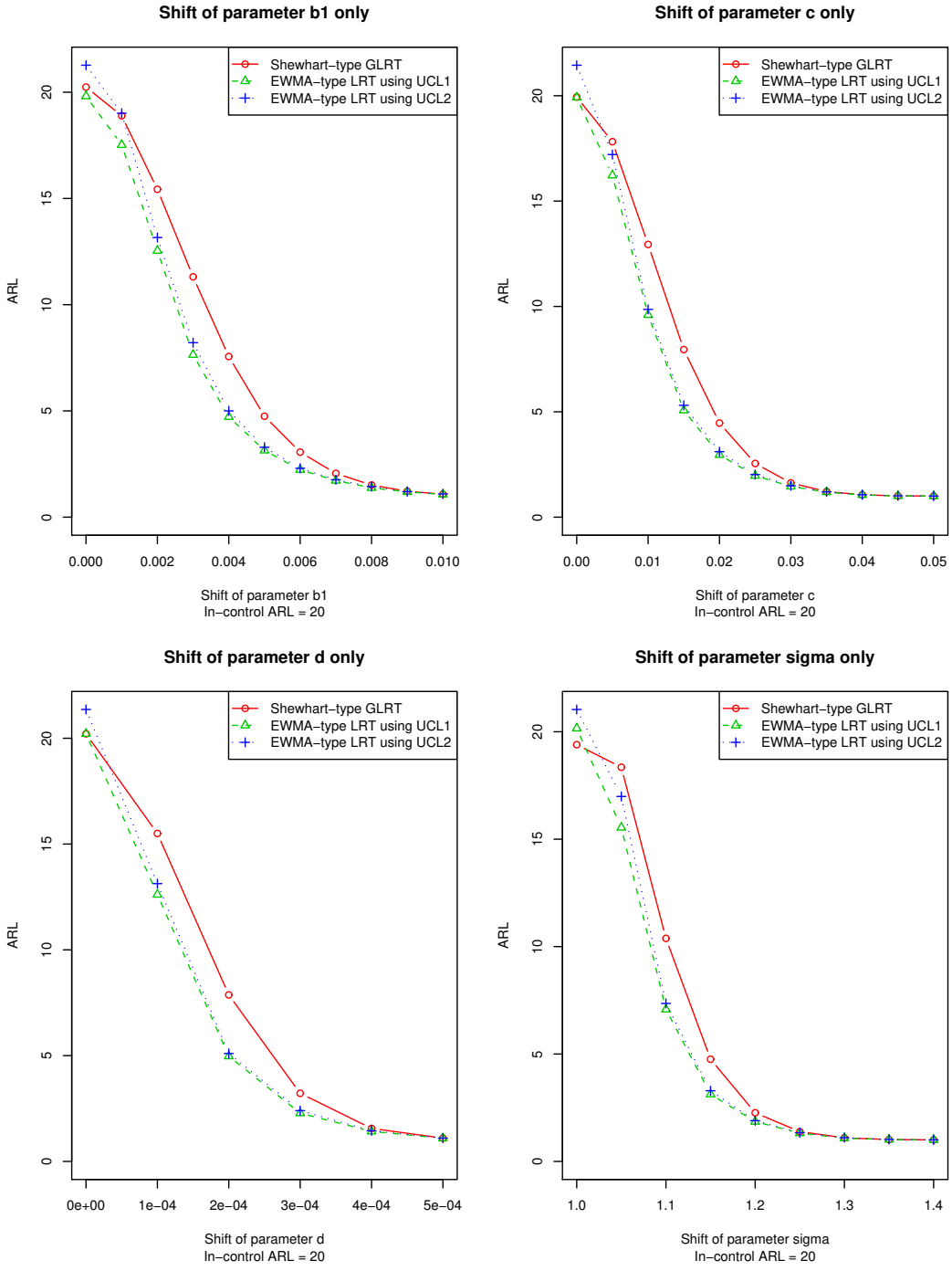


Fig. 5: Plot of the ARL Estimate of the Proposed GLRT Control Chart, for shift of parameters b_1, c, d, and σ of VDP profile

chart. The process in finding the desired UCL for the proposed GLRT-based chart are discussed. The UCL for Shewhart-type GLRT chart is obtained by parametric bootstrap Monte-Carlo simulation. The UCL of EWMA-type GLRT chart can be obtained by either bisectional search method or approximation method.

Using a "bathtub-shape" function to model a real-life VDP nonlinear profile, we demonstrate the ARL performance of the proposed control charts. The simulation results show that the proposed control chart can efficiently detect various types of shifts of on-line profiles. Especially, the EWMA-type chart outperform the Shewhart-type GLRT chart in giving out-of-control signals when the shifts of on-line profiles are small.

The proposed GLRT statistic is based on using nonparametric regression method to estimate the mean function of the on-line profiles. When nonparametric regression is used, it is critical to determine the bandwidth of estimation. In this paper, the default cross validation (CV) method is used to determine the bandwidth. The effect of bandwidth on the ARL performance of the proposed control charts worths further investigation. Also note that the smoothing spline is used here. Other nonparametric regression methods can be used in a similare manner.

References

Aly, A. A., Mahmoud, M. A., and Woodall, W. H. (2015). “A comparison of the performance of phase II simple linear profile control charts when parameters are estimated", *Communications in Statistics-Simulation and Computation*, 44(6), 1432-1440.

Agresti, A. (2002). *Categorical Data Analysis.* 2nd Edition, John Wiley & Sons, Inc., New York.

Cai, Z., Fan, J., and Li, R. (2000). “Efficient Estimation and Inference for Varying-Coefficient Models", *Journal of the American Statistical Association*, 95, 888-902.

Casella, G., and Berger, R. (2002). *Statistical Inference.* 2nd Edition, Duxbury Press Belmont, California.

Ding, Y., Zeng, L., and Zhou S. (2006). “Phase I Analysis for Monitoring Nonlinear Profiles in Manufacturing Processes", *Journal of Quality Technology*, 38, 199-216.

Fan, J., Zhang, C., and Zhang, J. (2001). “Generalized Likelihood Ratio Statistics and Wilks Phenomenon", *The Annals of Statistics*, 29, 153-193.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer-Verlag.

Kang, L., and Albin, S. L. (2000). “On-Line Monitoring When the Process Yields a Linear Profile", *Journal of Quality Technology*, 32, 418-426.

Kim, K., Mahmoud, M.A., and Woodall, W.H. (2003). “On the Monitoring of Linear Profiles", *Journal of Quality Technology*, 35, 317-328.

Knoth, S. (2005). “Accurate ARL computation for EWMA-S^2 control charts", *Statistics and Computing*, 15, 341-352.

Mahmoud, M.A. and Woodall, W.H. (2004). “Phase I Analysis of Linear Profiles with Calibration Applications", *Technometrics*, 46, 377-391.

cCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* 2nd Edition, Chapman & Hall, London.

Murphy, B.J. (1987). “Selecting Out of Control Variables With the T^2 Multivariate Quality Control Procedure", *The Statistician*, 36-5, 571-581.

Mahmoud, M.A., Parker, P.A., Woodall, W.H., and Hawkins, D.M. (2007). “A Change Point Method for Linear Profile Data", *Quality and Reliability Engineering International*, 23(2), 247-268.

Montgomery, D.C. (2005). *Introduction to Statistical Quality Control*. 5th Edition, John Wiley & Sons, Inc., New York.

Paynabar, K., Zou, C., and Qiu, P. (2016). “A Change-Point Approach for Phase-I Analysis in Multivariate Profile Monitoring and Diagnosis", *Technometrics*, 58(2), 191-204.

Rice, J. (1995). *Mathematical Statistics and Data Analysis*. 2nd Edition, Duxbury Press Belmont, California.

Wang, W and Lin, Dennis K.J. (2016). “Another Look at Run Length Distributions"., under revision for *Statistica Sinica*.

Woodall, W.H., Spitzner, D.J., Montgomery, D.C., and Gupta, S. (2004). “Using Control Charts to Monitor Process and Product Profiles", *Journal of Quality Technology*, 36, 309-320.

Xu, L., Wang, S., Peng, Y., Morgan, J. P., Reynolds Jr, M. R., and Woodall, W. H. (2012). “The monitoring of linear profiles with a GLR control chart". *Journal of Quality Technology*, 44(4), 348.

Yeh, A.B., Huwang, L.C., and Wu, Y.F (2004). “A Likelihood-Ratio-Based EWMA Control Chart for Monitoring Variability of Multivariate Normal Processes", *IIE Transactions*, 36, 865-879.

Zhang, J., Li, Z., and Wang, Z. (2009). “Control chart based on likelihood ratio for monitoring linear profiles". *Computational statistics and data analysis*, 53(4), 1440-1448.

Zhang, J., Zou, C., and Wang, Z. (2010). “A control chart based on likelihood ratio test for monitoring process mean and variability". *Quality and Reliability Engineering International*, 26(1), 63-73.

hu, J. and Lin, D.K.J. (2009). “Monitoring the Slopes of Linear Profiles", *Quality Engineering*, 22(1), 1-12.

Zou, C., Zhang, Y. and Wang, Z. (2006). “A Control Chart Based on a Change-Point Model for Monitoring Linear Profiles", *IIE Transactions*, 38, 1093-1103.

Zou, C., Qiu, P., and Hawkins, D. (2009). “Nonparametric control chart for monitoring profiles using change point formulation and adaptive smoothing". *Statistica Sinica*, 1337-1357.

Appendix

Estimation of the UCL for EWMA-type GLRT Chart

We assume the reader is familiar with R, a free statistical analysis software. Almost everything related with R can be found on its official website at http://www.r-project.org/ or on CRAN (The Comprehensive R Archive Network) at http://lib.stat.cmu.edu/R/CRAN/.

The spc package contributed by Sven Knoth is downloadable at the CRAN website http://lib.stat.cmu.edu/R/CRAN/web/packages/spc/index.html. In this package Dr. Knoth provides many useful functions in SPC area such as calculating the ARL for a corresponding UCL or vise versa, for EWMA or CUSUM control chart in monitoring process means or variances. The specific function in spc package that needs to be used to calculate UCL for EWMA-type GLRT control chart is called `sewma.crit`. To install the spc package, one simply type the command "`install.packages("spc")`" and then choose a CRAN mirror site. The install process will be finished automatically.

The calculation of the UCL using approximation method can be done using the following command lines:

```
> library(spc)  ## Load the spc package
> theta <- 0.2  ## This is the smooth parameter in EWMA chart
> ARL.IC <- 1/alpha  ## This the in-control ARL Value
> parm <-  19.57086    ## This is the estimated mean of GLRT statistic
> UCL.EWMA <- sewma.crit(l=theta, L0=ARL.IC, sigma0=sqrt(parm),
                         cu=NULL,hs=parm,df=parm)[[2]]
> UCL.EWMA
[1] 21.89268
```

The first line load the spc package; the next two lines give values of the θ (smoothing parameter) value and the in-control ARL value; then let the *parm* equal to the estimated mean value of the GLRT statistic (obtained by parametric bootstrap Monte-Carlo simulation); finally use the `sewma.crit` function to calculate the desired UCL corresponding to the in-control ARL.

Challenges in Monitoring Non-Stationary Time Series

Taras Lazariv and Wolfgang Schmid

Abstract In this paper different approaches for monitoring non-stationary processes are discussed. Despite 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, e.g., as well shifts as 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 attempt are used to obtain charts without reference parameters. All introduced schemes are compared with each other assuming that a univariate unit root process with drift is present. We make use of several performance measures of control charts like the average run length, the worst average delay and the limit average delay. Furthermore it is analyzed how sensitive the charts with reference value react on the choice of this quantity.

Key words: control charts; statistical process control; change-point detection; time series; state-space model

Taras Lazariv
Department of Statistics, European University Viadrina, PO Box 1786, 15207 Frankfurt(Oder), Germany, e-mail: lazariv@europa-uni.de

Wolfgang Schmid
Department of Statistics, European University Viadrina, PO Box 1786, 15207 Frankfurt(Oder), Germany, e-mail: schmid@europa-uni.de

1 Introduction

In the last 30 years monitoring problems have been discussed in many areas like, e.g., in economics (Frisén (2008)), medicine (Kass-Hout and Zhang (2010)), environmental sciences (Chou (2004)). It has turned out that there are many further topics beyond engineering, the original field of applications (e.g., Montgomery (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 these papers the underlying time series is assumed to be (weakly) stationary in the in-control state. Nowadays, it is mostly distinguished between residual charts and modified schemes in literature. Residual charts are based on the idea to transform the original data such 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 are making 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 among others by Alwan and Roberts (1988), Wardell et al (1994a,b), Lu and Reynolds (1999), modified charts are subject of, e.g., Nikiforov (1975) and Schmid (1995, 1997a,b).

In many applications, however, especially in economics, frequently the process of interest turns out to be close to non-stationarity or it is even non-stationary. It is not oscillating around a common mean or its variance and autocovariances are changing over time, respectively. The existing techniques fail while 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 up to now. 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 on the underlying in-control non-stationary process is given.

Schmid and Steland (2000) applied nonparametric kernel control charts to a non-stationary process to analyze 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 for modeling the underlying in-control process. These processes are very flexible and allow the modeling of a large family of non-stationary processes. Several control charts for detecting a mean shift were derived.

In the current paper we want to discuss various techniques for deriving control charts for non-stationary processes. Nonparametric techniques are not considered. One approach is based on differencing, i.e. the original data are transformed by successively calculating the differences of two successive observations. This procedure is applied until the resulting process is stationary. Such an approach is frequently

applied in econometrics and related to monitoring it has been studied among others by Steland (2005, 2007) 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. Here we consider this attempt as well. As in Lazariv and Schmid (2015) state-space models are 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 paper 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 attempts to handle 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)). This approach is briefly described in the next section. Another possibility is of course to directly model the non-stationary process by a suitable model. Here state-space models are frequently applied since they are on the one side very flexible and on the other side there exist computational techniques such that the statistical analysis of these processes can be done in fast time.

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)). This means in discrete time that it holds

$$Y_t = Y_{t-1} + \varepsilon_t, \quad t \geq 1 \tag{1}$$

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

$$X_t = \begin{cases} Y_t & \text{for} \quad 1 \leq t < \tau \\ Y_t + (t - \tau + 1)a & \text{for} \quad t \geq \tau \end{cases}. \tag{2}$$

Here $\{X_t\}$ denotes the observed process and $\{Y_t\}$ the target (in-control) process. 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 and 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 a behaviour is presented in Figure 1 where the daily closing prices of Facebook from May 18, 2012 to May 29, 2016 are plotted. The blue line shows the possible in-control behaviour.

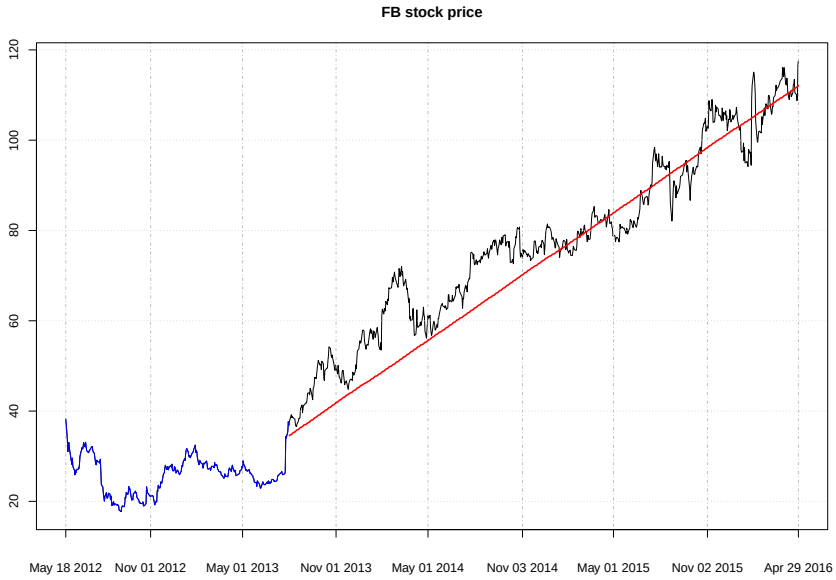


Fig. 1: Facebook share price with estimated trend ($\hat{a} = 0.1171$ in the out-of-control period).

2.2 State-Space Models

State-space models have been widely used in engineering. In recent years more applications in economics can be 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 following set of equations

$$\boldsymbol{Y}_t = \boldsymbol{G}_t \boldsymbol{S}_t + \boldsymbol{W}_t, \quad t = 1, 2, \ldots, \qquad \text{where} \tag{4a}$$

$$\boldsymbol{S}_{t+1} = \boldsymbol{F}_t \boldsymbol{S}_t + \boldsymbol{V}_t, \quad t = 1, 2, \ldots. \tag{4b}$$

The equation (4a) is called 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 that

(A1) Let for all $t \geq 1$

$$E\begin{pmatrix}\boldsymbol{V}_t \\ \boldsymbol{W}_t\end{pmatrix} = \boldsymbol{0}, \;\; E(\boldsymbol{V}_t \boldsymbol{V}_t') = \boldsymbol{Q}_t, \;\; E(\boldsymbol{W}_t \boldsymbol{W}_t') = \boldsymbol{R}_t, \;\; E(\boldsymbol{V}_t \boldsymbol{W}_t') = \boldsymbol{U}_t.$$

$\{\boldsymbol{Q}_t\}$, $\{\boldsymbol{R}_t\}$, and $\{\boldsymbol{U}_t\}$ are specified sequences of $q \times q$, $p \times p$ and $q \times p$ matrices, respectively.

(A2) Let $\boldsymbol{S}_1, (\boldsymbol{V}_1', \boldsymbol{W}_1')', (\boldsymbol{V}_2', \boldsymbol{W}_2')', \ldots$ be uncorrelated.

(A3) Let $E(\boldsymbol{Y}_0 \boldsymbol{V}_t') = \boldsymbol{0}$ and $E(\boldsymbol{Y}_0 \boldsymbol{W}_t') = \boldsymbol{0}$ for all $t \geq 1$.

The parameter matrices F_t, G_t, Q_t, R_t, 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 $\boldsymbol{Y}_0$,..., $\boldsymbol{Y}_{t-1}$ 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 i-th 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) and Durbin and Koopman (2012)).

$\hat{\boldsymbol{S}}_{t+1}$ can be rewritten as a linear combination of $Y_0, \ldots, Y_t$

$$\hat{\boldsymbol{S}}_{t+1} = \sum_{j=1}^{t} \boldsymbol{A}_{t+1,j} \boldsymbol{Y}_j + \boldsymbol{a}_{t+1}(\boldsymbol{Y}_0) \tag{7}$$

with $\boldsymbol{A}_{t+1,j} = \left(\boldsymbol{E}_t \cdots \boldsymbol{E}_{j+1}\right) \boldsymbol{\Theta}_j \Delta_j^{-1}$ and $\boldsymbol{E}_t = \boldsymbol{F}_t - \boldsymbol{\Theta}_t \Delta_t^{-1} \boldsymbol{G}_t$. $\boldsymbol{a}_{t+1}(\boldsymbol{Y}_0) = (\boldsymbol{E}_t \cdots \boldsymbol{E}_1)\, \hat{\boldsymbol{S}}_1$ is a function of $\boldsymbol{Y}_0$.

Similarly can be obtained the best linear predictor $\hat{\boldsymbol{Y}}_t$ of $\boldsymbol{Y}_t$ given $\boldsymbol{Y}_0$,..., $\boldsymbol{Y}_{t-1}$, 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 it holds that for $t \geq 1$

$$\boldsymbol{\Sigma}_t = \boldsymbol{G}_t \boldsymbol{\Omega}_t \boldsymbol{G}_t' + \boldsymbol{R}_t.$$

In the paper we assume that the parameters of the target process are known. In practice, however, they should be estimated using historical data. The influence of parameter estimation is an important question, but we will not discuss it in the present paper.

2.3 Modeling the Out-of-Control Process

In the following we want to consider a more general change-point model

$$X_t = \begin{cases} Y_t & \text{for} \quad 1 \leq t < \tau \\ Y_t + D_{t,\tau} a & \text{for} \quad t \geq \tau \end{cases}, \tag{9}$$

where $D_{t,\tau}$ denotes a known $p \times p$ matrix and $a \in \mathbb{R}^p$ an unknown parameter vector. Choosing $D_{t,\tau} = (t - \tau + 1)I$ we obtain the above drift model and setting $D_{t,\tau} = I$ a mean shift model is obtained. Here I stands for the $p \times p$ unity matrix. Of course it is also possible to take the standard deviation of the process into account. Then, e.g., we have to choose $D_{t,\tau} = diag(\sqrt{Var(Y_{t1})}, \ldots \sqrt{Var(Y_{tp})})$ for the shift model (see

Lazariv and Schmid (2015)). Of course it is also possible that there are components with a drift and others with a shift what can be handled by the approach as well.

3 Control Charts for Non-Stationary Processes

There are different approaches to derive control charts for non-stationary processes. The easiest attempt is to transform the original data such that the transformed data follow a stationary process. Then all well-known procedures for stationary processes can be applied to the transformed quantities. This method is briefly described in the next section. In Section 3.2 and Section 3.3 we introduce control charts for the generalized change point model assuming the in–control process is a state-space model. In Section 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 a which has to be replaced in practice by a suitable reference value. In Section 3.3 the generalized procedures are considered where the corresponding probability density is as well maximized over a so that the resulting chart does not depend on a.

3.1 The Transformation Approach

The transformation method works similar than 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 to detect a shift in a stationary problem which has been intensively discussed in 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 is 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 reasonable 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 $D_t = diag(\sqrt{Var(Y_{t1})}, ...\sqrt{Var(Y_{tp})})$ in Lazariv and Schmid (2015) by an arbitrary known matrix $D_{t,\tau}$, $t \geq \tau$, it is possible to obtain control charts for further out-of-control situations like, 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 additionally

($A2^*$) Let $\boldsymbol{S}_1$, $(\boldsymbol{V}_1', \boldsymbol{W}_1')'$, $(\boldsymbol{V}_2', \boldsymbol{W}_2')'$,... be independent.
(A4) Let $\boldsymbol{S}_1$, $(\boldsymbol{V}_1', \boldsymbol{W}_1')'$, $(\boldsymbol{V}_2', \boldsymbol{W}_2')'$,... be normally distributed.
(A5) Let Σ_t have a full rank for all $t \geq 1$.
For more details we refer to Lazariv and Schmid (2015).

Let us rewrite the densities of $\boldsymbol{X}_1$,..., $\boldsymbol{X}_n$ 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 ($1 \leq \tau \leq n$).

For detailed derivation of the control statistic follows similar as 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 a.

3.2.2 The Sequential Probability Ratio Chart

The sequential probability ratio test (SPRT) was introduced by Wald (1947). It was used by Page (1954) to derive a mean chart for independent samples. Lazariv and Schmid (2015) derived a 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 above a^* is a reference value of the unknown parameter a.

Note that the control statistic can be recursively calculated what dramatically simplifies its determination.

3.2.3 The Shiryaev-Roberts Chart

In this section we present the control chart based on a 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 a priori choice of these quantities. The optimal choice depends on the unknown quantities of the out-of-control model like, e.g., the size of the shift. Since frequently such information is not available the choice of the reference value is sometimes like a lottery. For that reasons 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

$$\sup_{\boldsymbol{a} \neq 0} \log f_\tau(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \longrightarrow \max,$$

is considered where the likelihood is maximized over all possible sizes of the shift.

This is the idea behind the Generalized LR (GLR), Generalized SPRT (GSPRT) and Generalized Modified SR (GMSR) schemes. The details are presented below. The derivation of the charts follows with the same arguments as in Lazariv and Schmid (2015)) where, however, the quantity D_t must be replaced by $D_{t,\tau}$. 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

$$N_{GLR}(c) = \inf\{n \in \mathbb{N} : \max\{0, \max_{1 \le i \le n}\left(-\sum_{t=i}^{n}(\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t + \frac{1}{2}\boldsymbol{M}_{t,i}\tilde{\boldsymbol{a}}_{i,n})'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{M}_{t,i}\tilde{\boldsymbol{a}}_{i,n}\right)\} > c\}, \tag{17}$$

where $\tilde{a}_{\tau,n}$ 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 that case the run length is obtained as

$$N_{GSPRT}(c) = \inf\left\{n \in \mathbb{N} : \max_{0 \le i \le n}\left(g_{n;GSPRT} - g_{i;GSPRT}\right) > c\right\} \tag{18}$$

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 exponential sums must be calculated. In order to avoid this problem we consider 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 ARL of all charts is compared with each other. In our study the reference value a^* takes values within the set $\{0.5, 1.0, ..., 3.0\}$. Moreover, an EWMA chart is applied to the first differences. The possible values of the smoothing parameter are lying in the set $\{0.1, 0.2, ..., 1.0\}$. Since there is no explicit formula for the ARL available it is estimated within 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.

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 1: ARLs for all control charts

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 approach 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 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.

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

Table 2: Average delays of all control charts

In literature mostly the limit of the average delay for $\tau \to \infty$ and the worst average delay over all τ are taken as performance measures. A further analysis shows that except 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 different since the average delay is increasing with τ and the results are worse. The chart seems to favor changes at the beginning but has problems to detect 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 scheme which are slightly better than the EWMA approach. 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 directly smaller (above) or larger (below) than the value leading to the minimum ARL. Note that in our analysis we have chosen $a^* \in \{0.5, 1.0, ..., 3.0\}$ 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 percent worth, above). For the choice $a^* = 2.0$ we obtain $ARL = 5.04$ (4.50 percent worth, below).

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%
	+6.63%	+5.98%	+1.29%	–	–	–	+1.13%
1.5	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%

Table 3: Influence of the wrong choice of reference parameter, for all control charts

The table shows that the charts react different on the choice of a^*. Nevertheless, the out-of-control ARLS are in all cases smaller than those of the GLR and the GMSR chart. 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 chart are very bad. Assuming $a = 0.5$ the worst ARL for the SPRT (EWMA) chart is 87.54 (115.11) and 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 that case.

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 considered the maximization of the sum of the logarithms of the likelihoods. This may lead to the deterioration. Moreover, the GSPRT scheme must be carefully applied since it favours changes at the beginning and it has huge problems to detect 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 carefully done 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 paper and check if the same issue arises for the present change-point model (9).

For this purpose we have calculated a table of frequencies, namely that the in-control run length will fall into certain intervals (see Table 4). The table shows the relative frequencies (in percent) 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 $P(5000 \leq N(c) \leq 10000)$. The results are based on simulating 10^5 independent random samples of a unit root process.

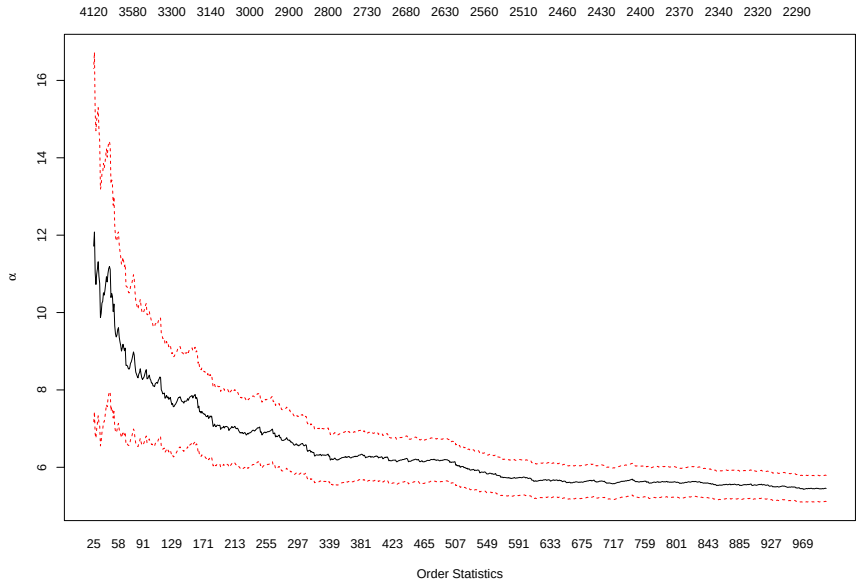
i	LR	SPRT	SR	GLR
1	0.00	0.00	0.00	0.01
2	0.00	0.01	0.00	0.04
3	0.05	0.03	0.00	0.05
4	0.09	0.07	0.00	0.09
5	0.07	0.11	0.02	0.05
...	...	...	...	...
[1000, 2000]	11.15	11.56	11.54	11.14
[2000, 3000]	1.57	1.53	1.76	1.10
[3000, 4000]	0.21	0.20	0.20	0.09
[4000, 5000]	0.01	0.03	0.00	0.00
[5000, 10000]	0.00	0.00	0.00	0.00

Table 4: Distributions of the in-control run lengths of the considered charts

The table shows that there is no evidence of heavy tails. The Hill and the Pickands plot (see, e.g., Resnick (2007)) are presented for the run length of the SPRT chart in order to analyze the tail behavior and to check the existence of the expectation of the run length. The run length is estimated within 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 to 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 paper 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 on the structure of the control statistics shows that the matrix $M_{t,\tau}$ heavily influences the control statistics of all 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 $M_{t,\tau}$ tends to 0 as a function of t and for fixed τ because of the quantity D_t. D_t models the variance of the target process in the univariate case and it



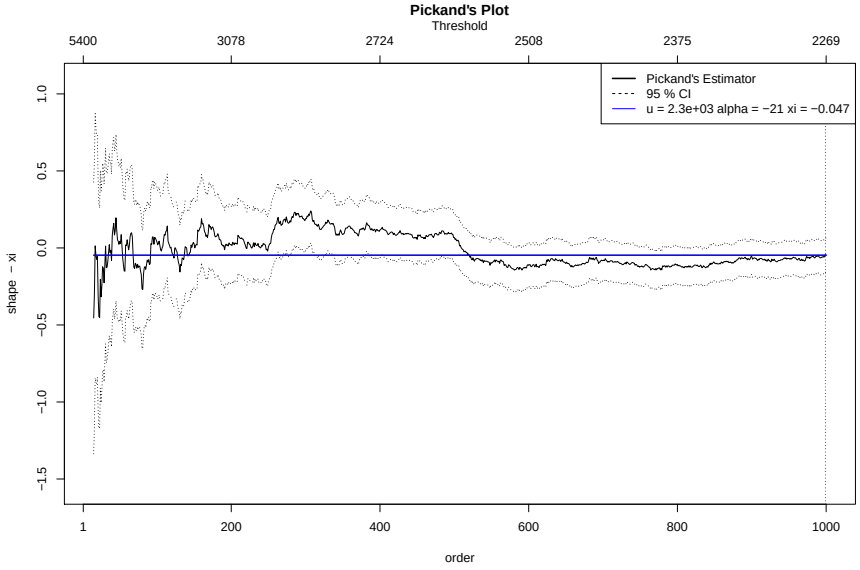


Fig. 2: Hill plot (above) and Pickands plot (below)

seems to tend to infinity for a target process as in Lazariv and Schmid (2015). In this paper, however, the quantity $D_{t,\tau}$ depends on τ as well and it holds that $M_{t,\tau} = -1$, i.e. it is constant.

6 Summary

In the present paper we discuss different attempts for monitoring non-stationary processes. We consider 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 of this procedure is that it only works for special type of processes like, 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 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 are derived. Applying the generalized likelihood ratio, the generalized sequential probability ratio and the generalized modified Shiryaev-Roberts procedure control schemes without reference values are obtained.

All charts are 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 are used to evaluate the introduced charts. Despite the average run length, the worst average delay and the limit of the average delay are considered. Moreover, it is analyzed, how the charts with a reference value react if not the optimal reference value leading to the smallest ARL is used but another, which is close to the optimal one or more far away. It is shown that the LR and the SPRT chart should be favored if some knowledges on the expected drift are given. Else, 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 the attempt of Lazariv and Schmid (2015). Using the Hill plot and the Pickands plot the tails of the run lengths of the introduced charts is analyzed and it is concluded that in the present case the average run length exists. The reason for the different behaviour lies in the consideration of another out-of-control model.

References

Alwan LC, Roberts HV (1988) Time-series modeling for statistical process control. Journal of Business & Economic Statistics 6(1):87–95

Brockwell PJ, Davis RA (2009) Time series: theory and methods. Springer

Chou CJ (2004) Groundwater monitoring: statistical methods for testing special background conditions. In: Wiersma GB (ed) Environmental monitoring, CRC press

Durbin J, Koopman SJ (2012) Time series analysis by state space methods. Oxford University Press

Frisén M (2008) Financial surveillance, vol 71. John Wiley & Sons

Hayashi F (2000) Econometrics. Princeton University Press

Kass-Hout T, Zhang X (2010) Biosurveillance: Methods and case studies. CRC Press

Koopman SJ (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 CW, Reynolds M (1999) Control charts for monitoring the mean and variance of autocorrelated processes. Journal of Quality Technology 31(3):259–274

Montgomery DC (2009) Introduction to statistical quality control, 6th edn. 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 pp 100–115

Resnick SI (2007) Extreme values, regular variation, and point processes. Springer Science & Business Media

Roberts S (1966) A comparison of some control chart procedures. Technometrics 8(3):411–430

Ruppert D (2004) Statistics and finance: an introduction. Springer Science & Business Media

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, Springer, pp 115–137

Schmid W, Steland A (2000) Sequential control of non-stationary processes by nonparametric kernel control charts. Allgemeines Statistisches Archiv (Journal of the German Statistical Assoc) Vol 84:315–336

Shiryaev AN (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. Wiley

Wardell DG, Moskowitz H, Plante RD (1994a) Run-length distributions of residual control charts for autocorrelated processes. Journal of Quality Technology 26(4):308–317

Wardell DG, Moskowitz H, Plante RD (1994b) Run-length distributions of special-cause control charts for correlated processes. Technometrics 36(1):3–17

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 paper, we describe the R package `dfphase1` which provides an implementation of many of recently suggested Phase I distribution-free methods. Indeed, becouse 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.

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

Giovanna Capizzi
Department of Statistical Sciences, University of Padua, Italy, e-mail: giovanna.capizzi@unipd.it

Guido Masarotto
Department of Statistical Sciences, University of Padua, Italy, e-mail: guido.masarotto@unipd.it

Preliminary version.

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 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 (on 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 becouse 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 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, Woodal, 2016).

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 practitioners 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 paper 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. All methods implemented in the package attain a desired false alarm probability (FAP) with no assumption on the underlying probability distribution of quality characteristics.

The paper is organized as follows. In Section 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 Section 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 Section 4. In Section 5, an example is discussed. Some concluding remarks are given in Section 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". 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 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 behaviour. 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, Chakraborti, 2011, for some reviews covering much of the recent SPM nonparametric literature).

Two possibile 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 the possible non-normality of the process distribution. According to this approach, the distribution-free design of control charts doe 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 since it is able to exactly achieve a prescribed FAP regardless of the underlying process distribution, at least for indipendent 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 showed 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).

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) chart, adapted also 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 *spatial signs or ranks* (Oja, 2010).
      (iv) Balanced combination of the previous Shewhart-type schemes.
   b. *Methods for change-point detection*
      (i) Sullivan and Woodall (2000) control chart, adapted also 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) or spatial signs or ranks (Oja, 2010).

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 Rcpp interface (Eddelbuettel, 2013).

Control statistics in Table 1 can be based on several estimates of the common process parameters. For example, as illustrated in Section 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 addresses the standard univariate and multivariate framework. However, as shown in Section 5, it can also be used in more complex situations where the monitored variables can be suitable features "extracted" from the original data,

such as principal components, model parameters, etc. Nevertheless, in order to guarantee the validy 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.,

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

where FAP_0 is a desired value of the FAP.

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

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

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.

5 An Example

5.1 Description of the data

The package can be loaded during an `R` session using

```
> library(dfphase1)
```

To illustrate its use, 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 the 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 1, showing the spectra collected during the first day, has been obtained with the following commands.

```
> wavelength <- rep(seq(900,1700,by=2),rep(12,401))
> samples <- reorder(rep(1:12,401),
+                    rep(c(9:12,5:8,1:4),401))
> xyplot(NIR[1:12,]~wavelength|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 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 following "regression-like" representation of the ith NIR spectrum

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

where $x_{i,j}$ is the jth principal component, and $\text{NIR}_i(nm)$, $\mu(nm)$, $\xi_j(nm)$ and $r_i(nm)$ are the logarithm of the absorbance, the log-absorbance mean, the jth eigenvector and the residual term for the nmth wavelength. Hence, because for spectra data, such as the gasoline spectra, the eigenvalues are relatively smooth functions (see Figure 3), 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.

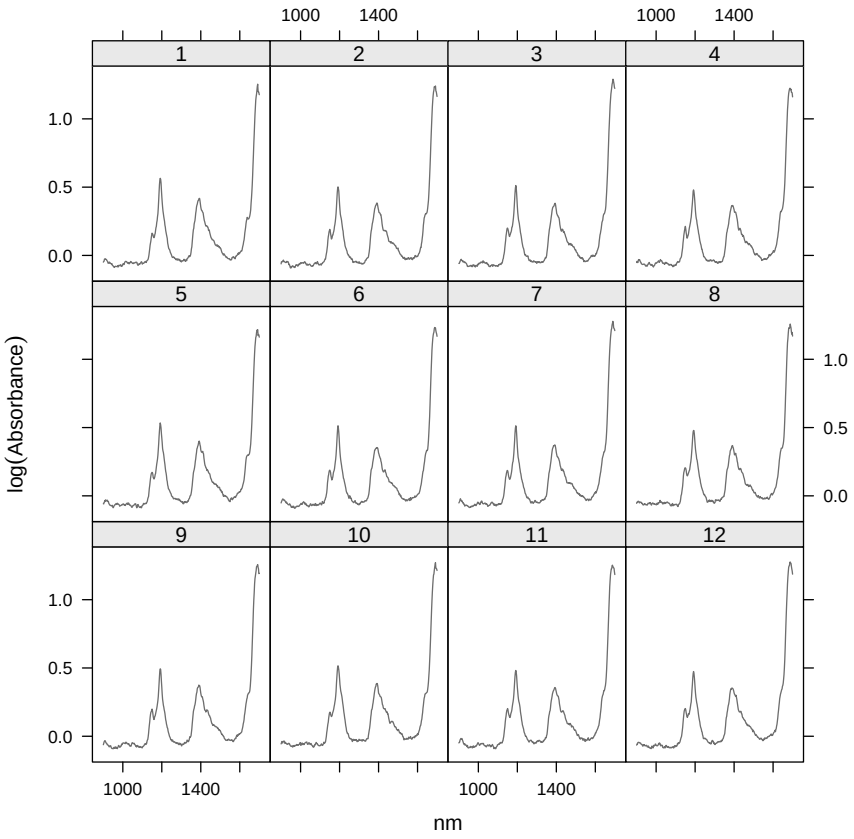


Fig. 1: The gasoline NIR spectra collected during the first day.

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 mantain their distribution-free properties.

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

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

and displayed in Figure 2, suggests to retain the first 4 principal components. Figure 3, displaying the corresponding eigenvectors, can be obained using the following commands.

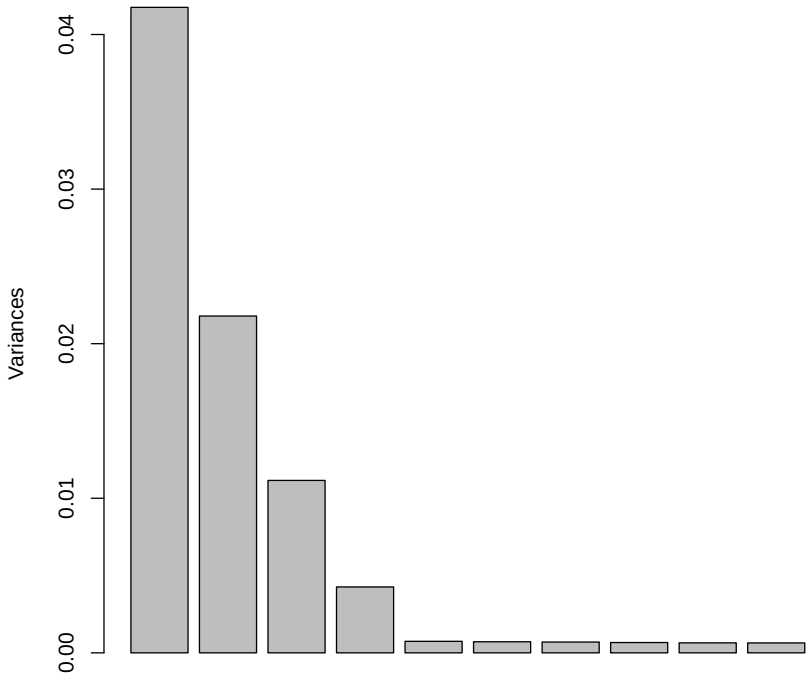


Fig. 2: Scree plot of the gasoline NIR spectra.

```
> wavelength <- rep(seq(900,1700,by=2),4)
> eigv <- gl(4,401,labels=4:1)
> xyplot(pca$rotation[,1:4]~wavelength|eigv,type="l",
+        xlab="nm",ylab=expression(log(Absorbance)),
+        layout=c(1,4))
```

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

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

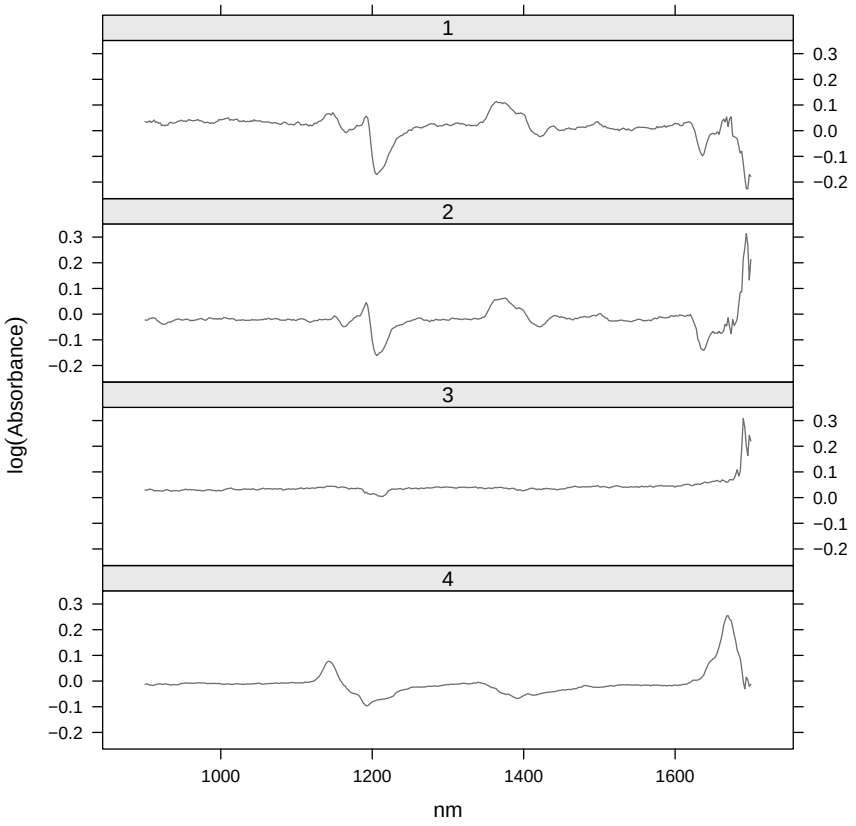


Fig. 3: First four eigenvectors of the gasoline near-infrared spectra.

which reflects the size of the residual term. The following code "extracts" the first 4 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)
```

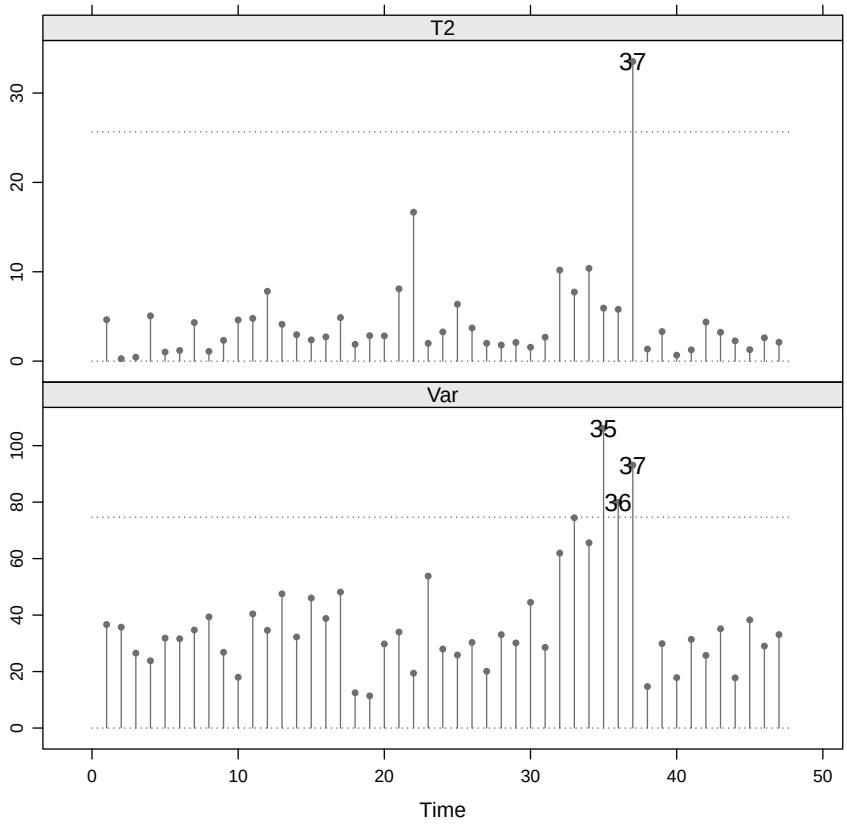


Fig. 4: Combination of an Hotelling T^2 control chart (upper panel) and a control chart for monitoring the stability of the covariance matrix (lower panel).

5.2 Phase I analysis

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 Figure 4. 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] 25.65993 74.66964
```

are computed by permutation so that the desired FAP is guaranteed for every multivariate distribution.

In `dfphase1`, the default value of the FAP is 5%, but, as illustrated in the following, it can be easily changed by users using the `FAP` argument.

Figure 4 suggests that the dispersion was probably OC during days 35, 36 and 37. Then, observe that control statistics of days 32, 33 and 34 are below the limit but larger than the values of the other "in-control" days (see the lower panel). This fact is even more evident replacing the standard estimates of multivariate location and dispersion (e.g. Montgomery, 2009, equations 11.17a-c) with the high-breakdown MCD estimates (e.g. Maronna et al, 2006, chapter 6). Indeed, Figures 6 and 7 show that, when the observations collected on days 35, 36 and 37 are deleted, days 32, 33 and 34 are flagged as OC for the dispersion.

In `dfphase1`, alternative estimates of process parameters can be used adding the optional argument `loc.scatter` in the call to `mshewhart`. The output is illustrated in Figure 5.

```
> mshewhart(x,loc.scatter="MCD")
```

Figures 6 and 7 are produced using the following commands.

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

As illustrated by the following code, multivariate Shewhart-type control charts based on the spatial signs or ranks can be obtained using the `score` argument.

```
> mshewhart(x,score="Spatial Signs")
> mshewhart(x,score="Spatial Ranks")
```

Results, displayed in Figures 8 and 9, show that the control statistics based on these two transformations fail to detect any OC situation.

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 tested applying separately one or more univariate Phase I control charts to the five variables. For reason of space, we will show only 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. The command

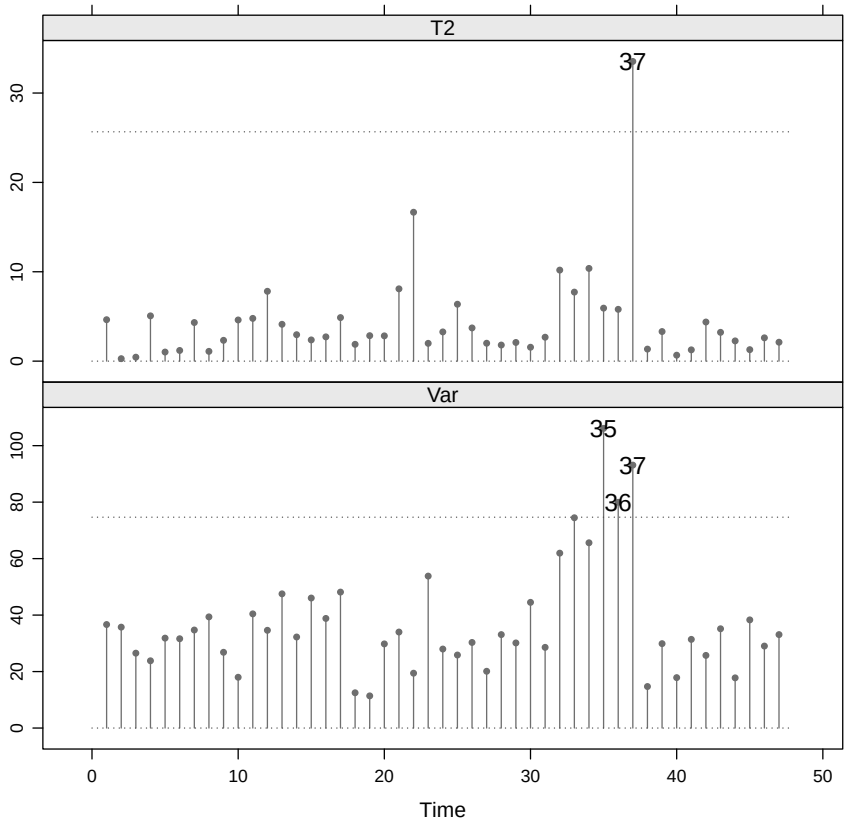


Fig. 5: Combination of an Hotelling T^2 control chart (upper panel) and a control chart for monitoring the stability of the covariance matrix (lower panel) based on the MCD estimates.

```
> shewhart(x[2,,],FAP=0.01)
```

produces Figure 10 displaying the combined $\overline{X} - S$ control chart with limits computed by permutation.

The `stat` argument of the `dfphase1` function can be used to select the control statistic[s]. For example, the command

```
> shewhart(x[2,,],stat="Rank")
```

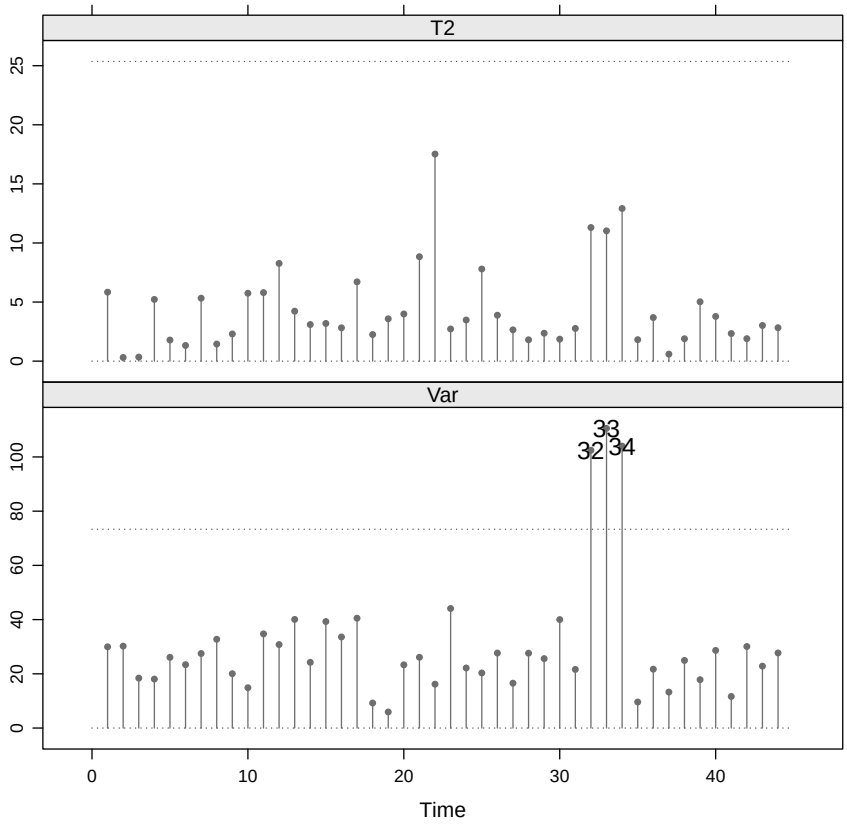


Fig. 6: Combination of an Hotelling T^2 control chart (upper panel) and a control chart for monitoring the stability of the covariance matrix (lower panel): days 35, 36 and 37 have been deleted.

"asks" for 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. The corresponding output is shown in Figure 11.

The `rsp` function implements the RS/P method suggested by Capizzi and Masarotto (2013b). Figure 12 can be obtained with the following command.

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

Figures 10-12 indicate an increased variability of 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).

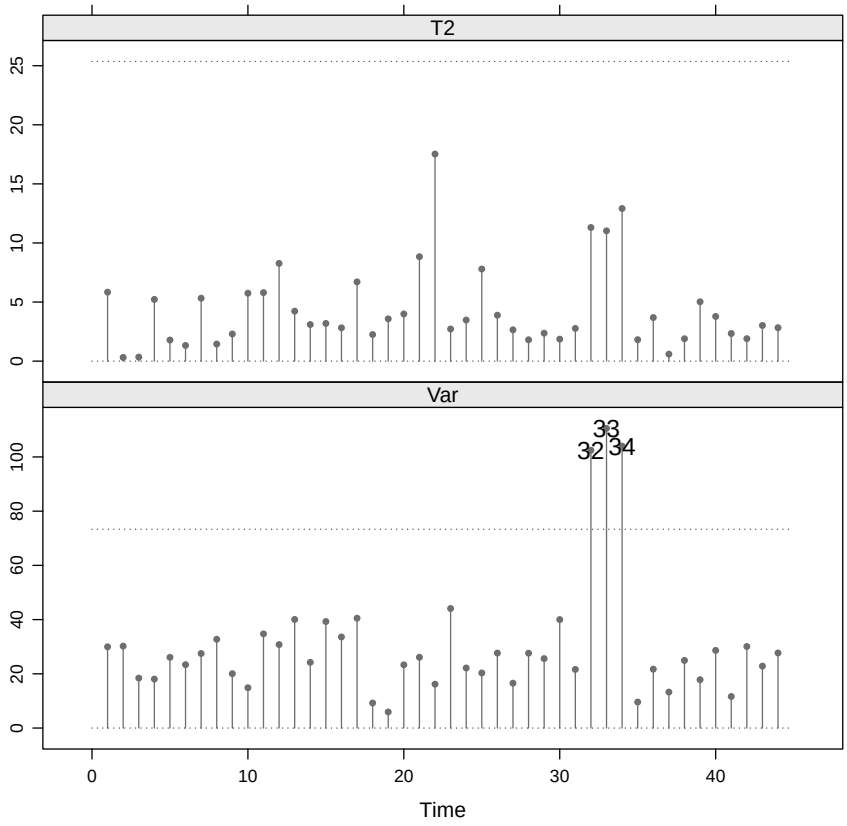


Fig. 7: Combination of an Hotelling T^2 control chart (upper panel) and a control chart for monitoring the stability of the covariance matrix (lower panel) based on the MCD estimates: days 35, 36 and 37 have been deleted.

Globally speaking, the application of univariate and multivariate control charts signals the presence of an OC condition in the interval [32;37]. The unstability 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 capability and design a Phase II control chart for prospectively monitoring the process.

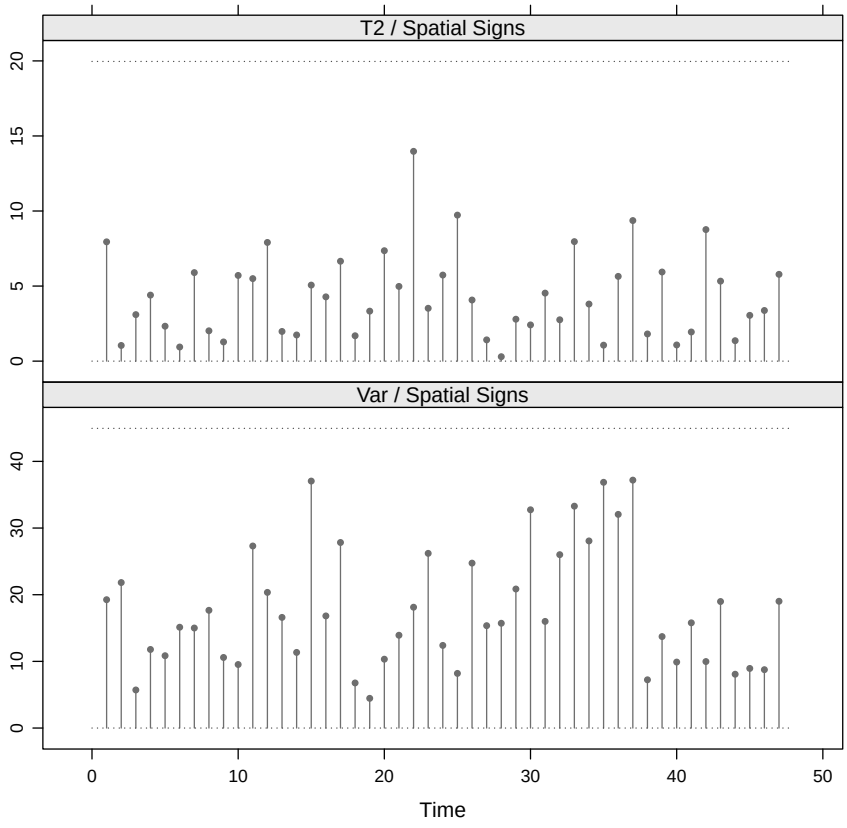


Fig. 8: Multivariate Shewhart control charts based on the spatial signs for testing for the stability of location (upper panel) and dispersion (lower panel) over time.

6 Conclusions

We have illustrated the structure and the use of an R package developed for the distribution-free Phase I analysis of univariate and multivariate data. After some additional testing, and the implementation of further methods, the package will be available from the *The Comprehensive R Archive Network* (CRAN).

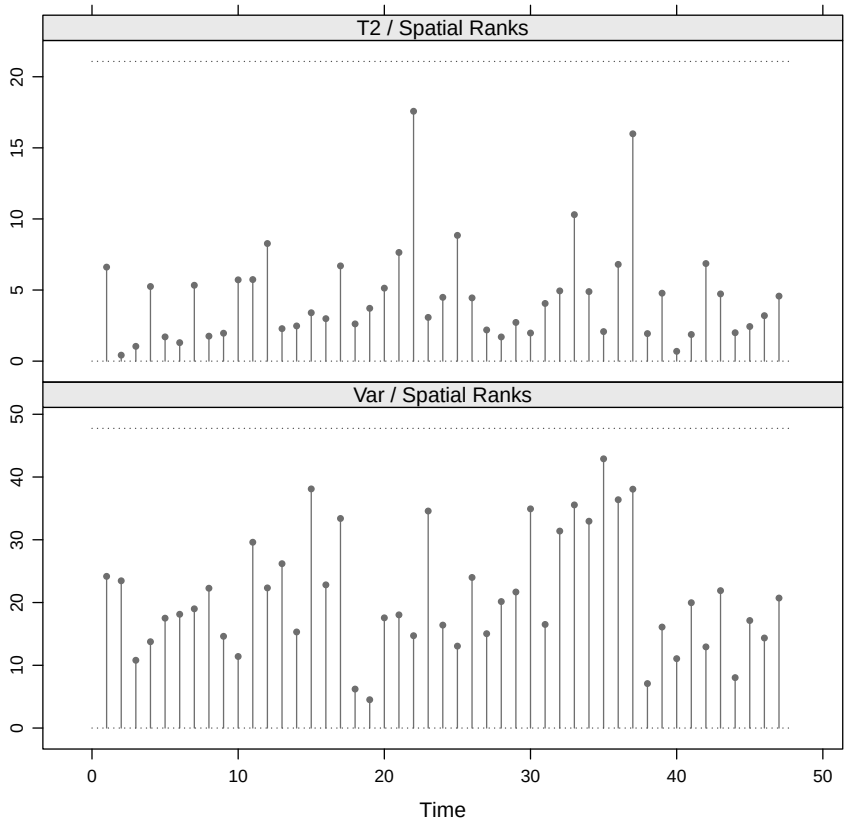


Fig. 9: Multivariate Shewhart control charts based on the spatial ranks for testing for the stability of location (upper panel) and dispersion (lower panel) over time.

References

Bell RC, Jones-Farmer LA, 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 (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

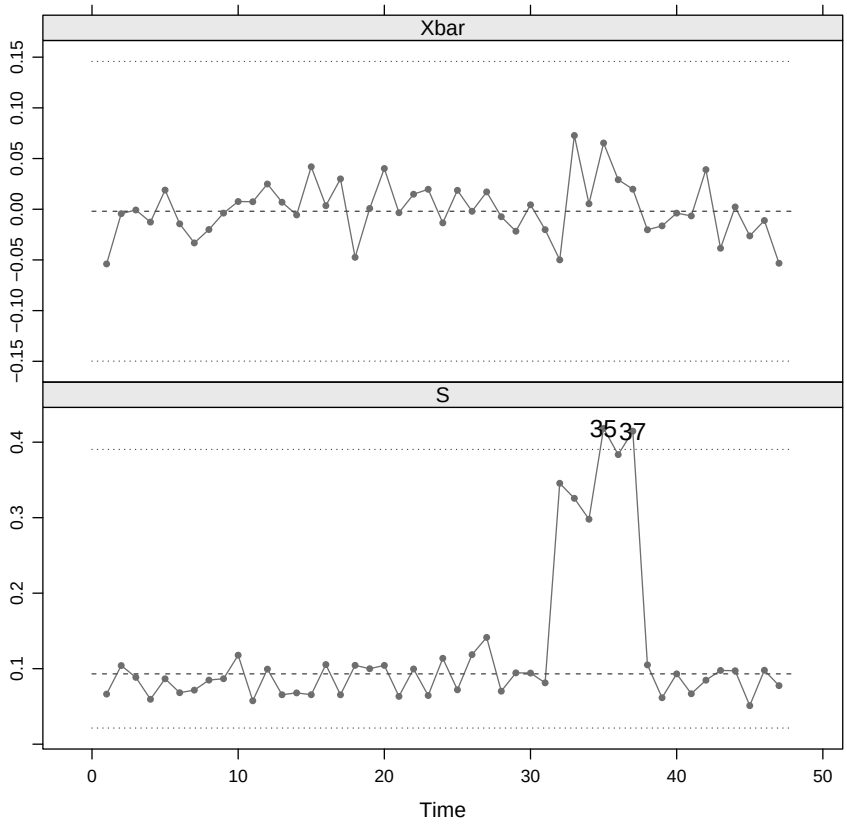


Fig. 10: Combined $\overline{X} - S$ control chart applied to the second principal component.

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

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

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

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

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

Cheng CR, Shiau JJH (2015) A distribution-free multivariate control chart for Phase I applications. Quality and Reliability Engineering International 31:97–111

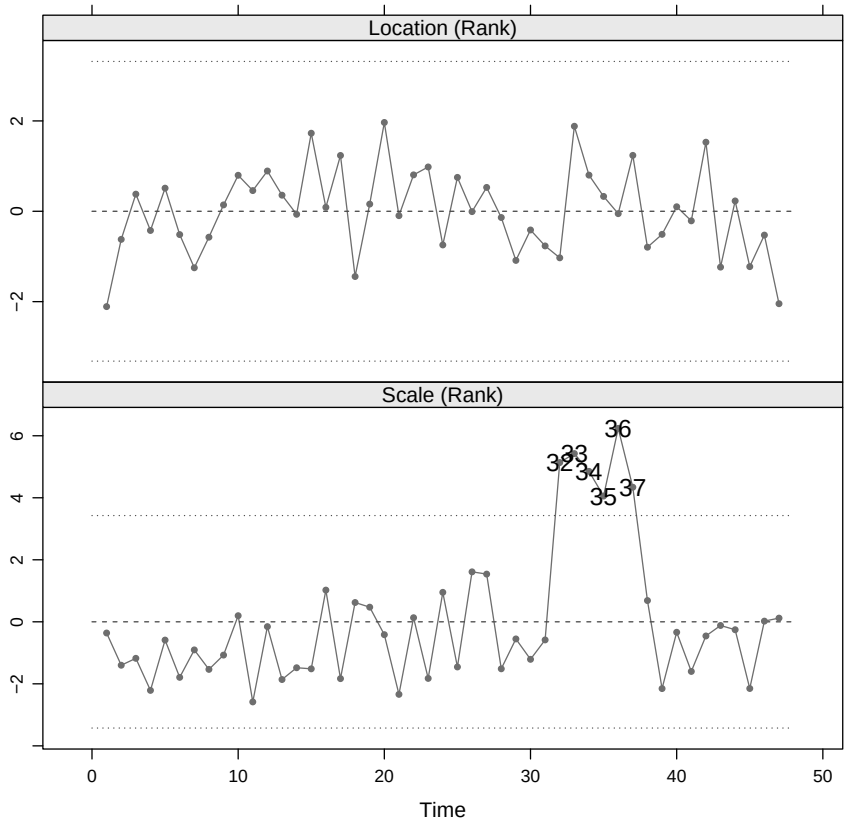


Fig. 11: Two univariate rank-based control charts applied to the second principal component.

Eddelbuettel D (2013) Seamless R and C++ Integration with Rcpp. Springer, New York, NY

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

Human SW, Chakraborti S, Smit CF (2010) Shewhart-type control charts for variation in Phase I data analysis. Computational Statistics & Data Analysis 54:863–874

Jensen WA, Birch JB, Woodall WH (2007) High breakdown estimation methods for Phase I multivariate control charts. Quality and Reliability Engineering International 23(5):615–629

Jones-Farmer LA, Champ CW (2010) A distribution-free Phase I control chart for subgroup scale. Journal of Quality Technology 42:373–387

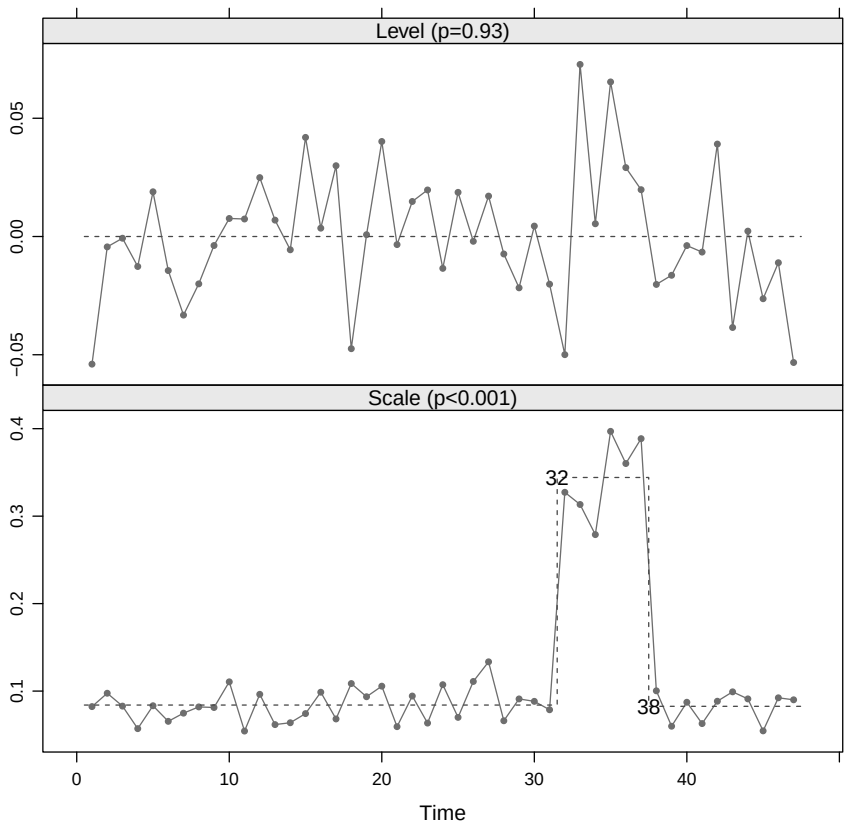


Fig. 12: Phase I analysis of the second principal component using the RS/P method.

Jones-Farmer LA, Woodall WH, Steiner SH, Champ CW (2014) An overview of Phase I analysis for process improvement and monitoring. Journal of Quality Technology 46:265–280

Jones-Farmer LA, Jordan V, Champ CW (2009) Distribution-free Phase I control charts for subgroup location. Journal of Quality Technology 41:304–316

Killick R, Eckley IA (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. URL https://CRAN.R-project.org/package=spc, R package version 0.5.3

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

Maronna RA, Martin RD, Yohai VJ (2006) Robust Statistics. Theory and Methods. Wiley, Hoboken, NJ

Montgomery DC (2009) Introduction to Statistical Quality Control, 6th edn. Wiley, New York, NY

Oja H (2010) Multivariate Nonparametric Methods with R. An Approach Based on Spatial Signs and Ranks. Springer, New York, NY

Qiu P (2013) Introduction to Statistical Process Control. Chapman & Hall/CRC Press, Boca Raton, FL

Ramsay J, Silverman B (2005) Functional Data Analysis,. Springer, New York, NY

Ramsay JO, Hooker G, Graves S (2009) Functional Data Analysis with R and MATLAB. Springer, New York, NY

Ross J Gordon (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 JH, Woodall WH (1996) A control chart for preliminary analysis of individual observations. Journal of Quality Technology 28:265–278

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

Woodal WH (2016) Bridging the gap between theory and practice in basic statistical process monitoring (with discussion). Quality Engineering , forthcoming

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

Big Data Analytics and System Monitoring & Management

Kwok Leung Tsui and Yang Zhao

Abstract Due to the advancement of computation power and data storage/collection technologies, the field of data modelling and applications have been evolving rapidly over the last two decades, with different buzz words as knowledge discovery in databases (KDD), data mining (DM), business analytics, big data analytic, etc. There are tremendous opportunities in interdisciplinary research and education in data science, system informatics, and big data analytics; as well as in complex systems optimization and management in various industries of finance, healthcare, transportation, and energy, etc. In this paper, we will present our views and experience in the evolution of big data analytics, challenges and opportunities, and some applications in system monitoring and management.

Key words: Big data analytics; system surveillance; prognostics and health management; large scale simulation

1 Introduction

With the fast development of information technology, social media, data collection capacity and data storage, big data analytics field is now rapidly expanding in all science and engineering domains. Real-world applications such as telecommunications, health care, pharmaceutical or financial businesses generate massive amounts of data round the clock (del Río et al., 2014). Taking web social media alone for instance, today's customer is estimated to generate 2.5 quintillion bytes of data per day

Kwok Leung Tsui
Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong, e-mail: kltsui@cityu.edu.hk

Yang Zhao
Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong, e-mail: yangzhao9-c@my.cityu.edu.hk

between tweets, likes, comments, blogs, videos and images (Strong, 2015). These big data streams contain enormous information stored in the form of hidden patterns and unknown correlations. Analyzing big data that was previously untapped and inaccessible enables new insights resulting in better and faster decisions.

The process of examining big data to uncover hidden patterns, unknown correlations and other insights is referred to as big data analytics (SAS Institute Inc., 2012). The primary goal of big data analytics is to help make intelligent decisions through analyzing large data streams from multiple sources. Big data analytics has benefited many industries in various aspects, and created many opportunities for research (Russom, 2014, Wu et al., 2014, Chen et al., 2012). At the same time, more challenges have been raised along with opportunities, such as increased noise in large data, and under-developed policy for protecting individual privacy and security (Hilbert, 2016).

System monitoring and management refers to the framework of continuous surveillance, analysis and interpretation of related data for system maintenance, management and strategic planing. This framework is essential to ensure that the entire system is stable and in control. The concept of system is generally defined as 'an organized set of detailed methods, procedures and routines created to carry out a specific activity or solve a problem', and has been successfully applied to many domains, ranging from mechanical systems to public health systems (Tsui et al., 2008, 2014, Plett, 2006). Like many other applications and research fields, big data analytics has permeated in the domains of system monitoring and management, and has been verified to be an effective approach in practice, such as syndromic surveillance (Manyika et al., 2011, Ginsberg et al., 2009), electronics-rich system management (Pecht and Jaai, 2010), emergency departments simulation and optimization in medical system (Guo et al., 2016) and mass transit planning (Wang et al., 2014).

This paper attempts to review the issues associated with big data analytics in a general sense, by discussing the evolution of big data analytics, categorizing data types based on the data sources and collection processes, and providing some insights on research opportunities and challenges brought by big data. Specifically, we provide an overview and discussion on system monitoring and management driven by big data analytics.

2 Evolution of big data analytics

The origins of big data analytics can be traced back to 1970s or before, when research communities in computer science (CS) and statistics, such as machine learning and statistical computing played a major role in data mining. In the following decade, the scale and volume of data had grown dramatically due to the capability of computing power and automation. To distinguish these large data from the conventional ones, they were referred to as 'very large data base(VLDB)' or 'massive data (MD) sets' among the CS and statistics research communities. In 1990s, we witnessed an

unprecedentedly fast development and maturation of the methodology and theoretical foundations of data analytics across various disciplines from data mining, statistical learning, to knowledge discovery in database (KDD), and we label this as the first wave of big data analytics. In this period, most of the development of data analytics activities fell primarily in the realm of academic.

Humans have never stopped pushing the boundary of their knowledges forward. After the first wave, the giant success in methodology and theory development of data analytics quickly traveled to every corner of the research world, and even industry. More and more people realized the value of big data, and discovered the potential to change and improve society and humans lives. Since 2000, big data analytics has successfully developed in a lot more disciplines, such as business analytics (in business and management schools); and informatics in science and engineering (including bioinformatics, health informatics, systems informatics, etc.). We label this period the second wave, during which there was a parallel development in big data analytics among academic, education and industry. Figure 1 depicts the brief history we discussed above, which provides a clear picture of the evolution of big data analytics.

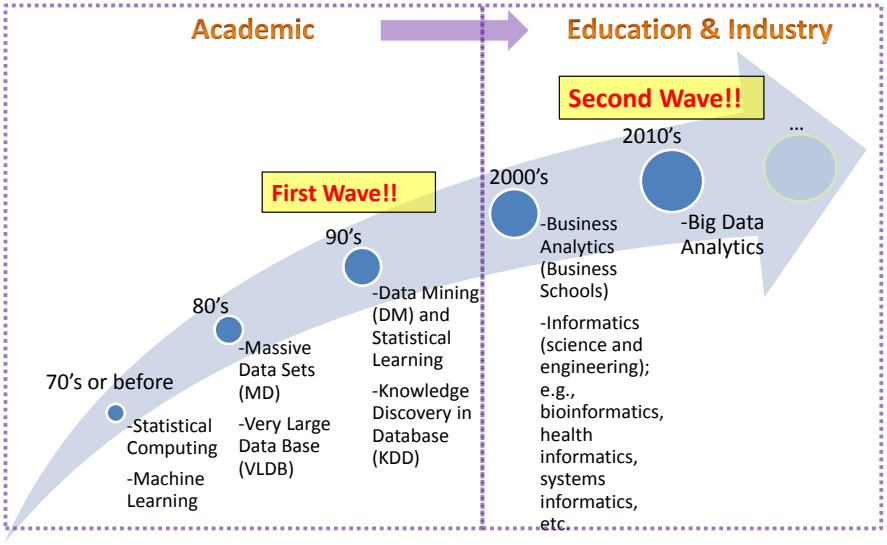


Fig. 1: A brief history of big data analytics

3 Opportunities and risks of big data analytics

3.1 Active and passive data

Data sources are critical to the effectiveness of data analytics. Depending on the sources and collecting processes, we divide big data into two main categories: Active data and passive data.

Active data (or primary data) refer to the data that are collected to study the scientific, health, engineering or business problems at hand, through a well-designed or planned data generation or collection mechanism for the purpose of study. This is similar to Design of Experiment (DOE) study in manufacturing applications. Examples of active data can be found in many applications, such as digital survey of the sky in astronomy, monitoring and surveillance in risk management in finance and banking industry, sensor data for prognostics and systems health management (PHM), transportation management, computer and communications management, etc.

On the contrary, passive data (or secondary data) refer to the data that are readily available or naturally collected for various other purposes but are potentially useful for addressing our current questions of interest. This is similar to production data in manufacturing applications. Examples of passive data include customer transaction data, electronic medical records, web searches, social media data, etc.

In real applications, active and passive data are complementary to each other. It is most effective to make use of both active data and passive data in real life data analytics applications. In the next section, we will present some data analytics examples which take advantage of both data types.

3.2 Opportunities of big data analytics research

The availability of big data in many new areas introduces new research opportunities in statistical modeling. In particular, enormous and detailed data at various locations and time domains make it possible to develop models for both population and individual levels. For example, Alyass et al. (2015) provided a thorough discussion on the feasibility and challenges of personalized medicine. A global approach to personalized medicine might be to model population heterogeneity in real time, as well as to integrate and manage various data sources and types to improve patient treatment. Another example is to develop in-situ automotive prognostics by tracking and analyzing user-specific driving records over the life of an automobile (Ji et al., 2013).

An integration of active and passive big data leads to a new challenge in big data analytics. Instead of solely relying on active data, researchers are now trying to incorporate existing passive data that might enhance the models for data analytics, prediction and decision making. For example, in wind turbine applications, engi-

neers attemped to improve wind power prediction by integrating wind speed (active data) and environmental factors (passive data), such as wind direction, air density, humidity, turbulence intensity, etc (Lee et al., 2015). In public health applications, one would like to improve the accuracy of predicting weekly counts for influenza like illness (ILI) by integrating ILI activity reports (active data) from Centers for Disease Control & Prevention (CDC) and internet search data (passive data) (Yang et al., 2015).

There are a lot more research opportunities, such as design of sensor locations, development of new modeling methods for incorporating web data as predictor variables and scaling with large data size and dimensionality, forecasting with process parameters, etc.

3.3 Risks in big data analytics

While big data creates a lot of opportunities, we should be aware of the potential risks as well.

Below we illustrate the opportunity and risk of big data analytics through the famous Google Flu Trend (GFT) example. GFT is a data analytics model developed by Google for predicting weekly reported ILI rate using instant query data (Lazer et al., 2014, Ginsberg et al., 2009). ILI is defined as a influenza like clinical syndrome, such as fever and cough, without a known cause. It is regarded as an indicator of influenza activity level around the region. CDC reports weekly ILI rates in the US with state-level detail, but there is always a one to three weeks delay in the report. It is recognized that a timely detection of acute disease outbreak means more days gained, more lives saved and more resources saved. Therefore, an accurate prediction of ILI before CDC's release report would be helpful for developing intervention strategies and remedies. In 2008, researchers from Google developed a web service GFT, claiming that they could accurately predict (nowcast) the ILI rate by modeling instant search queries. However, as reported in Lazer et al. (2014), Butler (2013), GFT failed by predicting more than double the proportion of doctor visits for ILI than the CDC report in the 2012-2013 season. Figure 2 depicts the trend of GFT prediction on ILI and actual CDC data over time (Lazer et al., 2014). As shown in Figure 2, GFT reported overly high flu prevalence from 21 August 2011 to 1 September 2013.

GFT's failure highlights a number of potential risks in prediction and forecasting models based on big data analytics. For example, the number of predictive variables can change over time, the impact of individual variables may change as well, and thus it is important to update the prediction model over time. In fact, the GFT prediction model has run ever since 2009, with a few changes announced in October 2013 (Ginsberg et al., 2009). GFT's failure has led to a large number of research papers aiming at improving the prediction performance of GFT (Copeland and et al., 2013, Yang et al., 2015, Santillana et al., 2014). One representative method is ARGO proposed by Yang et al. (2015), which not only incoprates the seasonality in historical

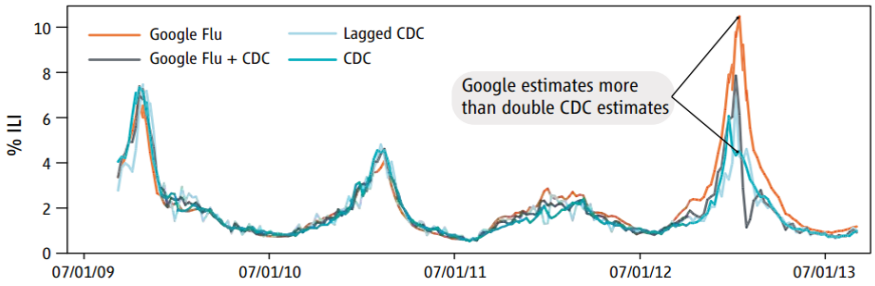


Fig. 2: GFT overestimation in the 2012-2013 season (Lazer et al., 2014)

ILI rates, but also captures changes in peoples's online search behavior over time. Comparing GFT's failure and all the subsequent revisions including ARGO in ILI prediction, there are some lessons we can learn. First, it is important to investigate and understand why the search terms are predictive. Second, inference relying on big data sources only may be misleading, one should investigate information from big data together with traditional knowledge.

In addition to the GFT example, there are plenty of other examples that illustrate the potential risks of big data analytics. In medical research, one has reported that forty percent of the experiments reported in research journal can not be reproduced. In finance hedge fund companies, many consultants have claimed that they have outperformed the market by applying their investment models to historical data. The truth is that their claimed successes were mainly caused by noises rather than signals most of the time.

4 Big data analytics in system monitoring and management

In this section, we will discuss some applications in system monitoring and management through big data analytics. Specifically, we will illustrate the role of big data analytics in three application domains, namely public health and healthcare surveillance, prognostics and systems health management (PHM), and large-scale simulations.

4.1 Public health and healthcare surveillance

Public health surveillance aims to systematically collect, analyze, and interpret public health data to understand trends; detect changes in disease incidence and death rates; and plan, implement, and evaluate public health practices (Tsui et al., 2008). Big data provides great promise for public health, as certain critical activities, such as

monitoring population health status and evaluating population-based health service quality, require the ability to collect volumes of information, rapidly interpret data, and monitor data for long periods of time (Thorpe and Gray, 2015).

At population level, traditional public health data includes information from vital statistics registries and hospital admission statistics. In the last few decades, more health data has been assimilated from electronic medical records, over the counter (OTC) drug purchase records, geographical positioning systems, social media and beyond (Wyber et al., 2015). At individual level, personal health data has been diversified and enriched as well. New techniques, such as wireless wearable electrocardiogram (ECG) sensor and wearable glucose monitor enable the cost-effective collection of abundant detailed personal health information (Zheng et al., 2014). These new data sources, together with traditional ones, provide great opportunities for developing effective systems for improving healthcare and public health surveillance. Below we will address three common surveillance systems.

4.1.1 Syndromic surveillance

The objective of syndromic surveillance is to make early detection of disease outbreaks (natural or an intended bioattack) by monitoring data that is related to the outbreak, such as influenza-like illness (ILI) symptoms, over-the-counter (OTC) drug sales, hospital telephone hotline calls, and emergency room visits (Tsui et al., 2008). The data streams are usually in temporal, spatial or spatiotemporal form. By monitoring various disease-related indicators, an outbreak can be detected earlier than by conventional reporting of confirmed cases, so that countermeasures can be implemented effectively (Rolka et al., 2007, Shmueli and Burkom, 2010).

As an example of syndromic surveillance, Figure 3 shows the time series plots of syndromic data taken from a large Maryland county during influenza season of 2004 (Rolka et al., 2007). The plot depicts the trend of several disease-related indicators, including counts of respiratory diagnoses from visits to civilian physician offices ('Office Visits'), military clinic visits ('MILITARY'), hospital emergency departments ('ED-UI' and 'ED ILI'), and sales of related OTC remedies ('OTC'). As shown in Figure 3, there is a sharp rise among the indicator beginning in late November 2003, which is confirmed by positive laboratory influenza tests. It should be noted that the September increase in OTC sales is not consistent with the other data streams. In fact, there were sporadic influenza cases documented in October and early November, which were reflected by OTC sales while not by the clinical visits.

Early detection of outbreaks is not trivial in multivariate data streams environment. Decision making involves analytics in terms of which data sources or combinations to test, which algorithm to use, how to define an outbreak with respect to numerous types of outbreak patterns, etc. Various independent syndromic surveillance systems have been developed, such as Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE) (Manyika et al., 2011), Early Aberration Reporting System (EARS) (Hutwagner et al., 2003) and GFT (Ginsberg et al., 2009). A detailed review concerning syndromic surveillance

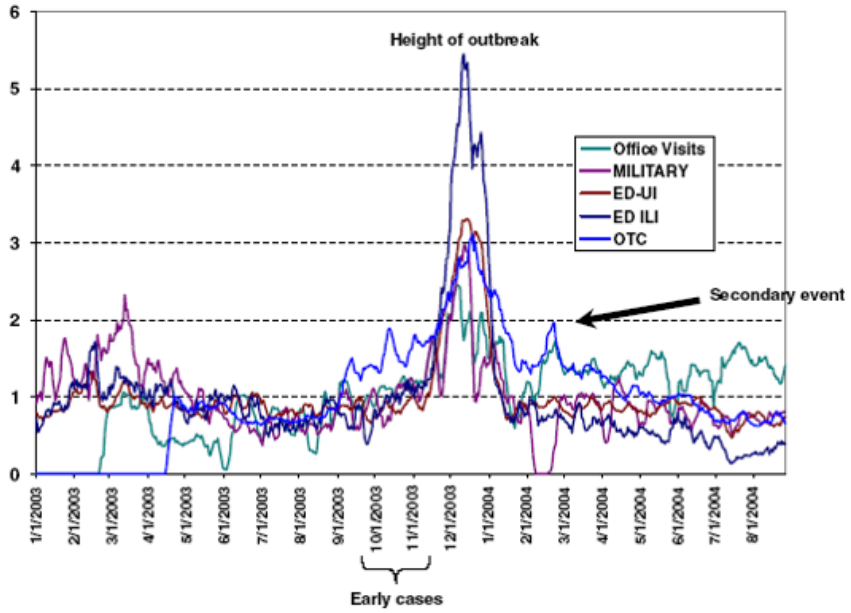


Fig. 3: An example of respiratory syndrome data (Rolka et al., 2007)

systems can be found in (Tsui et al., 2008) and (Yan et al., 2006). However, as pointed out by Tsui et al. (2013), how to integrate disparate data sources as well as unify with other surveillance systems to provide accurate detection of outbreaks remains a major challenge for the existing syndromic surveillance systems.

4.1.2 Public health surveillance

The objective of public health surveillance is to examine trends, detect changes in disease incidence and death rates, and to plan, implement, and evaluate public health practice by systematically collecting, analyzing, and interpreting public health data (chronic or infectious diseases).

Understanding the challenges to nations' public health system and how those challenges are shifting over time is of crucial importance for policymaker to establish effective strategies. In general, databases containing sufficient information about mobility and mortality across regions, time, age, and gender are prerequisite for informed analytics. Many public health organizations have made great efforts to maintain such databases, such as the Global Burden of Disease (GBD) project by the World Health Organization (WHO) for quantifying health loss from hundreds of diseases, injuries, and risk factors (Forouzanfar et al., 2015), and a wide

array of disease database by the Centers for Disease Control and Prevention (CDC) (http://www.cdc.gov/DataStatistics/).

In public health surveillance, the volume and velocity of data streams have been dramatically growing since the last decades. Taking the sample-based mortality surveillance system in China as an example, the surveillance population has increased from 6% to 24% of the Chinese population from 1978 to 2013 (Liu et al., 2016). In spite of data volume, the advanced information technology has made the collection of cause-of-death data in a more timely manner. Since 2008, information on individual deaths in all population catchment areas in China has been reported in real time via a Internet-based reporting system (Wang et al., 2008).

The availability of public health big data may provide a comprehensive picture of health system status in terms of what causes significant change in population, what the underlying risks are, how the pattern of health loss changes, etc. Plenty of efforts have been done for monitoring and evaluating population's health by taking advantage of public health big data. For example, GBD 2013 Mortality and Causes of Death Collaborators (2015) provided a systematic analysis of the levels and trends for age-sex-specific all-cause and cause-specific mortality for 240 causes of death; Zhou et al. (2015) studied the effect of ambient air pollution on adult respiratory mortality in China at city level; and Pluemper and Neumayer (2006) investigated the impact of armed conflict on gender structure in life expectancy.

4.1.3 Personal health surveillance

The objective of personal health surveillance is to monitor personal health performance, such as medical history, real-time health information and vital signs, for understanding individual health conditions, early detection of health risks and providing effective medical care to individuals.

In the field of personal health surveillance, the big data approach may facilitate the development of effective medical care system and enable more precise management of individuals to improve the health of entire populations. Researchers have been actively seeking for innovative solutions that could improve the quality of patient care via big data analytics.

One example is the use of unobtrusive sensing and wearable devices for personal health monitoring. For patients, the devices can provide real-time information and facilitate timely remote intervention to acute events such as stroke and heart attack. This type of implementation would be effective particularly in rural areas where expert treatment may be unavailable (Zheng et al., 2014). Additionally, for healthy population, unobtrusive and wearable monitoring can track their health and fitness closely, which will enable detecting any health risk and facilitating the implementation of preventive measures at an earlier stage.

At population level, the data collected at individual level can be aggregated for understanding and evaluating the medical care effectiveness of entire cohort. Various analytics methods have been proposed for monitoring patient disease conditions, such as sets-based methods (Chen, 1978) and risk adjustment methods (Steiner

et al., 2000). Woodall (2006) provides detailed discussions of these methods for healthcare applications.

4.2 Prognostics and systems health management (PHM)

PHM is the process of real time monitoring and accessing the extent of deviation and degradation of a system for predicting its future 'effective reliability' (Pecht and Jaai, 2010). In recent years, PHM has emerged as an essential approach for achieving competitive advantages in the global market by improving reliability, maintainability, safety, and affordability (Tsui et al., 2014). In industrial and system engineering, modern systems are often built with overwhelming complexities. Monitoring component performance through sensors (e.g. vibration, current/voltage, etc.) have been widely deployed within these systems, which enable monitoring data streams at both macro and micro scale.

Evaluating system reliability by analyzing the monitoring data in real time is crucial for managing system. Taking battery management system (BMS) as an example, BMS plays a vital role in improving battery performance and optimizing system operation in a safe and reliable manner. Various sensors are installed in the battery pack for data acquisition at the monitoring layer, and then the real-time collected data are used to maintain the system safety and determine the battery state (Xing et al., 2011). Battery state, mainly including state of health (SOH) and state of charge (SOC), is indicator of the health status of batteries, which can be used for determining the charge time, discharge strategy, cell equalization, and thermal management among the cells. Plenty of research works have been done for system monitoring and maintenance in BMS, such as remaining useful life (RUL) prediction (He et al., 2011, Si, 2015, Saha et al., 2009) and state of charge (SOC) estimation (Chen et al., 2014, Omar et al., 2013, Plett, 2006). A comprehensive review on PHM approaches in BMS and electronics-rich systems can be found in the work of Plett (2006) and Xing et al. (2011).

Besides BMS, there are many PHM applications taking advantage of large monitoring data streams. For example, fault diagnosis on gear crack development (Lei and Zuo, 2009, You et al., 2010), predicting RUL of rotational bearings (Chen and Tsui, 2013, Mahamad et al., 2010) and equipment maintenance in large-scale smart manufacturing facilities (O'Donovan et al., 2015).

4.3 Large-scale simulation for public safety and disaster management

Real time data streams have become more and more available across pervasive public networks, which creates opportunities in intelligent management and operation under emergencies. Large-scale simulation, as a powerful tool for imitating the operation of

a real-world process or system, has been significantly advanced by the accessibility of big data for system representation and simulation model development. Large-scale simulation is of crucial importance in many applications, such as public safety and disaster management, where the in-situ data during emergency are difficult to collect, and tremendous cost is involved.

One example is simulating large-scale crowd evacuation under emergencies. Liu et al. (2013) proposed a simulation approach for detailed analysis of passenger flows and assessing the crowdedness level of metro stations based on field surveys. Wang et al. (2015) studied and quantified the impact of the crowd physiological and psychological factors on large-scale evacuation, and provided a probabilistic description of crowd route selection. Developing such simulation model requires an extensive understanding of human behaviors and ambient environment, such as effective analysis of personnel movement, overall traffic situation during evacuation, as well as uncertainties in individual behavioral reactions under pressure and tension (Liu et al., 2014, 2015). All these factors play an important role in evaluating evacuation plans and predicting evacuation time under emergencies.

Another example is simulating disease propagation in spatiotemporal domain, which provides a useful tool for establishing and evaluating preventive strategies. Figure 4 shows a simulated map for disease propagation in Atlanta metropolitan area. The size and color coding of the circles represent the proportion of infected people in the county, which will grow over time. It illustrates advent event scenarios with respect to spatial locations, enabling graphical display of various key disease spread and severity in a systematic and aggregate format, and analysis on the correlations of diseases with symptoms, such as age groups, contact history and cluster events. With advanced syndromic surveillance techniques and real-time monitoring data, the dynamical spatiotemporal spread of disease can be simulated precisely. Many simulation models and methods have been proposed for studying disease propagation. For example, Wong et al. (2016) developed a simulation model of an influenza pandemic with a localized population structure to study the effect of individual school closure strategies on influenza pandemic in Hong Kong. More example of disease spread simulation approaches can be found in the work of Angulo et al. (2013), Perez and Dragicevic (2009) and Yang et al. (2011).

5 Conclusions

Nowadays, the concept of big data is prevailing in many application and research domains. In this paper, we reviewed the evolution of big data analytics and discussed its research opportunities and challenges. Based on the data sources and collecting processes, we divide the types of big data into two main categories, i.e. active data and passive data. In real applications, active and passive data are complementary to each other. It is most effective to make use of both active data and passive data in real life data analytics applications. For illustration purpose, we discussed and

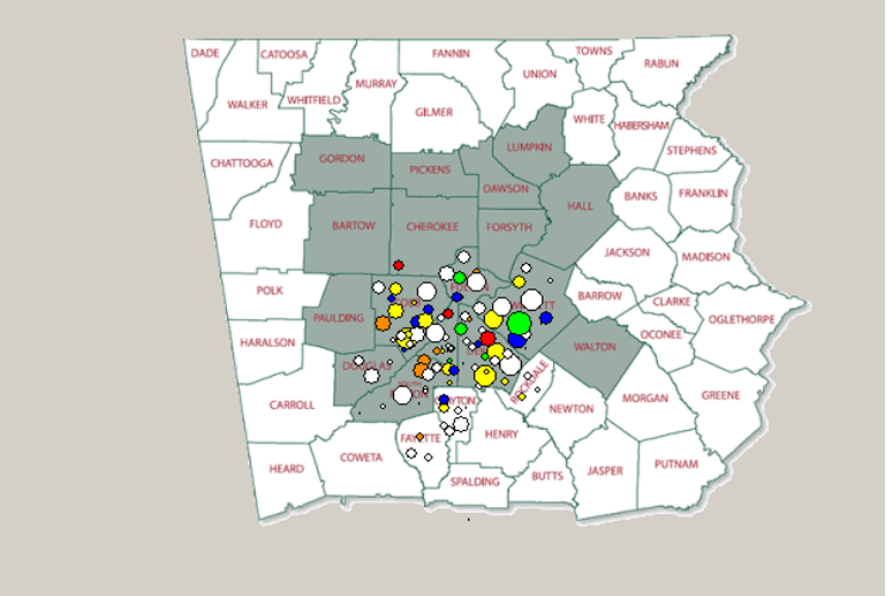


Fig. 4: A simulated maps for disease propagation in Atlanta metropolitan area

illustrated some applications in systems monitoring and management through big data analytics.

While there exist challenges and barriers, big data analytics has drawn much attention from researchers and practitioner in many fields. It is clear that it will continue to grow in new dimensions and areas under different challenges. This will provide opportunities and risks to academics and industries.

Acknowledgements Kwok L. Tsui gratefully acknowledges support from RGC Theme-based Research Scheme No. T32-101/15-R and No. T32-102/14-N, and the National Natural Science Foundation of China (NSFC) No. 11471275 and No. 71420107023.

References

Alyass, A.; Turcotte, M.; and Meyre, D. (2015). "From big data analysis to personalized medicine for all: challenges and opportunities". *BMC Medical Genomics*, 8.

Angulo, J.; Yu, H.-L.; Langousis, A.; Kolovos, A.; Wang, J.; Madrid, A.; and et al. (2013). "Spatiotemporal Infectious Disease Modeling: A BME-SIR Approach". *PLoS ONE*, 8(9).

Butler, D. (2013). "When Google got flu wrong". *Nature*, 494, pp. 155.

Chen, H.; Chiang, R. H. L.; and Storey, V. C. (2012). "Business Intelligence and Analytics: From Big Data to Big Impact". *MIS Q.*, 36(4), pp. 1165–1188.

Chen, N. and Tsui, K. L. (2013). "Condition monitoring and remaining useful life prediction using degradation signals: revisited". *IIE Transactions*, 45(9), pp. 939–952.

Chen, R. (1978). "A Surveillance System for Congenital Malformations". *Journal of the American Statistical Association*, 73(362), pp. 323–327.

Chen, X.; Shen, W.; Cao, Z.; and Kapoor, A. (2014). "A novel approach for state of charge estimation based on adaptive switching gain sliding mode observer in electric vehicles". *Journal of Power Sources*, 246, pp. 667–678.

Copeland, P. and et al. (2013). "Google disease trends an update".

del Río, S.; López, V.; Benítez, J. M.; and Herrera, F. (2014). "On the use of MapReduce for imbalanced big data using Random Forest". *Information Sciences*, 285(0), pp. 112 – 137.

Forouzanfar, M. H.; Alexander, L.; Anderson, H. R.; Bachman, V. F.; Biryukov, S.; and et. al. (2015). "Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013". *The Lancet*, 386, pp. 2287–2323.

GBD 2013 Mortality and Causes of Death Collaborators (2015). "Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013". *The Lancet*, 385, pp. 117–171.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski1, M. S.; and Brilliant, L. (2009). "Detecting influenza epidemics using search engine query data". *Nature*, 457, pp. 1012–1014.

Guo, H.; Goldsman, D.; Tsui, K.-L.; Zhou, Y.; and Wong, S.-Y. (2016). "Using simulation and optimisation to characterise durations of emergency department service times with incomplete data". *International Journal of Production Research*, pages 1–18.

He, W.; Williard, N.; Osterman, M.; and Pecht, M. (2011). "Prognostics of lithium-ion batteries based on Dempster-Shafer theory and the Bayesian Monte Carlo method". *Journal of Power Sources*, 196(23), pp. 10314–10321.

Hilbert, M. (2016). "Big Data for Development: A Review of Promises and Challenges". *Development Policy Review*, 34, pp. 135–174.

Hutwagner, L.; Thompson, W.; Seeman, G. M.; and Treadwell, T. (2003). "The bioterrorism preparedness and response Early Aberration Reporting System (EARS)". *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 80(Suppl 1), pp. i89–i96.

Ji, B.; Pickert, V.; Cao, W.; and Zahawi, B. (2013). "In Situ Diagnostics and Prognostics of Wire Bonding Faults in IGBT Modules for Electric Vehicle Drives". *IEEE Transactions on Power Electronics*, 28(12), pp. 5568–5577.

Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. (2014). "The Parable of Google Flu: Traps in Big Data Analysis". *Science*, 343, pp. 1203–1205.

Lee, G.; Ding, Y.; Genton, M. G.; and Xie, L. (2015). "Power Curve Estimation With Multivariate Environmental Factors for Inland and Offshore Wind Farms". *Journal of the American Statistical Association*, 110(509), pp. 56–67.

Lei, Y. and Zuo, M. J. (2009). "Gear crack level identification based on weighted K nearest neighbor classification algorithm". *Mechanical Systems and Signal Processing*, 23(5), pp. 1535 – 1547.

Liu, S.; Lo, S.; Ma, J.; and Wang, W. (2014). "An Agent-Based Microscopic Pedestrian Flow Simulation Model for Pedestrian Traffic Problems". *IEEE Transactions on Intelligent Transportation Systems*, 15(3), pp. 992–1001.

Liu, S.; Wu, X.; Lopez, A. D.; Wang, L.; Cai, Y.; Page, A.; Yin, P.; Liu, Y.; Li, Y.; Liu, J.; You, J.; and Zhou, M. (2016). "An integrated national mortality surveillance system for death registration and mortality surveillance, China". *Bulletin of the World Health Organization*, 94(1), pp. 46–57.

Liu, S. B.; Lo, S. M.; Tsui, K. L.; and Wang, W. L. (2015). "Modeling Movement Direction Choice and Collision Avoidance in Agent-Based Model for Pedestrian Flow". *Journal of Transportation Engineering*, 141(6), pp. 04015001.

Liu, S. B.; Lo, S. M.; Wang, W. L.; Ma, J.; and Yuen, J. K. K. (2013). "Crowding in Metro Stations : Passenger Flow Analysis and Simulation". *the 92nd annual meeting of the Transportation Research Board*, pages 1–11.

Mahamad, A. K.; Saon, S.; and Hiyama, T. (2010). "Predicting remaining useful life of rotating machinery based artificial neural network". *Computers and Mathematics with Applications*, 60(4), pp. 1078 – 1087.

Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; and Byers, A. H. (2011). "Big Data: The Next Frontier for Innovation, Competition, and Productivity".

O'Donovan, P.; Leahy, K.; Bruton, K.; and O'Sullivan, D. T. J. (2015). "An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities". *Journal of Big Data*, 2(1), pp. 25.

Omar, N.; Van den Bossche, P.; Coosemans, T.; and Van Mierlo, J. (2013). "Peukert revisited-critical appraisal and need for modification for lithium-ion batteries". *Energies*, 6(11), pp. 5625–5641.

Pecht, M. and Jaai, R. (2010). "A prognostics and health management roadmap for information and electronics-rich systems". *Microelectronics Reliability*, 50(3), pp. 317–323.

Perez, L. and Dragicevic, S. (2009). "An agent-based approach for modeling dynamics of contagious disease spread". *International Journal of Health Geographics*, 8(1), pp. 1–17.

Plett, G. L. (2006). "Sigma-point Kalman filtering for battery management systems of LiPB-based HEV battery packs. Part 1: Introduction and state estimation". *Journal of Power Sources*, 161(2), pp. 1356–1368.

Pluemper, T. and Neumayer, E. (2006). "The Unequal Burden of War: The Effect of Armed Conflict on the Gender Gap in Life Expectancy". *International Organization*, 60(3), pp. 723–754.

Rolka, H.; Burkom, H.; Cooper, G.; Kulldorff, M.; Madigan, D.; and Wong, W.-K. (2007). "Issues in applied statistics for public health bioterrorism surveillance

using multiple data streams: research needs". *Statistics in Medicine*, 26, pp. 1834 – 1856.

Russom, P. (2014). "Big Data Analytics". *TDWI Best Practices Report, Fourth Quarter*.

Saha, B.; Goebel, K.; Poll, S.; and Christophersen, J. (2009). "Prognostics Methods for Battery Health Monitoring Using a Bayesian Framework". *Instrumentation and Measurement, IEEE Transactions on*, 58(2), pp. 291–296.

Santillana, M.; Zhang, W.; Althouse, B. M.; and Ayers, J. W. (2014). "What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?". *Article in American Journal of Preventive Medicine*, 47, pp. 341-347.

SAS Institute Inc. (2012). "Big Data Meets Big Data Analytics".

Shmueli, G. and Burkom, H. (2010). "Statistical Challenges Facing Early Outbreak Detection in Biosurveillance". *Technometrics*, 52(1), pp. 39–51.

Si, X. (2015). "An Adaptive Prognostic Approach via Nonlinear Degradation Modeling: Application to Battery Data". *Industrial Electronics, IEEE Transactions on*, 62(8), pp. 5082–5096.

Steiner, S. H.; Cook, R. J.; Farewell, V. T.; and Treasure, T. (2000). "Monitoring surgical performance using risk-adjusted cumulative sum charts". *Biostatistics*, 1(4), pp. 441–452.

Strong, C. (2015). *Humanizing Big Data: Marketing at the Meeting of Data, Social Science & Consumer Insight*. Kogan Page, 1st edition.

Thorpe, J. H. and Gray, E. A. (2015). "Big Data and Public Health: Navigating Privacy Laws to Maximize Potential". *Public Health Reports*, 130, pp. 171–175.

Tsui, K. L.; Chen, N.; Zhou, Q.; Hai, Y.; and Wang, W. (2014). "Prognostics and Health Management : A Review on Data Driven Approaches". *Mathematical Problems in Engineering., Hindawi Publishing Corporation*, 2015.

Tsui, K.-L.; Chiu, W.; Gierlich, P.; Goldsman, D.; Liu, X.; and Maschek, T. (2008). "A Review of Healthcare, Public Health, and Syndromic Surveillance". *Quality Engineering*, 20(4), pp. 435–450.

Tsui, K.-L.; Wong, Z. S.-Y.; Goldsman, D.; and Edesess, M. (2013). "Tracking Infectious Disease Spread for Global Pandemic Containment". *IEEE Intelligent Systems*, 28(6), pp. 60–64.

Wang, J.; Sun, J.; and Lo, S. (2015). "Randomness in the evacuation route selection of large-scale crowds under emergencies". *Applied Mathematical Modelling*, 39(18), pp. 5693 – 5706.

Wang, L.; Wang, Y.; Jin, S.; Wu, Z.; Chin, D. P.; Koplan, J. P.; and Wilson, M. E. (2008). "Emergence and control of infectious diseases in China". *The Lancet*, 372, pp. 1598–1605.

Wang, W. L.; Lo, S. M.; and Liu, S. B. (2014). "Aggregated Metro Trip Patterns in Urban Areas of Hong Kong: Evidence from Automatic Fare Collection Records". *Journal of Urban Planning and Development*, 141(3), pp. 10.

Wong, Z. S.-Y.; Goldsman, D.; and Tsui, K.-L. (2016). "Economic Evaluation of Individual School Closure Strategies: The Hong Kong 2009 H1N1 Pandemic". *Plos One*, 11(1), pp. e0147052.

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

Wu, X.; Zhu, X.; Wu, G.-Q.; and Ding, W. (2014). "Data Mining with Big Data". *IEEE Trans. on Knowl. and Data Eng.*, 26(1), pp. 97–107.

Wyber, R.; Vaillancourt, S.; Perry, W.; Mannava, P.; Folaranmi, T.; and Celi, L. (2015). "Big data in global health: Improving health in low- and middle-income countries". *Bulletin of the World Health Organization*, 93(3), pp. 203–208.

Xing, Y.; Ma, E. W. M.; Tsui, K. L.; and Pecht, M. (2011). "Battery management systems in electric and hybrid vehicles". *Energies*, 4(11), pp. 1840–1857.

Yan, P.; Zeng, D.; Chen, Hsinchun", e. S.; Zeng, D. D.; Chen, H.; Thuraisingham, B.; and Wang, F.-Y. (2006). *A Review of Public Health Syndromic Surveillance Systems*, pages 249–260. Springer Berlin Heidelberg, Berlin, Heidelberg.

Yang, S.; Santillana, M.; and Kou, S. C. (2015). "ARGO: a model for accurate estimation of influenza epidemics using Google search data". *in Proceedings of the National Academy of Sciences*.

Yang, Y.; Atkinson, P. M.; and Ettema, D. (2011). "Analysis of CDC social control measures using an agent-based simulation of an influenza epidemic in a city". *BMC Infectious Diseases*, 11(1), pp. 1–10.

You, M. Y.; Liu, F.; Wang, W.; and Meng, G. (2010). "Statistically Planned and Individually Improved Predictive Maintenance Management for Continuously Monitored Degrading Systems". *IEEE Transactions on Reliability*, 59(4), pp. 744–753.

Zheng, Y. L.; Ding, X. R.; Poon, C. C. Y.; Lo, B. P. L.; Zhang, H.; Zhou, X. L.; Yang, G. Z.; Zhao, N.; and Zhang, Y. T. (2014). "Unobtrusive Sensing and Wearable Devices for Health Informatics". *IEEE Transactions on Biomedical Engineering*, 61(5), pp. 1538–1554.

Zhou, M.; He, G.; Liu, Y.; Yin, P.; Li, Y.; Kan, H.; Fan, M.; Xue, A.; and Fan, M. (2015). "The associations between ambient air pollution and adult respiratory mortality in 32 major Chinese cities, 2006–2010". *Environmental Research*, 137, pp. 278 – 286.

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 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 denomination "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). In many cases, the variable approach, contradicting the intuition, may also result in faster detection of special causes.

Eugenio K. Epprecht
Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro, Brazil, e-mail: eke@puc-rio.br

Francisco Aparisi
Departamento de Estadística e I.O. Aplicadas y Calidad, Universidad Politécnica de Valencia, Spain, e-mail: faparisi@eio.upv.es

Omar Ruiz
Centro de Investigaciones Biotecnológicas del Ecuador, Escuela Superior Politécnica del Litoral, Ecuador, e-mail: oruiz@espol.edu.ec

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 the process may be 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). In many cases, the variable dimension approach, contradicting the intuition, may also result in faster detection of 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 first one to appear was the *variable-dimension* T^2 (*VDT2*) chart (Aparisi et al., 2012). The purpose of this paper is to review and compare the several variants of the approach.

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 denomination "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 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.

Four process control charts based on the 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 paper 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 either 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 and understand by the practitioners.

For the details, the reader is referred to the paper (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, 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 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%. It is quite relevant (ranging 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 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 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; it 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 $T^2_p - T^2_{p_1}$.

Similarly to 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., 2016). 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. Given a specified average sample size n_0, $n_1 < n_0 < n_2$. 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 paper cited, the authors considered that this first sample is of small dimension and size.

The chart is illustrated in the picture below, reproduced from Aparisi et al. (2014).

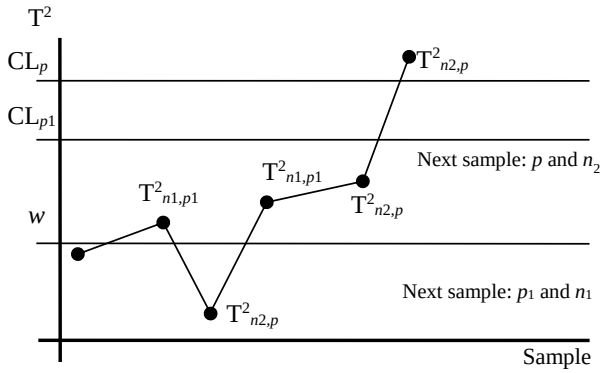


Fig. 1: VSSVDT2 chart from Aparisi et al. (2014).

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: n_1, n_2, CL_p, (and CL_{p1} for the chart with two control limits), w. Four or five parameters. 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 is 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, this gain is not so directly related with that proportion, 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

$$\frac{\%p \cdot (a \cdot n_2 - n_1) + 100 n_1}{n_0 \cdot a}$$

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

$$\frac{\%p \cdot (a \cdot n_2 - n_1) + 100 n_1}{\%p \cdot n_0 (a-1) + 100 n_0}$$

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% to 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 paper.

The ARL performance of the VSSVDT2 chart can never match the ARL performance of the VSST^2 chart 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 paper, the choice of proceeding to the smoothing first was justified by the performance analysis, carried out by the 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-T^2 chart (Epprecht et al., 2016) 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 $F_{T^2_p}(T^2)$ in the case of the samples with p variables, where $F_{T^2_{p_1}}(\cdot)$ and $F_{T^2_p}(\cdot)$ 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. Next, to make easier the operation of the chart, the cumulative probabilities 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.

To avoid extra operational complexity, and given the findings in Aparisi et al. (2012) that the VDT2 chart with only one control limit and one warning line performed in practice as well as the chart with two control limts and two warning lines, the EWMA procedure was applied with just 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 bounds for one-sided EWMA charts was proven effective by Gan (1993) and adopted since by other authors.

A VDEWMA-T2 chart is depicted below, where the big dots correspond to sample points from samples with all p variables.

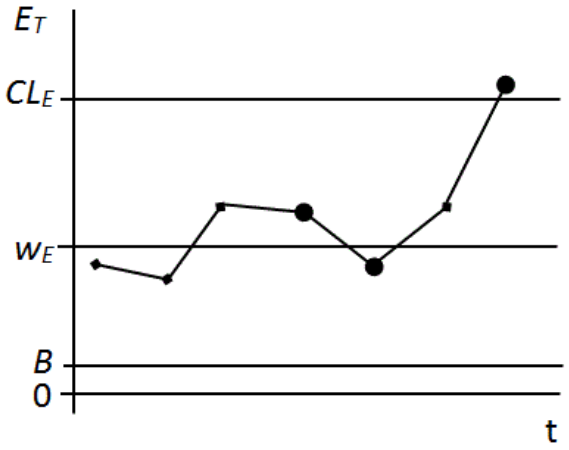


Fig. 2: VDEWMA-T2 chart.

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 limt is 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. (2016) 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 fo freedom (according to the sample dimension) and the cumulative probability of that T^2 value is converted to a Z score by:

$$z_t = \Phi^{-1}\left(F_{\chi^2_\nu}(T^2_\nu(t))\right)$$

It is the Z score which is smoothed into an EWMA statistic:

$$E_t = \max\{B, r z_t + (1-r)E_{t-1}\}$$

where r is the smoothing constant and $E_0 = B$.

After a signal and intervention, when resuming the monitoring, the EWMA is returned to the initial value $E_0 = B$.

A difference between the single control limit of this chart and the single control limit of the VDT2 chart is that the latter is active only with samples of p variables, whereas the former 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 (computation of the corresponding cumulative probability) and Z score.

Just for register, 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 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 paper, 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 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 optimization criterion leads to solutions in which the p variables are measured only a small proportion of the times. This contrasting behavior of the optimization 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 faster out-of-control conditions.

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 data from the user. The solutions satisfying this constraint still have smaller out-of-control ARLs than 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 $AATS = (ARL - 0.5)h$. The product

$$(ACS/h) \cdot AATS = \left(1 + a \cdot \frac{\%p}{100}\right) C_{p_1} (ARL - 0.5)$$

(note how h cancels out) 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. (2016) for more details and extensive tables of results, but in synthesis, 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).

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 details, the reader is referred to the original papers.

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, pp. 2853-2862.

Aparisi, F., and Haro, C. (2001). "Hotelling's T^2 Control Chart with Variable Sampling Intervals". *International Journal of Production Research*, 39, 14, pp. 3127-3140.

Aparisi, F., Epprecht, E.K., and Ruiz, O. (2012). "T^2 Control Charts with Variable Dimension". *Journal of Quality Technology*, 44, 4, pp. 375-393.

Aparisi, F., Epprecht, E.K., Carrión, A., and Ruiz, O. (2014). "The variable sample size variable dimension T^2 Control Chart". *International Journal of Production Research*, 52, 2, pp. 368-383.

Bodnar, O. and Schmid, W. (2005). "Multivariate control charts based on a projection approach". *Allgemeines Statistisches Archiv*, 89, pp. 75-93.

Champ, C.W. and Aparisi, F. (2008). "Double Sampling Hotelling's T^2 Charts". *Quality and Reliability Engineering International*, 24, pp. 153-166.

Costa, A. F. B. (1994). "$\bar{X}$ Charts with Variable Sample Size". *Journal of Quality Technology* 26(3), pp. 155-163.

Costa, A.F.B. (1999). "$\bar{X}$ charts with variable parameters". *Journal of Quality Technology* 31, 4, pp. 408-416.

Costa, A. F. B. and De Magalhães, M. S. (2005). "Economic design of two-stage $\bar{X}$ charts: the Markov-chain approach." *International Journal of Production Economics*, 95, 1, pp. 9-20.

Croasdale, R. (1974). "Control charts for a double-sampling scheme based on average production run length". *International Journal of Production Research* 12, pp. 585–592.

Daudin, J.J. (1992). "Double sampling charts". *Journal of Quality Technology* 24, 2, pp. 78–87.

Epprecht, E.K., Aparisi, F., Ruiz, O., and Veiga, A. (2013). "Reducing Sampling Costs in Multivariate SPC with a Double-Dimension T^2 Control Chart". *International Journal of Production Economics*, 144, 1, pp. 90-104.

Epprecht, E.K., Aparisi, F. and Ruiz, O. (2016). "A Variable-Dimension EWMA Chart for Multivariate Statistical Process Control." *Unpublished paper*.

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, pp. 311-325.

Ferrer, A. (2014). "Latent Structures-Based Multivariate Statistical Process Control: A Paradigm Shift." *Quality Engineering*, 26, pp. 72-91.

Gan, F. F. (1993). "Exponentially weighted moving-average control charts with reflecting boundaries." *Journal of Statistical Computation and Simulation*, 46, 1-2, pp. 45-67.

González, I., and Sánchez, I. (2010). "Variable selection for multivariate statistical process control". *Journal of Quality Technology*, 42, 3, pp. 242-259.

Jackson, J.E. (1980). "Principal components and factor analysis: part I – principal components". *Journal of Quality Technology*, 12, 4, pp. 201-213.

Jackson, J.E. (2003). *A User's Guide to Principal components.* Wiley.

Jackson, J.E., and Mudholkar, G.S. (1979). "Control Procedures for Residuals Associated with Principal Components Analysis". *Technometrics*, 21, 3, pp. 341-349.

Kourti, T., and MacGregor, J.F. (1996). "Multivariate SPC Methods for Process and Product Monitoring". *Journal of Quality Technology*, 28, 4, pp. 409-428.

Lowry, C.A., Woodall, W.H., Champ, C.W., and Rigdon, S. E. (1992). "A Multivariate Exponentially Weighted Moving Average Chart". *Technometrics*, 34, pp. 46-53.

Nomikos, P., and MacGregor, J.F. (1995). "Multivariate SPC Charts for Monitoring Batch Processes". *Technometrics*, 37, 1, pp. 41-59.

Prabhu, S. S.; Runger, G. C.; and Keats, J. B. (1993). "An Adaptive Sample Size Chart". *International Journal of Production Research*, 31, pp. 2895-2909.

Reynolds, M.R., Jr., Amin, R.W., Arnold, J.C., and Nachlas, J.A. (1988). "$\bar{X}$ charts with variable sampling intervals". *Technometrics* 30, 2, pp. 181-192.

Runger, G.C. (1996). "Projections and the U-squared multivariate control chart." *Journal of Quality Technology*, 28, pp. 313-319.

Rodrigues, A. A. A., Epprecht, E. K. and De Magalhães, M. S. (2011). "Double sampling control charts for attributes." *Journal of Applied Statistics*, 38, 1, pp. 87-112.

Steiner, S.H. (1999). "Confirmation sample control charts". *International Journal of Production Research*, 37, 4, pp. 737-748.

Woodall, W. H. and Mahmoud, M. A. (2005). "The Inertial Properties of Quality Control Charts." *Technometrics*, 47, 4, pp. 425-436.

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 that includes 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.

Presentation slides

. . . follow on the next pages.

Emmanuel Yashchin
IBM, Thomas J. Watson Research Ctr., Box 218, Yorktown Heights, NY 10598, USA, e-mail: yashchi@us.ibm.com

Statistical monitoring of multi-stage processes

Emmanuel Yashchin

IBM Research, Yorktown Heights, New York

XII Conference in Intelligent SQC, Hamburg, Germany, August 2016

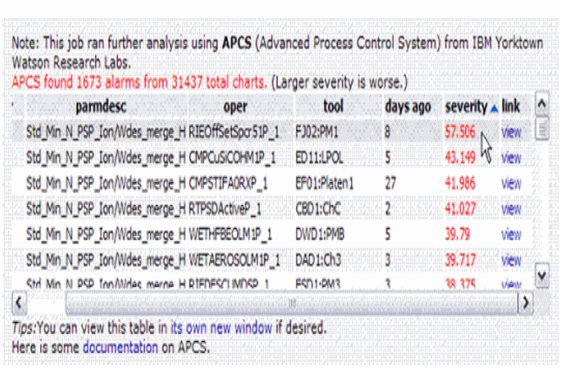
Outline

- Statistical Process Monitoring
- Multi-stage data flows
- A system for monitoring multi-stage processes
- Target setting
- Problem phase management
- Concluding Remarks

Special thanks to IBM colleagues: R. Baseman, A. Civil, W. Hoffman, D. Jensen, J. Komatsu, R. O. Corral, S. Ruegsegger, A. Spielberg, J. Wargo, B. White, P. Zulpa.

1. Selected IBM activities in Statistical Process Monitoring

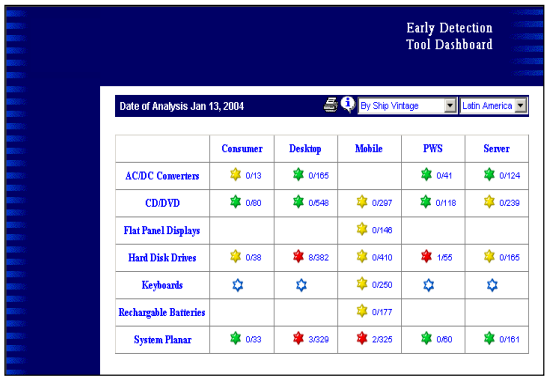
- Semiconductors
- Pension fund
- Supply chain
- Warranty data



Note: This job ran further analysis using **APCS** (Advanced Process Control System) from IBM Yorktown Watson Research Labs.
APCS found 1673 alarms from 31437 total charts. (Larger severity is worse.)

parmdesc	oper	tool	days ago	severity ▲	link
Std_Min_N_PSP_Ion/Wdes_merge_H	RIEOffSetSpcr51P_1	FJ02:PM1	8	57.506	view
Std_Min_N_PSP_Ion/Wdes_merge_H	CMPCuSiCOHM1P_1	ED11:LPOL	5	43.149	view
Std_Min_N_PSP_Ion/Wdes_merge_H	CMPSTIFA0RXP_1	EF01:Platen1	27	41.986	view
Std_Min_N_PSP_Ion/Wdes_merge_H	RTPSDActiveP_1	CBD1:ChC	2	41.027	view
Std_Min_N_PSP_Ion/Wdes_merge_H	WETHFBEOLM1P_1	DWD1:PMB	5	39.79	view
Std_Min_N_PSP_Ion/Wdes_merge_H	WETAEROSOLM1P_1	DAD1:Ch3	3	39.717	view
Std_Min_N_PSP_Ion/Wdes_merge_H	RIEDESCUNDXP_1	EC01:PM3	3	38.375	view

*Tips:*You can view this table in its own new window if desired.
Here is some documentation on APCS.

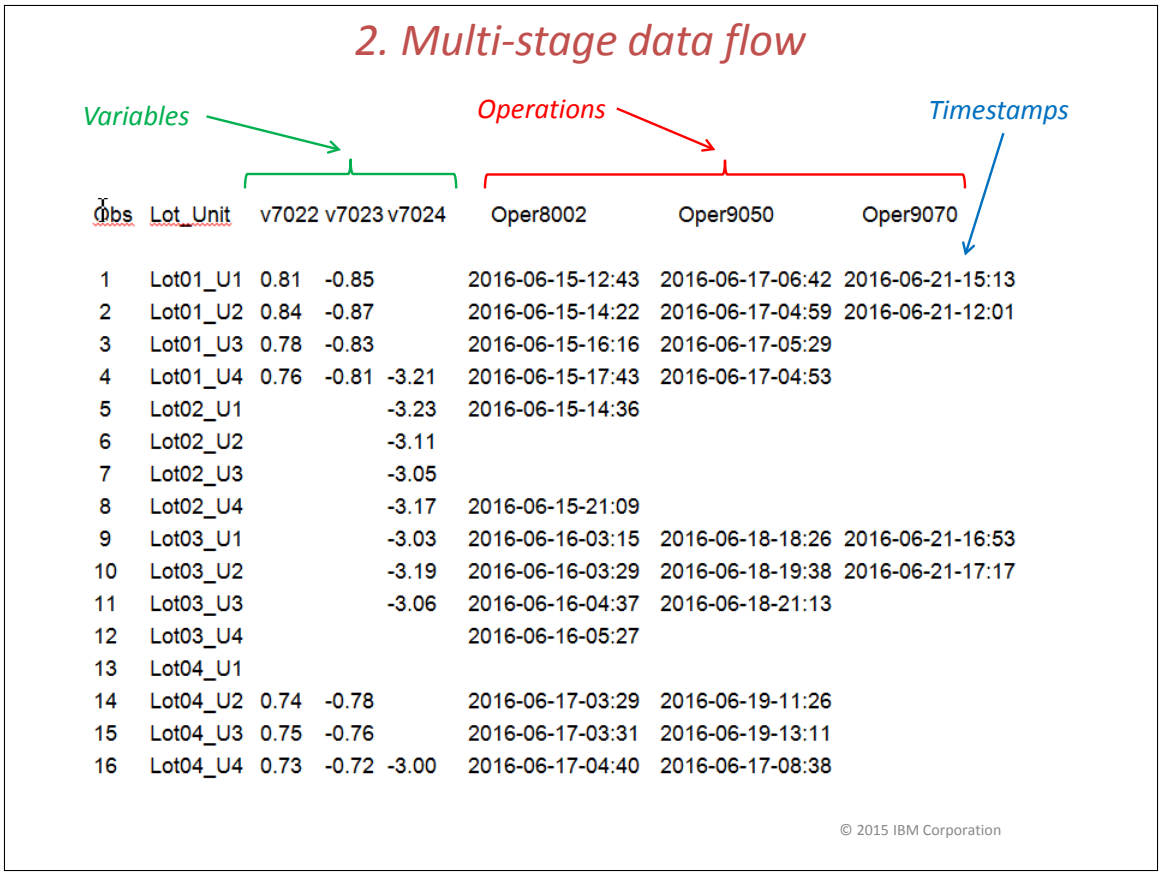


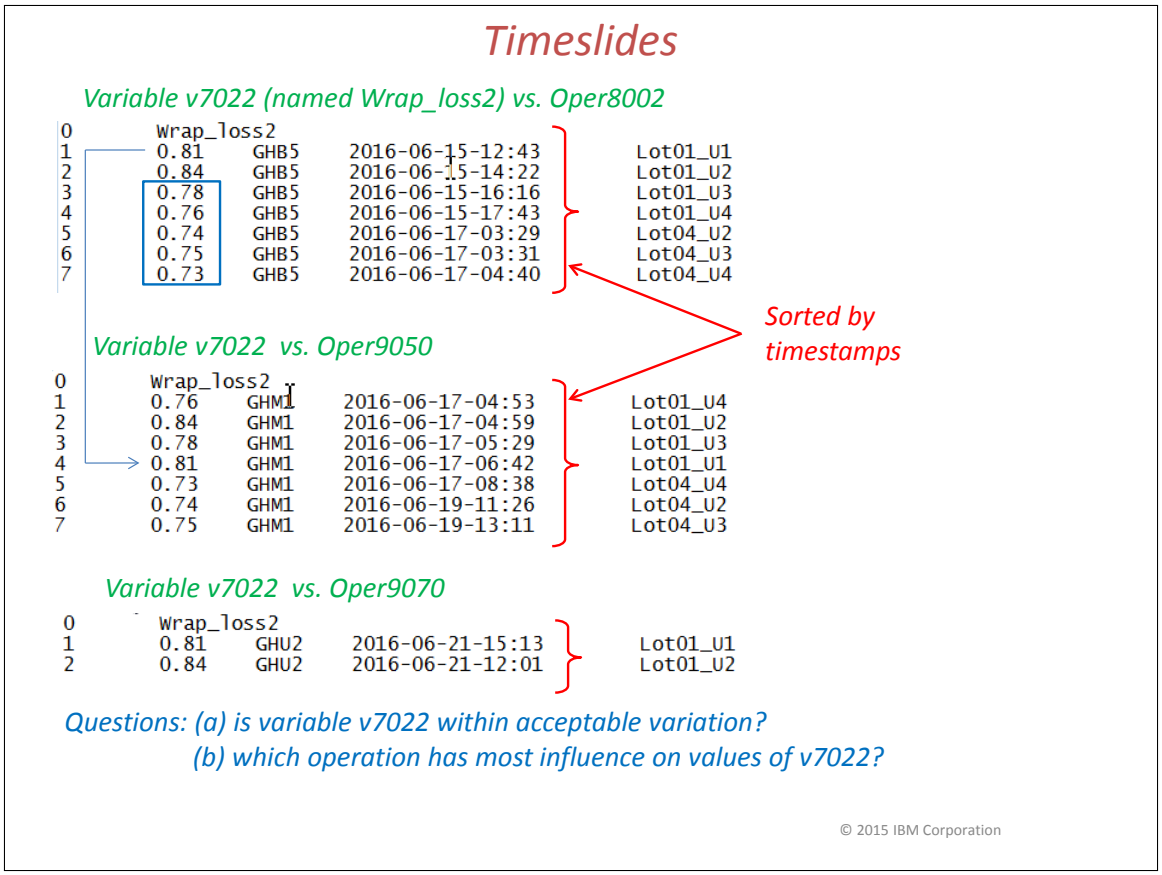
Early Detection Tool Dashboard

Date of Analysis Jan 13, 2004 — By Ship Vintage — Latin America

	Consumer	Desktop	Mobile	PWS	Server
AC/DC Converters	0/13	0/165		0/41	0/124
CD/DVD	0/80	0/548	0/297	0/118	0/239
Flat Panel Displays			0/146		
Hard Disk Drives	0/38	8/382	0/410	1/55	0/165
Keyboards			0/250		
Rechargable Batteries			0/177		
System Planar	0/33	3/329	2/325	0/80	0/161

Part Number	Code Name	Description	Step Name	Trend Stoplight	AVT Stoplight	QEWS Stoplight	Drill to Cube
000002000563	HMS	HMS	BLFO				
000002002626	GUAYIN_MAOKONG	GUAYIN_MAOKONG	BLFO				
000002002929	YILAN PLANAR	YILAN PLANAR	BLFO				
000002003346	YUNGAN PLANAR	YUNGAN PLANAR	BLFO				NO CUBE AVAILABLE
000002000882	BEITOU	BEITOU	BLFO				
000002000900	YUSHAN	YUSHAN	BLFO				
000002002928	HARUHI	HARUHI	BLFO				
000002000478	MEIYA	MEIYA	BLFO				
000002000035	PIRANHA	PIRANHA	BLFO				


2. Multi-stage data flow
Variables
Operations
Timestamps
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
© 2015 IBM Corporation


Timeslides
Variable v7022 (named Wrap_loss2) vs. Oper8002
0 Wrap_loss2
1 0.81 GHB5 2016-06-15-12:43 Lot01_U1
2 0.84 GHB5 2016-06-15-14:22 Lot01_U2
3 0.78 GHB5 2016-06-15-16:16 Lot01_U3
4 0.76 GHB5 2016-06-15-17:43 Lot01_U4
5 0.74 GHB5 2016-06-17-03:29 Lot04_U2
6 0.75 GHB5 2016-06-17-03:31 Lot04_U3
7 0.73 GHB5 2016-06-17-04:40 Lot04_U4
Sorted by timestamps
Variable v7022 vs. Oper9050
0 Wrap_loss2
1 0.76 GHM1 2016-06-17-04:53 Lot01_U4
2 0.84 GHM1 2016-06-17-04:59 Lot01_U2
3 0.78 GHM1 2016-06-17-05:29 Lot01_U3
4 0.81 GHM1 2016-06-17-06:42 Lot01_U1
5 0.73 GHM1 2016-06-17-08:38 Lot04_U4
6 0.74 GHM1 2016-06-19-11:26 Lot04_U2
7 0.75 GHM1 2016-06-19-13:11 Lot04_U3
Variable v7022 vs. Oper9070
0 Wrap_loss2
1 0.81 GHU2 2016-06-21-15:13 Lot01_U1
2 0.84 GHU2 2016-06-21-12:01 Lot01_U2
Questions: (a) is variable v7022 within acceptable variation?
(b) which operation has most influence on values of v7022?
© 2015 IBM Corporation

3. A system for monitoring multi-stage data

Objective: Analyze multiple timeslides*; detect, as early as possible, unfavorable conditions; identify operations that merit engineering attention*

- Statistically designed control schemes
  - ✓ *Probabilistic modeling of timeslide data*
  - ✓ *Targets and process windows (driven by business objectives)*
  - ✓ *Low rate of false alarms*
  - ✓ *Detection capability*
- Massively parallel deployment
- Monitoring system functions as a *Search Engine*
- *Log of analysis* updated on a regular basis
- Multi-layer *dashboard*
- Analytic/Graphical support for *diagnostics* and *alarm-related decisions,* incl. *identification of operations* associated with unfavorable conditions

4. Approach to Monitoring

- Specify the *segments* of multi-stage data for analysis
- For each *segment* specify the *variables* (e.g., v7022, v7023) to be monitored, and the *operations* (e.g., Oper8002, Oper9050) that have a potential to be associated with changes in these variables, esp. in an adverse way
- Specify *timeslides* for data analysis in a *segment*
- *Characterize* the distribution of variables parametrically. E.g. assume that variables are Gaussian, subject to changes in mean, μ. I.e., if the values of the variable are $\{X_i, i = 1, 2, \ldots\}$, focus on $\mu = E(X_i)$
- Specify: *Target*
  - Nuisance parameters (e.g. σ)
  - Type of control (*1-sided or 2-sided*)
  - Acceptable & unacceptable *levels* (μ_{accept}, $\mu_{unaccept}$)
  - Acceptable *rate of false alarms*
- Compute *Cusum-Shewhart (Page's) scheme parameters*
- *Apply scheme* to every timeslide; *flag* it if out-of-control conditions are detected; deliver attributes of each analysis

Main test: Cusum-Shewhart (Page's) scheme

Suppose: the *timeslide* at time *T* contains *N* data points
$\{X_i\}$, $i = 1, 2, \ldots, N$

Define: the set of scheme values $\{S_i, i = 1, 2, \ldots, N\}$ as follows:

where $S_0 = 0, \quad S_i = \max[0, S_{i-1} + (X_i - k)] \quad (evidence\ curve)$

$k \approx (\mu_{accept} + \mu_{unaccept}) / 2$

Define $S = max\ [S_1, S_2, \ldots, S_N]$;

Flag the data set at time *T* if $S > h$, where *h* is chosen via:

Average Run Length $\{ \mu = \mu_{accept} \} = ARL_0$ (False alarm rate)

Notes: (a) $ARL_0 > 50*N$ recommended

(b) *2-sided* test = combo of two *1-sided* tests

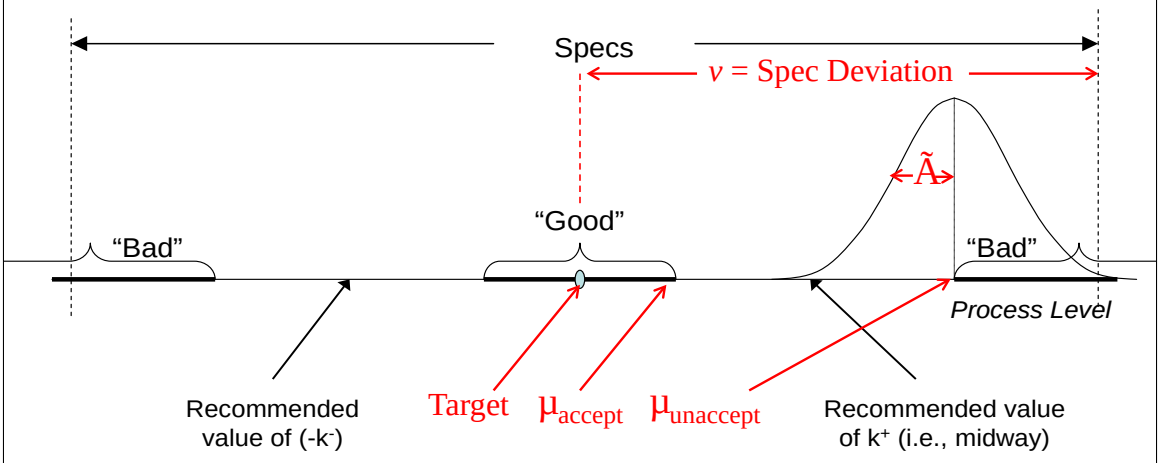
(c) *Supplemental Shewhart* test is needed:

Flag the data set at time *T* if $X_i > c$ for some $i \le N$

5. Target-setting when Specs are present

Usually available: Target, capability Indices, data intensities
Need to find: acceptable/unacceptable mean levels μ_{accept}, $\mu_{unaccept}$

Process Window Configuration (2-sided case)



Target-setting process: Let us make a deal

Performance of the monitoring system will *depend on level of input* that the user (real or virtual) is willing to provide

Complete or partially complete input: Target, Ã, Acceptable/Unacceptable levels, type of control, False Alarm (FA) rate.

spcvar	target	pop. sigma	input format (Abs/Del)	Acceptable level	Unacceptable level	type	FA rate	Unacc. Sigma	spec deviation (for suggesting)	parm description
v001	0	10		-2	-10	2 sym. / 1 side	10	1.5	?:	DMIW ptile Std_N_
v002	18	.15		18.1	19	2 sym. / 1 side	5	1.5	?:	DMIW ptile PSRO_(
v003	5	.01		5.1	5.2	2 sym.	10	1.5	?:	DMIW meas v003 N

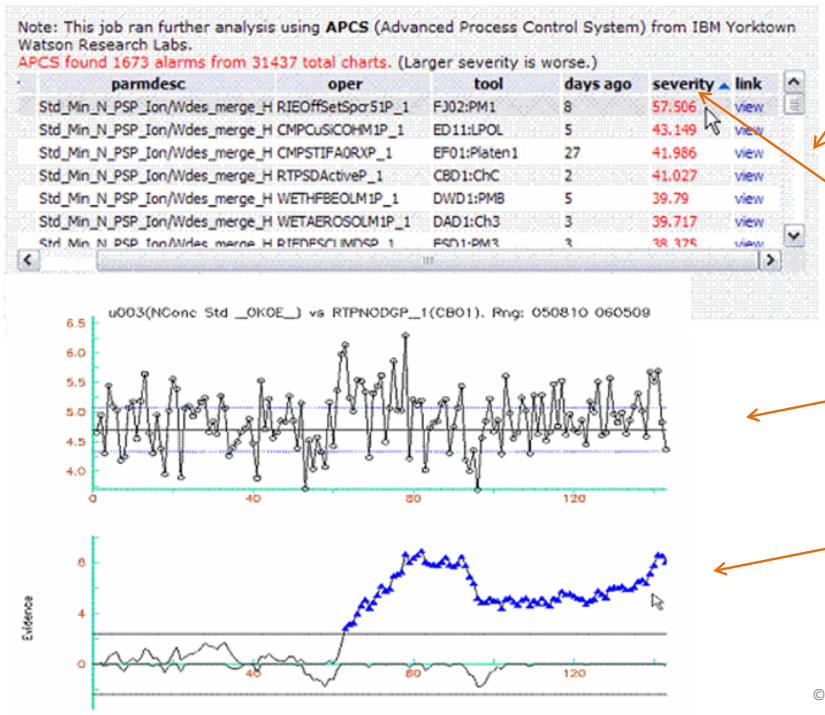
Minimal input: type of control (A), FA rate and "spec deviation" (B)

A. B.

spcvar	target	pop. sigma	input format (Abs/Del)	Acceptable level	Unacceptable level	type	FA rate	Unacc. Sigma	spec deviation (for suggesting)	parm description
v001						2 sym. / 1 side	1000	1.5	?: 45	DMIW ptile Std_N_M1_Ioff . . . . .

Follow-up run and tuning

Inspection of the *dashboard* and of *individual charts*



Flagged cases (*alarms*) are selected

Alarm ranking attributes

One of *flagged* cases (variable u003)

Evidence (Page's) plot

6. Analysis Attributes: Search Engine Log

logb_sample_4.xlsx

	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

Last bad point observed 2675 minutes ago

"Richter scale" magnitude of violation

Forgiveness Achieved (9 = highest)

Forgiveness based on last 12 points

Last violation direction: down

Log "feeds" the Dashboard & Alarm Prioritization

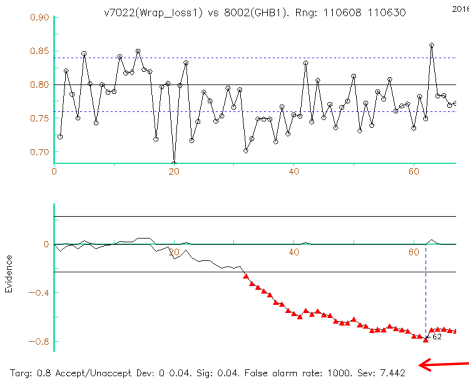
Attribute #1: Severity

Severity: reflects degree of threshold(s) violation

Severity = combination of $-\log_{10}$ (p-value) of the main and supplemental tests. E.g. Severity = $-\log_{10}[\min\{p_1,p_2,p_3\}]$

Can be interpreted as value on "Richter Scale"

Estimated via: Markov Chain, Brownian Motion approximations



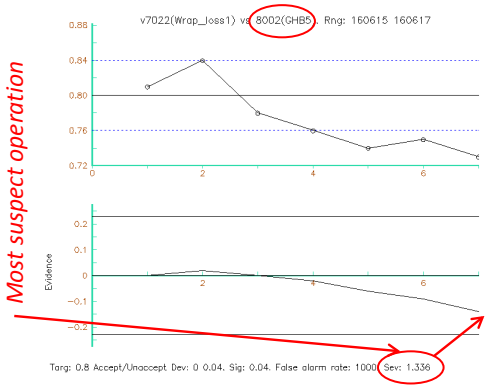
2016

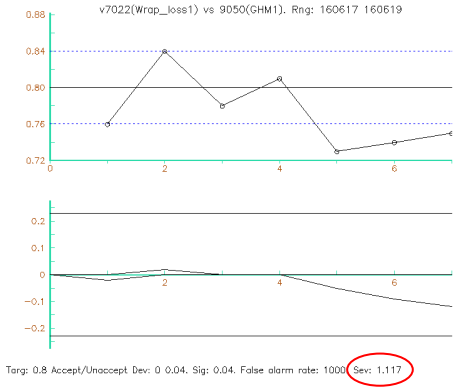
logb_sample_4.xlsx

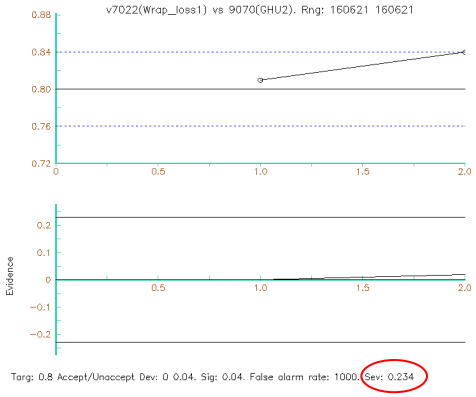
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Variable	Operation	Tool	LastBad	Severity	Bad2End	Npoints	Avg	Stdv	Forgiv_ Ind	Forgiv_ Depth	LastBad_ Ind	LastBad_ Cond
2	v7022	8002	GHB1	-2675	7.442	-387	67	0.774	0.0361	1	5	62	-1
3	v7022	8002	GHB5	0	1.336	0	7	0.773	0.0236	0	7	0	0
4	v7023	8002	GHB1	-2286	10.19	-387	67	-0.832	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

Severity can also be helpful in identifying "guilty" stage

Most suspect operation





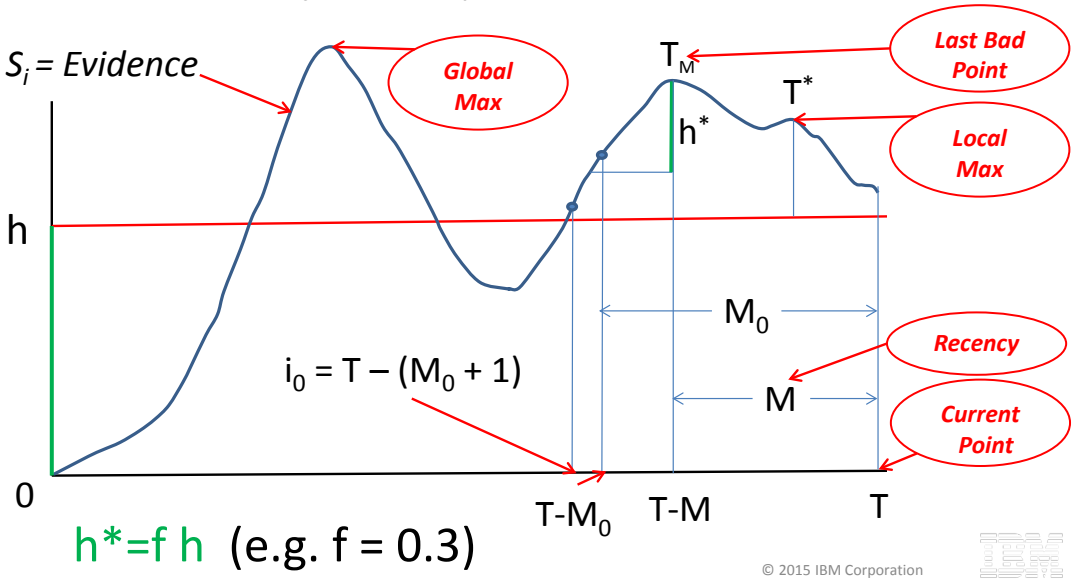


Higher *severity* => a factor in declaring operation 8002 as "more suspect" in having an impact on variable v7022 than other operations

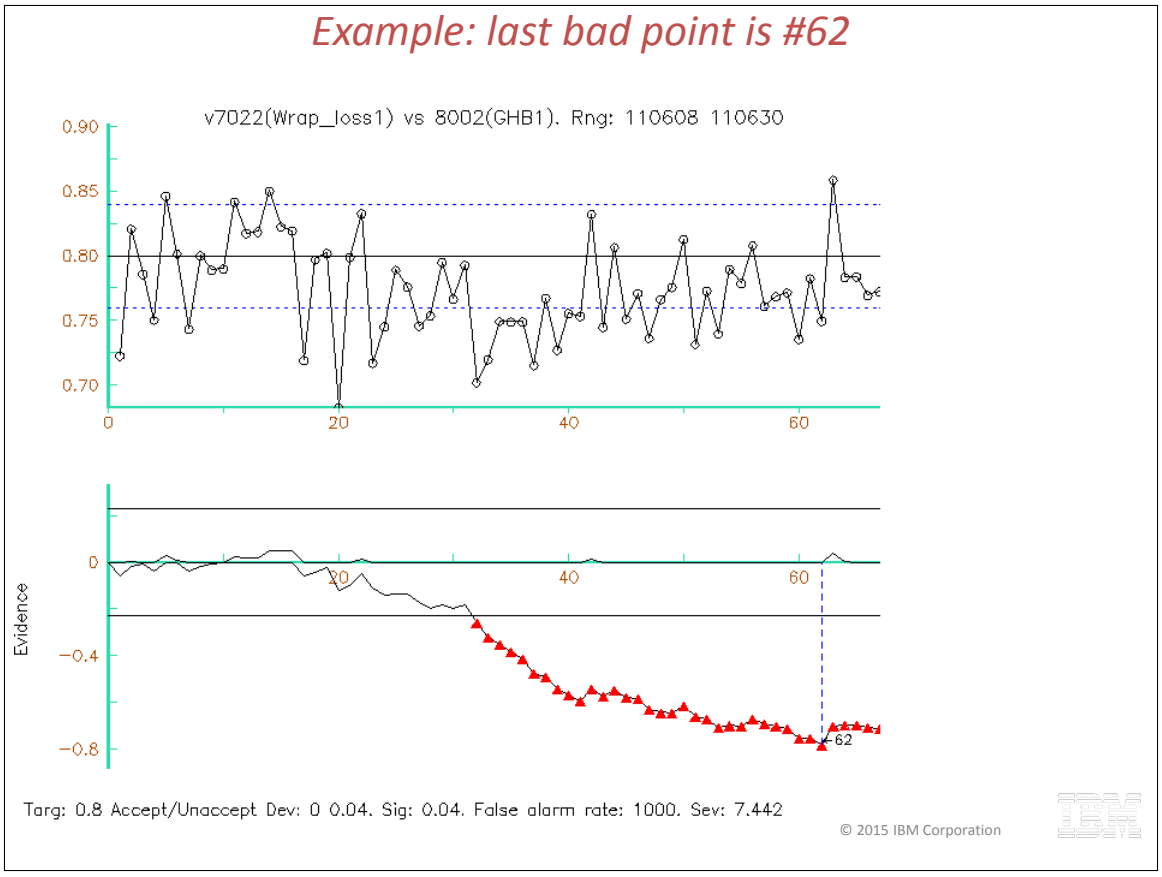
Attribute #2: Recency

Recency: reflects degree of *"newsworthiness"*

Recency = number of periods (or points) elapsed since conditions consistent with unfavorable process level were observed.



$h^* = f\,h$ (e.g. f = 0.3)


Example: last bad point is #62
v7022(Wrap_loss1) vs 8002(GHB1). Rng: 110608 110630
0.90
0.85
0.80
0.75
0.70
0
20
40
60
Evidence
0
−0.4
−0.8
62
Targ: 0.8 Accept/Unaccept Dev: 0 0.04. Sig: 0.04. False alarm rate: 1000. Sev: 7.442
© 2015 IBM Corporation

7. Return to normal

Measured in terms of *Forgiveness Index*

Required: statistical proof that the current process level is acceptable

Acceptability criterion (1-sided upper control): mean μ satisfies

$H_1 : \mu \le \mu_{0\delta} = \mu_{accept} + \delta^*(k - \mu_{accept})$ = *forgiveness level*

Procedure: (a) set *Level of confidence:* $1 > \varepsilon_r > 0$, e.g., $\varepsilon_r = 0.05$.

(b) Compute the *last bad point* T_M (and corresponding *window* M)

(c) For *m = 1, 2, ... M compute*

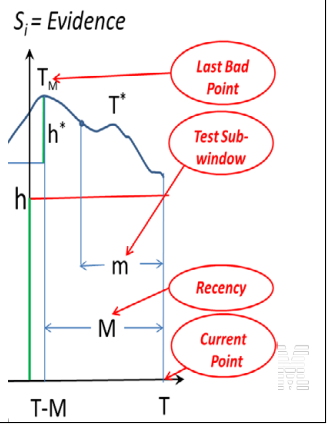
$$Z_m = \frac{\bar{X}_{[m]} - \mu_{0\delta}}{\sigma / \sqrt{m}}, \quad where\ \bar{X}_{[m]} = \frac{X_N + X_{N-1} + \ldots + X_{N-m+1}}{m}$$

and corresponding p-values $p[m] = \Phi(Z_m)$

(c) Denote m^* = argmax $\{p[m]\}$

(d) Assign *forgiveness index* on the scale 0 - 9 based on $p(m^*)$ and p(M). E.g. set *index = 9* if $p(M) < \varepsilon_r$ and $p(m^*) < 0.5$

"Complete" statistical forgiveness



8. Engineering considerations

1. Identification of *operation* most closely associated with emergence of unfavorable conditions can be achieved based on the provided set of analysis attributes – however, fully automating the process can be challenging
2. *Analysis Attributes* require special considerations prior to their use in business decisions, incl. *Alarm prioritization* and *Dashboard logic.*
3. Ex1. Business impact of *high-severity* event typically depends on other factors, that are not reflected directly in search engine.
4. Ex2. Translation of *recency* index to time frame(s) used in decisions can be done in several ways, depending on objectives.
5. Ex3. Statistical *forgiveness* index is only one factor when establishing the degree of *"return-to-normal".*
6. Outlier management policies can be implemented at different stages, affecting decisions.

Concluding Remarks

1. In the environment of massive/intensive data streams *statistical performance* of *decision rules* is a major factor.
2. Special challenge: keeping *False Alarm* rate in check, esp. when amounts of data in *timeslides* are highly variable
3. Approach based on 3-zone *Likelihood Ratio approach* enables efficient deployment of early warning systems based on practical significance of alarms.
4. *Analysis Attributes* play a key role in decision process. Determining the set of attributes sufficient for a given business process is essential. Relative importance of attributes depends on business objectives.
5. *Decision to forgive* is of different nature than decision to alarm
6. Rich field for statistics research, esp. hypothesis testing, time series, decision theory, Change-point theory, sequential analysis, Monte Carlo, Bayesian methods.

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 paper 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 paper touches on statistical sampling plans, statistical process monitoring, and the design and analysis of experiments.

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 and engineers develop solutions based on the insights gained from these approximations. Ultimately, however, people must interact with the concrete universe to confirm the

G. Geoffrey Vining
Department of Statistics, Virginia Tech, Blacksburg, VA 24060, USA, e-mail: vining@vt.edu

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 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 des 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

Bayesian analysis requires a likelihood function to describe the random behavior of the data given the paprameters, a prior distribution on at least one parameter to describe the prior belief about the parameter and to define the uncertainty associated with this prior belief, and a loss function that forms the basis for making decisions. The key to Bayesian analysis is the posterior distribution of the data, which has the form

$$f(\mathbf{y}, \beta) = \int f(\mathbf{y}|\beta) g(\beta) d\beta$$

where

- $g(\beta)$ is the prior distribution on the parameter vector β
- $f(\mathbf{y}|\beta)$ is the likelihood function

Bayesian analysis requires strong distributional assumptions, unlike ordinary least squares. First, the Bayesian analysis assumes a strong distributional form for the fundamental likelihood function. It then adds another strong distribution 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 some optimized expected posterior loss of the form

$$E\left[\ell(\mathbf{y})\right] = \int \ell(\mathbf{y}) f(\mathbf{y}|\beta) g(\beta) d\beta$$

where $\ell(\mathbf{y})$ is an appropriate loss function. The "best" parameter estimates optimize the expected loss function. For example, a Bayesian optimal experimental design optimizes the appropriate loss 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 variance of σ^2 and with a normal prior distribution with mean θ 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. Let $\overline{y}_n$ be the sample mean for a random sample of n observations. With quite a bit of algebra, one can show assuming a quadratic loss function that

$$\mu_p = \frac{\sigma^2\theta + n\tau^2\overline{y}_n}{\sigma^2 + n\tau^2} \tag{1}$$

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. The key point 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. 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 Sampling and Process Monitoring

I worked with Bayesian acceptance sampling plans as a graduate student in the 1980s. Their use was quite limited; however, they did provide insights about the use of Bayesian approaches within industrial statistics.

Consider a simple illustrative example involving a batch production process. The sampling plan focuses on the sample mean of a continuous characteristic. The organization 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 prior distribution is also normal, and Equation (1) gives the resulting posterior mean.

An important question is at what point does the sample mean begin to overwhelm the prior distribution. 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 $1 \leq k \leq 5$. Suppose that the organization has m batches as a base period, and let $\overline{y}_{..}$ is the sample mean for the quality characteristic over the base period. It can be shown that the variance of $\overline{y}_{..}$ is

$$\frac{1}{m}\left[k\sigma^2 + \frac{\sigma^2}{n}\right].$$

For simplicity, assume that $m > 50$ and that n is of moderate size. A reasonable approximation for τ^2 is

$$\tau^2 = \frac{k\sigma^2}{m}.$$

The posterior mean then becomes

$$\mu_p = \frac{\mu_0 + \frac{k}{m} n\overline{y}}{\frac{k}{m} n + 1}$$

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 weight to the sample mean and the mean of the prior ($\nu = .5$). The resulting sample size is $n = 20$. Giving the sample mean only equal weight is not really dominating. Consider the same scenario with $\nu = .9$. The resulting sample size is $n = 90$. The situation gets worse 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 sampling plan is to protect the organization from shipping a bad batch.

The reality is that all production processes are truly dynamic except from for 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 sampling plans ability to detect serious problems. of course, one could use a diffuse prior, but at that point the resulting sampling plan is little more than the frequentist version.

Originally, I had hoped to discover a much richer literature on Bayesian statistical monitoring procedures. I was disappointed to find very little. I should have paid more attention to Woodall and Motgomery (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. 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 one 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 s eries 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 discover 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 logisitc 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. and Watss, D.G. (1988). *Nonlinear Regression Analysis and Its Interpretation*. New York: John Wiley and Sons.

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. and Draper, N.R. (1975). Robust Designs. *Biometrika*, 62, pp. 347-352.

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., and Vining, G.G. (2013). A Tutorial on the Planning of Experiments. *Quality Engineering*, 25, pp. 315-332.

Kiefer, J. (1975). Optimum Design: Variation in Structure and Performance under Change in Criterion. *Biometrika*, 62, pp. 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. and Montgomery, D.C. (2014). Some Current Directions in the Thery and Application of Statistical Process Monitoring. *Journal of Quality Technology*, 46, 1, pp. 78-94.