Research Report
Statistical Research Unit
Department of Economics
Göteborg University
Sweden

Consequences of using the probability of a false alarm as the false alarm measure.

David Bock

Research Report 2007:3
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Statistical Research Unit	Nat: 031-786 12 74	Nat: 031-786 00 00	http://www.statistics.gu.se/
P.O. Box 640 SE 405 30 Göteborg Sweden	Int: +46 31 786 12 74	Int: +46 31 786 00 00	

Consequences of using the probability of a false alarm as the false alarm measure

DAVID BOCK*
Statistical Research Unit, Goteborg University, Sweden

Abstract In systems for on-line detection of regime shifts, a process is continually observed. Based on the data available an alarm is given when there is enough evidence of a change. There is a risk of a false alarm and here two different ways of controlling the false alarms are compared: a fixed average run length until the first false alarm and a fixed probability of any false alarm (fixed size). The two approaches are evaluated in terms of the timeliness of alarms. A system with a fixed size is found to have a drawback: the ability to detect a change deteriorates with the time of the change. Consequently, the probability of successful detection will tend to zero and the expected delay of a motivated alarm tends to infinity. This drawback is present even when the size is set to be very large (close to 1). Utility measures expressing the costs for a false or a too late alarm are used in the comparison. How the choice of the best approach can be guided by the parameters of the process and the different costs of alarms is demonstrated. The technique is illustrated by financial transactions of the Hang Seng Index.

Key words: *Monitoring; Surveillance; Repeated decisions; Moving average; Shewhart method.*

* Correspondence: David Bock, Statistical Research Unit, Goteborg University, P.O. Box 640, SE 405 30 Goteborg, Sweden. E-mail: david.bock@statistics.gu.se

1. Introduction

Online detection of an important change in the underlying process is important in many areas. In economics and finance, we are interested in detecting turning points in the business cycle (Andersson (2002) and Andersson et al. (2006)) and changes in volatility in financial asset returns (see e.g. Schipper and Schmid (2001)), e.g. for timely trading of assets (Bock et al. (2005)). In medicine and public health, we aim at quick detection of e.g. kidney failures (Smith and West (1983)), the most fertile phase of the menstrual cycle (Royston (1991)), a foetal lack of oxygen (Frisén (1992)), and an increased disease incidence (Sonesson and Bock (2003) and Radaelli (1992)). In quality control, if a manufacturing process produces contaminated products, we want to detect it early (Abujiya and Muttlak (2004)).

In a situation where we have repeated decisions, the methodology of statistical surveillance is appropriate. Repeated decisions are also made in sequential analysis, but surveillance is different since even when we conclude that no change has happened, the monitoring is not stopped but continued (the null hypothesis is never accepted). Methods for on-line detection have been developed in different areas (e.g. econometrics and quality control). Much of the work has emerged from the work of Shewhart (1931) and is often referred to as statistical process control or statistical surveillance. In this field the false alarms are often characterized by measures reflecting the timeliness of these, for example the average run length to the first false alarm. For a review of statistical surveillance, see Frisén and de Maré (1991), Srivastava and Wu (1993), Lai (1995), Frisén and Wessman (1999) and Frisén (2003).

On-line detection problems are receiving increasing attention in the econometric literature. In e.g. Chu et al. (1996), Leisch et al. (2000), Carsoule and Franses (2003), Zeileis et al. (2004) and Bock (2006) hypothesis tests for retrospective detection of structural change are combined with the prospective aspect of surveillance, i.e. a hypothesis is repeatedly tested each time a new observation becomes available. The false alarms are controlled by a fixed size during an infinitely long surveillance period (asymptotic size).

In this paper the aim is to compare the behavior of monitoring methods where the false alarms are controlled in either of two ways: by using a fixed asymptotic size or a fixed measure reflecting the timeliness of false alarms. This was briefly discussed in Frisén (1994) and Bock (2006) but no thorough analysis has previously been made.

In on-line detection it is important that the change is detected quickly without having too many false alarms. Therefore, the behavior is investigated in terms of the timeliness of motivated alarms and for different specifications of utility, expressing the different costs for the gain of a motivated alarm and the loss of a false alarm.

The plan of this paper is as follows. Notations and specifications are given in section 2. In section 3 different ways of evaluating surveillance systems are presented and in section 4 the methods under study are presented. In section 5 a comparison is made between the two approaches. We discuss drawbacks and advantages of the different approaches in different situations and specifications of the utility. Some concluding remarks are given in section 6.

2. Notations and specifications

The process under surveillance, denoted by X, is measured at discrete time points t=1, 2, …, where X may be an average or some other derived statistic. We consider a regime shift that occurs in the expected value μ of the process

$$X(t) = \mu(t) + \varepsilon(t) \quad (1)$$

from an acceptable level μ_0 to an unacceptable level μ_1 where μ_0 and μ_1 are constant, $\mu_1 > \mu_0$ and $\varepsilon(t) \sim$iid N[0, σ^2], t=1, 2, ... Model (1) might be too simple for economic time series However the model is often used, and here it is used to emphasize the inferential issues of statistical surveillance. Surveillance of autocorrelated and multivariate processes are studied in e.g. Kalgonda and Kulkarni (2004) and Pan (2005).

The shift occurs at an unknown time point, denoted by τ such that $\mu(t)=\mu_0$ and $\mu(t)=\mu_1$ for $t< \tau$ and $t\geq\tau$, respectively. When $\mu=\mu_0$ the process is said to be in control whereas when $\mu=\mu_1$ it is said to be out-of-control. Other types of changes treated in the literature are e.g. in the monotonicity of μ (Andersson (2002)), in σ^2 (Yeh et al. (2003)) or both μ and σ^2 (Costa and Rahim (2004) and Wu et al. (2005)).

The parameters μ_0, μ_1 and σ^2 are regarded as known. Without loss of generality we impose $\mu_0 = 0$ and $\sigma=1$, i.e. the size of the shift is specified by μ_1. The variable τ is random with a constant intensity $\nu=P(\tau=t| \tau \geq t)$ that is τ has a Geometric distribution on t=1, 2, …, which is a common assumption in the literature, see e.g. Shiryaev (1963) and Frisén and Wessman (1999).

At each decision time s, s=1, 2, …, we make a decision whether there has been a regime shift or not. In statistical surveillance this is expressed as discriminating between two events, C(s) and D(s), where C(s) is the critical event implying that the process is out-of-control and D(s) implies that it is in-control. The two events can be specified in various ways and different methods are optimal for different specifications. For the situation when it is important to see whether there has been a change since the start of the surveillance, the following specification is used

$$C(s)=\{\tau \leq s\} \text{ and } D(s)=\{\tau > s\}.$$

When the monitoring is done from a repeated hypothesis testing angle, then at each time s that a new observation becomes available, we formulate it as a testing of a null hypothesis

$$H_0(s): \text{No change has occurred up to time s,} \quad (2)$$

i.e. $\mu(1)=\mu(2)= \ldots =\mu(s)=\mu_0$. This $H_0(s)$ corresponds to $D(s)=\{\tau > s\}$. The event $C(s)=\{\tau \leq s\}$ corresponds to the alternative hypothesis

$$H_A(s): \text{A change has occurred at some time point } t \leq s,$$

i.e. $\mu(1)=\mu(2)= \ldots =\mu(t-1)=\mu_0$ and $\mu(t)= \ldots =\mu(s)=\mu_1$. Hence, there is a different null and alternative hypothesis for each s.

An alarm set A(s) is constructed, with the property that as soon as $X_s=\{X(1), \ldots, X(s)\} \in A(s)$ we infer that a change has occurred. The alarm set consists of a function $p(X_s)$ and a limit g(s), where the time of an alarm, t_A, is defined as

$$t_A = \min\{s: p(X_s) > g(s)\}.$$

The alarm limit g(s) is determined in order to control the false alarms and this can be done in various ways to be described below.

3. Evaluation in on-line monitoring

The monitoring situation is characterized by repeated decisions as well as not having fixed hypotheses and an increasing sample size. With repeated decisions, it is important to consider the timeliness aspect. In the traditional hypothesis testing framework the behavior of the procedure under the alternative hypothesis is usually characterized by the power. There is however no information in the power about when the alarm was called in relation to the regime shift, for example how long after the shift the alarm was given. A natural evaluation measure in a monitoring situation is instead the delay of a motivated alarm. Desirable properties of a surveillance method are that the delay between the time of the alarm, t_A, and the time of the change, τ, is short and that there are not too many false alarms.

As mentioned above, monitoring is often made by repeatedly testing a hypothesis each time a new observation becomes available. If we define the alarm set such that at each decision time the type I error probability is fixed to e.g. 5%, then the probability of ever falsely rejecting the null hypothesis will tend to 1 as we repeat the test. This has sought to be avoided by instead constructing alarm sets in such a way that this probability is fixed below one.

The probability that a false alarm is given before time i, as $i\rightarrow\infty$, is hereafter referred to as the asymptotic size or α. It is defined as

$$\lim_{i\rightarrow\infty} \alpha(i) = \alpha$$

where $\alpha(i)=P(t_A \leq i|H_0)$ and H_0 is defined in (2), i.e. $\alpha(i)=P(t_A\leq i|\tau>i)$. A sequence of alarm sets is constructed resulting in $\alpha < 1$. It is hence a situation with strict significance testing. When $\alpha<1$ the false alarm probabilities, $P(t_A=i|\tau > i)$, will not sum to 1 and then t_A is not a random but a generalized random variable.

In the methodology of statistical surveillance the type I error is characterized by the run length distribution of the false alarms. Usually in the quality control literature the average run length, conditional of no change, $ARL^0=E[t_A \mid \tau=\infty]$, summarizes the information. A similar measure is the median run length conditional of no change, $MRL^0=Median[t_A \mid \tau=\infty]$. Another summarizing measure is the probability of a false alarm (PFA) where the expectation is taken with respect to the distribution of τ,

$$PFA=P(t_A<\tau)=E_\tau[P(t_A< \tau| \tau=t)].$$

There are several measures which reflect the timeliness of a motivated alarm. In some applications, an alarm that comes too late is of no value. The probability of successful detection within d time units measures how good a method is when we only have a limited time for action. It is defined as

$$PSD(t, d)=P(t_A-\tau<d|t_A \geq \tau, \tau=t)$$

where $d \geq 1$. Timeliness can also be reflected by the delay of a motivated alarm, here presented as the conditional expected delay (CED),

$$CED(t) = E[t_A-t|t_A \geq \tau, \tau=t].$$

An evaluation measure that is often used is the average run length, given a change at the start of the monitoring, $ARL^1=E[t_A \mid \tau=1]$, which equals CED(1)+1. A widely used optimality criteria in the literature on quality control is that of a minimal ARL^1 for a fixed ARL^0. This criterion might be suitable in some situations but there are however some drawbacks with this optimality criterion, see e.g. Frisén (2003).

Another important aspect when evaluating a method, is the trust you should have in an alarm at a specific time. The predictive value of an alarm at time t $PV(i)=P(C(i)|t_A=i)$, suggested by Frisén (1992) reflects the trust of an alarm.

In the utility treated by Girshick and Rubin (1952) and Shiryaev (1963) the gain of an alarm is a linear function of the expected delay and the loss associated with a false alarm is a function of the same difference. The utility is

$$u(t_A, \tau) = \begin{cases} h(t_A - \tau) & , t_A < \tau \\ a_1 \cdot (t_A - \tau) + a_2, & t_A \geq \tau \end{cases} \qquad (3)$$

where $h(t_A-\tau)$ is an arbitrary function. In a situation where the intensity of a change is constant, the full likelihood ratio method (LR, described in section 4.1) maximizes the expected value of the utility, $E[u(t_A, \tau)]$ (see Shiryaev (1963) and Frisén and de Maré (1991)). If $h(t_A-\tau)$ is a constant, $a_1<0$ and PFA is fixed then $E[u(t_A, \tau)]$ is maximized for a minimal expected delay (ED), defined as $E_\tau[ED(t)]$ where $ED(t)=CED(t)\cdot P(t_A \geq t)$. This is sometimes referred to as the expected delay criterion.

4. Methods

4.1. The Shewhart and the Moving average methods in statistical surveillance

It was shown by Frisén and de Maré (1991) that the optimal method for discriminating between events C(s) and D(s) is based on the likelihood ratio (LR) between the events, and an alarm is given when

$$f_{X_s}(x_s|C(s)) \big/ f_{X_s}(x_s|D(s)) = \sum_{t=1}^{s} w(t) \cdot L(s, t) > g(s),$$

where $L(s, t)=f_{X_s}(x_s|\tau=t)/f_{X_s}(x_s|D)$ is the partial likelihood ratio when $\tau=t$, $w(t)=P(\tau=t)/P(\tau \leq s)$ is the weight for L(s, t) and g(s) is a time dependent limit equal to $k\cdot P(\tau \leq s)/P(\tau > s)$, $k > 0$.

Many methods are based on the LR, where the difference depends on how the partial likelihood ratios are weighted. When $C(s)=\{\tau=s\}$ the LR method simplifies to the Shewhart approach which puts all weight to the last partial likelihood ratio and signals an alarm as soon as L(s, s) exceeds the alarm limit. For independent variables with a Gaussian distribution the Shewhart approach gives an alarm as soon as

$$X(s)-\mu_0 > g, \qquad (4)$$

where g is a constant.

When $C(s)=\{\tau=s-p+1\}$ and $D(s)=\{\tau>s\}$, the LR method simplifies to the Moving average (MA) approach which puts all weight on the partial likelihood ratio L(s, s-p+1). For independent variables with a Gaussian distribution, the MA approach gives an alarm as soon as

$$\sum_{i=s-p+1}^{s}\left(x(i)-\mu_0\right) > g, \tag{5}$$

where g is a constant. This approach was studied in e.g. Wong et al. (2004) and Yu and Chen (2005). The methods in (4) and (5) are hereafter referred to as ShewSur and MASur, respectively in order to distinguish methods derived in the literature on surveillance from those of the next section.

In the culture of statistical surveillance, when we compare several methods, their respectively alarm limits are adjusted to yield the same false alarm property (e.g. ARL^0=100). For the ShewSur and MASur methods in (4) and (5), respectively, the probability of exceeding the alarm limit is the same for each decision time s, given that all observations used in the statistic is from the same state. Consequently, $\lim_{i\to\infty} P\left(t_A \le i | \tau > i\right)=1$, i.e. a false alarm will be given with probability 1.

4.2. Shewhart and MA methods modified to allow false alarms controlled by a fixed asymptotic size

If we want a system that satisfies $\alpha < 1$, the alarm limit should not be a constant as above. Leisch et al. (2000) suggested the following alarm limit for decision time s

$$g(s) = \begin{cases} c & , \ s \le e \\ c\cdot\sqrt{\ln s} & , \ \text{else} \end{cases}$$

where and $c>\sqrt{2}$ is a constant to be determined and e is the natural logarithmic base used to ensure that $g(s)\ge c$.

The methods based on the moving sum in (5) but where the constant alarm limit g in (5) is replaced by the limit g(s) above are for p=1 (only the last observation) and p≥2 in (5) hereafter referred to as ShewTest and MATest, respectively.

Theorem: ShewTest with $c>\sqrt{2}$ yield $\alpha < 1$.

Proof: According to theorem 4.1 in Frisén and de Maré (1991), it holds that α<1 if and only if $P(t_A=s|\tau>s,\ t_A\ge s)<1$ for all s and $\sum_{s=1}^{\infty} P\left(t_A = s \mid t_A \ge s,\ \tau > s\right) < \infty$. We have that $P(t_A=s|\tau>s,\ t_A\ge s)=1-\Phi(g(s))<1$ since $\Phi(g(s))>0$ for all s, where $\Phi(\cdot)$ is the standard Normal probability distribution function.

$$\sum_{s=1}^{\infty} P\left(t_A = s \mid t_A \ge s,\ \tau > s\right) = \sum_{s=1}^{\infty}\left(1-\Phi\left(g(s)\right)\right) = \sum_{s=1}^{\infty}\left(2\cdot\pi\right)^{-1/2}\cdot\int_{g(s)}^{\infty}\exp\left(-z^2/2\right)dz$$

$$\le \sum_{s=1}^{\infty}\left(2\cdot\pi\right)^{-1/2}\cdot\int_{g(s)}^{\infty}\frac{z}{g(s)}\cdot\exp\left(-z^2/2\right)dz = \sum_{s=1}^{\infty}\left(2\cdot\pi\right)^{-1/2}\cdot\exp\left\{-g^2(s)/2\right\}\Big/g(s)$$

$$= \left(2/\pi\right)^{1/2}\cdot c^{-1}\cdot e^{-c^2/2} + \sum_{s=3}^{\infty}\left(2\cdot\pi\right)^{-1/2}\cdot\left(c\cdot\sqrt{\ln s}\right)^{-1}\cdot s^{-c^2/2}.$$

The last sum converges for $c>\sqrt{2}$ by Abel's convergence test since the sequence $\left\{\left(c\cdot\sqrt{\ln s}\right)^{-1}\right\}$ is monotone and converges to zero for $c\ne 0$, $s>1$ and $\sum_{s=1}^{\infty} s^{-c^2/2}$ is convergent for $c>\sqrt{2}$. Therefore $\alpha<1$ for $c>\sqrt{2}$.

Leisch et al. (2000) gave a related theorem in continuous time.

5. A comparison between the two approaches

In this section we discuss the two approaches for controlling the false alarms, a fixed $\alpha < 1$ and a fixed ARL^0. We will demonstrate the consequences of these two approaches in terms of the timeliness of alarms. The in- and out-of-control properties are investigated in section 5.1 and 5.2, respectively. The predictive value and the utility of alarms are discussed in section 5.3 and 5.4, respectively.

Chu et al. (1996) assume that sampling under the null hypothesis is costless, whereas resetting the monitoring system after a false alarm is expensive, i.e. false alarms are severe and from this point of view we should set α to a small value, e.g. α=0.10. In a situation where the cost of a false alarm is low, we can instead set α to a large value, e.g. 0.90.

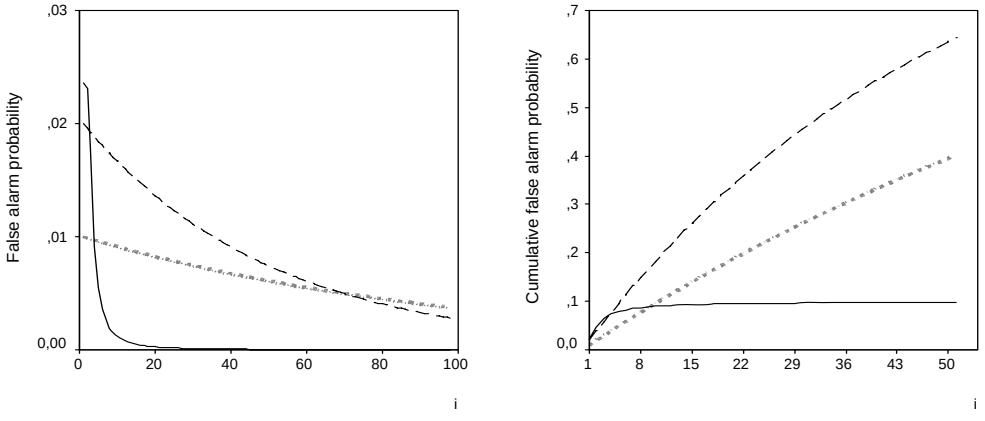
The respective alarm limits of ShewTest and MATest are adjusted to give α={0.10, 0.90}. A low value of ARL^0 can be interpreted as a situation where observations are made seldom and a high value with more frequent observations. ShewSur is adjusted to give ARL^0={50, 100, 250}. The limit of MASur is adjusted to give ARL^0 ={50, 100}. For MASur and MATest, p=2 is considered as in e.g. Vanbrackle and Williamson (1999) and Yu and Chen (2005) and simulations determine the alarm limits.

For all approaches data is collected from time t=1. The alarm statistic of the moving average approach is based on the likelihood ratio L(s, s-p+1) where p=2 and can hence not be constructed at t=1. Therefore, we start the monitoring at t=2 as in Ryan (2000) and Wetherhill and Brown (1991).

To distinguish between the same methods with different values of ARL^0 or α, the value will be given as argument, e.g. ShewSur(50) and ShewTest(0.10).

5.1. In-control properties

In this section, we analyze the in-control behaviour that is the false alarm distributions. In Fig. 1 below, the false alarm probability and the cumulative false alarm probability are shown for ShewSur and the ShewTest.



a b

Figure 1. Panel a: False alarm probability. Panel b: Cumulative false alarm probability. ShewTest(0.10) (———), ShewSur(50) (– – –), ShewSur(100) (▪ ▪ ▪).

The probability of a false alarm for the ShewTest becomes small very fast and almost all alarms are located at early time points (panel a) that is the size level 0.10 is quickly reached (panel b). The pronounced left-skewness in the false alarm density of methods that use a fixed asymptotic size has been pointed out by Chu et al. (1996), Leisch et al. (2000) and Zeileis et al. (2004). The tendency to give early alarms for the test approach is an important difference to the surveillance approach as will be evident in the next section.

The PFA in section 3 summarizes the false alarm distribution in Fig. 1 and is shown in Fig. 2 below. For a constant intensity ν, PFA=1-ν/(1-(1-ν)·Φ(g)) for ShewSur.

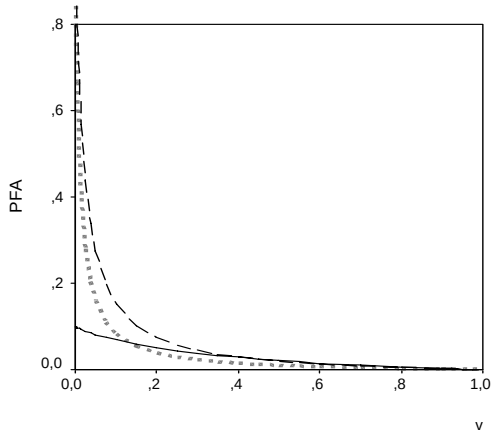


Figure 2. The probability of a false alarm, as a function of the intensity ν. ShewTest(0.10) (——), ShewSur(50) (– –), ShewSur(100) (▪ ▪ ▪).

The difference in level between the two surveillance methods (ShewSur) is due to the difference in the value of ARL^0. When ν→1, PFA tends to zero. The reason is that ν close to 1 implies that the density of τ will be much concentrated to the left and only alarm probabilities at early time points influence PFA. When ν = 1 it follows that P(τ=1)=1 which implies that PFA = 0. When ν→0, the density of τ tends to a uniform distribution, i.e. the regime shift is equally likely to occur early as very late. Most of the alarms are therefore false. When ν→0, PFA for the test and the surveillance methods tends to α and 1, respectively, as is seen in Fig. 2.

Apart from that the alarm probability by construction is zero at the first time point and very high at the first decision time 2, the shapes of the curves of the false alarm probability and the cumulative false probability for MASur and the MATest are very similar to those of the Shewhart approaches (Fig. 1) and therefore not depicted here.

5.2. Out-of-control properties

In this section, we analyze the out-of-control behaviour that is the ability to detect a change. It was proved by Frisén (1994) that methods which have α<1 have a low probability of a late false alarm (a false alarm long after the monitoring has started):

$$\lim_{i\to\infty} \alpha(i)=\lim_{i\to\infty} \sum_{j=1}^{i} P\left(t_A = j \mid t_A \geq j\right)\cdot P\left(t_A \geq j\right)=\alpha < 1$$

$$\Rightarrow \lim_{j\to\infty} P\left(t_A = j \mid t_A \geq j\right) = 0 \text{ since } \alpha > \lim_{i\to\infty} P\left(t_A \geq i\right)\cdot \sum_{j=1}^{i} P\left(t_A = j \mid t_A \geq j\right).$$

This explains the shapes of the false alarm probability of the test approaches in Fig. 1. That the false alarm probability is low might at a first glance seem like a good property. But if $\lim_{j\to\infty} P(t_A=j|t_A \geq j)=0$, then probability to detect a change that happens a long time after the monitoring has started also tends to zero. This was pointed out by Pollak and Siegmund (1975) and Frisén (1994). The reason is that $\lim_{j\to\infty} P(t_A=j|t_A \geq j)=0$, implies that the alarm limit tends to infinity as $j\to\infty$. Therefore also $\lim_{j\to\infty} P(t_A=j|t_A \geq j, \tau=j)=0$. Consequences of this will be illustrated below.

A change occurs at the same time as the surveillance was started ($\tau=1$) is the most widely considered case for evaluation in literature. ARL^1 is the average run length, given a change at the start of the monitoring. This corresponds to $\tau=1$ and $\tau=2$ for the Shewhart and MA approach with p=2, respectively. For ShewSur, $ARL^1=1/(1-\Phi(g-\mu_1))$. The ARL^1 is given the graphs of CED in fig. 4 below (since $ARL^1= CED(1)+1$).

The test approaches yield the smallest ARL^1. Thus in terms of ARL^1, the test approaches are better and the reason is that they allocate the alarms early. This is especially pronounced when $\alpha=0.90$, where the false alarm rate is high as a result of the low alarm limit and this low alarm limit, in turn, results in a short ARL^1. The trust of these early alarms are however low (see section 5.3).

When $\alpha < 1$, the probability of successful detection, PSD in section 3, tends to zero as the time of the change tends to infinity. We have that $PSD(t, d) = \sum_{j=0}^{d-1} P\left(t_A = t+j \mid t_A \geq t, \tau = t\right)$ and for the test methods, $\lim_{t\to\infty} P(t_A=t+j|t_A \geq t, \tau=t)=0$, $j=0, 1, \ldots$ Therefore $\lim_{t\to\infty} PSD(t, d) = 0$ for any $d \geq 1$. For ShewSur and ShewTest PSD(t, d) equals to $1-\Phi(g-\mu_1)^d$ and $1-\prod_{j=0}^{d-1} \Phi\left(g(t+j)-\mu_1\right)$, respectively. For ShewTest, the PSD(t, d) is decreasing (not always strict) with t, since the alarm limit is increasing (i.e. $PSD(t, d) \geq PSD(t+1, d)$ for all t and d, since $g(t) \leq g(t+1)$ for all t and μ_1). The PSD(t, d) curves are shown in Fig. 3 for d=2 and $\mu_1=1$.

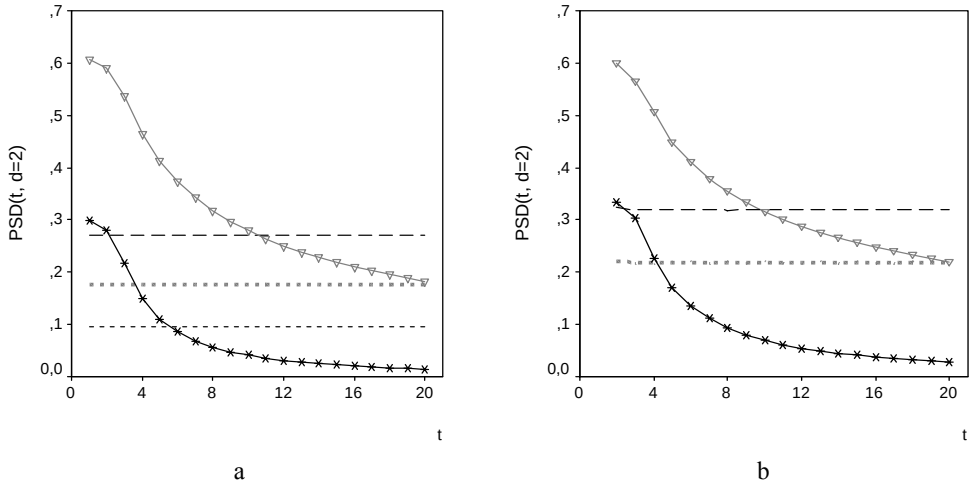


Figure 3. The probability of successful detection PSD(t, d=2) for different values of the time of the change when $\mu_1=1$. Test(0.10) (—— *), Test(0.90) (——▽), Sur(50) (– –), Sur(100) (▪ ▪ ▪), Sur(250) (----). Panel a: ShewSur and Shewtest. Panel b: MASur and MATest.

Since $\lim_{t\to\infty} PSD(t, d) = 0$ for the test methods, these methods have very little chance of detecting a change that occurs late. This drawback can not be overcome by changing α. As seen in Fig. 3, the behavior is the same for α=0.10 and α=0.90 and the difference is mainly in the level but not in the general shape of the curve.

As the probability of a motivated alarm becomes smaller the later the change occurs, the delay of alarms will consequently be higher the later the change occurs, as was pointed out by Pollak and Siegmund (1975). This was in fact noticed by Chu et al. (1996), Leisch et al. (2000) and Zeileis et al. (2004) from simulation experiments. However, it was not recognized as a direct consequence of having $\alpha < 1$ but as a consequence of the way the alarm limit changed with time.

Chu et al. (1996) motivated using a $\alpha < 1$ in terms of the cost of false alarms, but the cost of the delay of motivated alarms was not considered. Here the delay is summarized by the CED in section 3. The CED functions are shown in Fig. 4 for $\mu_1=3$. For ShewSur, $CED(t)=ARL^1-1$ and for ShewTest,

$$CED(t)=t\cdot\left(1-\Phi\left(g(t)-\mu_1\right)\right)+\sum_{i=t+1}^{\infty} i\cdot\left(1-\Phi\left(g(i)-\mu_1\right)\right)\cdot\prod_{j=t}^{i-1}\Phi\left(g(j)-\mu_1\right)-t\,.$$

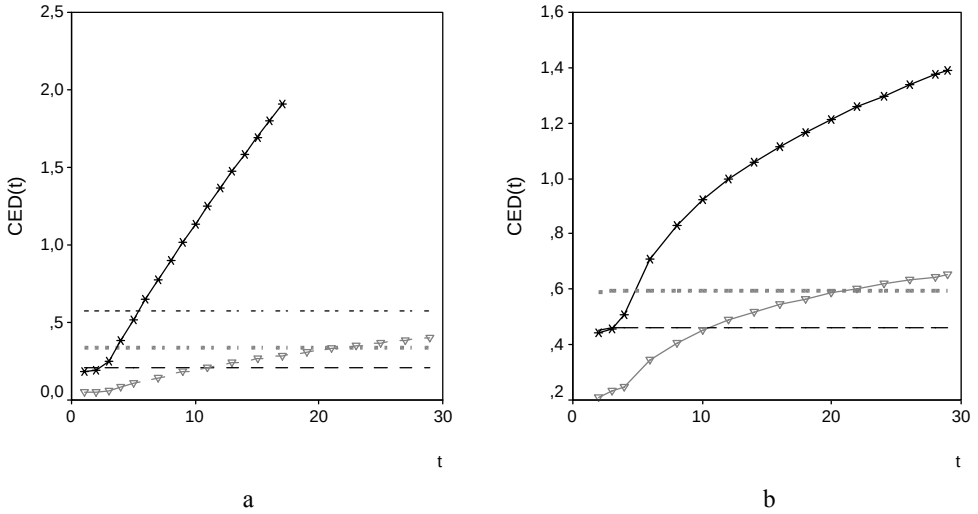


Figure 4. The conditional expected delay CED(t) for different values of t (the time of the change) when $\mu_1=3$. Test(0.10) (—— ✳), Test(0.90) (——▽),Sur(50) (– –), Sur(100) (▪ ▪ ▪),Sur(250) (----). Panel a: ShewSur and Shewtest. Panel b: MASur and MATest.

The CED(t) of the test approaches are seen to increase with t and we confirm what was pointed out by Pollak and Siegmund (1975) and later proved by Frisén (1994); the delay of alarms will be higher the later the change occurs.

Generally, the CED(t) can be written as $\sum_{d=0}^{\infty} P\left(t_A - t > d \mid t_A \geq \tau, \tau = t\right)$, which is the same as $\sum_{d=1}^{\infty}\left(1-PSD(t, d)\right)$. For the ShewTest we have that $PSD(t, d) \geq PSD(t+1, d)$ for all t and d, and then it follows that $CED(t) \leq CED(t+1)$, i.e. CED(t) is increasing with t. Since $\lim_{t\to\infty} PSD(t, d) = 0$ when $\alpha < 1$, CED(t) will tend to infinity as $t\to\infty$ for the test approaches. Comparing the PSD and CED curves of the test approaches for α=0.10 and 0.90, there is a large difference in level but not substantially in the shape. The limited ability to detect changes that occur late remains at any level of α. Though

different alarm limits can increase the detection power at later time points the probability of a motivated alarm will still tend to zero.

5.3. Predictive value

The predictive value at time i, PV(i), reflects the trust of an alarm at that time and can be expressed as

PV(i)=PMA(i)/(PMA(i)+PFA(i))

where PFA(i)=$P(t_A=i|\ i < \tau)\cdot P(\tau > i)$ and PMA(i)=$\sum_{j=1}^{i} P(\tau = j)\cdot P(t_A = i|\tau = j)$ are probabilities of a false and a motivated alarm at time i, respectively. For ShewSur expressions for PMA(i), PFA(i) and the asymptote $\lim_{i\to\infty} PV(i)$ are given in Frisén (1992) for a two-sided case, but they can easily be expressed for our one-sided case. For ShewTest, PFA(i)=$(1-\nu)^i \cdot (1-\Phi(g(i)))\cdot \prod_{j=1}^{i-1}\Phi(g(j))$ and

$$PMA(i)=\sum_{j=1}^{i} \nu\cdot(1-\nu)^{j-1}\cdot\prod_{t=1}^{j-1}\Phi(g(t))\cdot\prod_{t=j}^{i-1}\Phi(g(t)-\mu_1)\cdot(1-\Phi(g(i)-\mu_1)).$$

The PV is shown as a function of the time of the alarm in Fig. 5. The shapes of the curves of the PV for MASur and the MATest are very similar to those of the Shewhart approaches and therefore not depicted here.

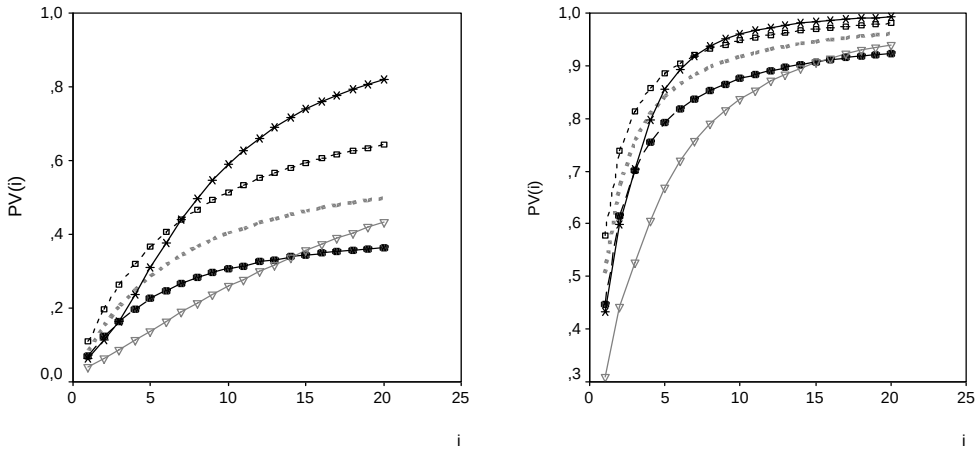


Figure 5. The predictive value as a function of i when μ_1=1. Test(0.10) (—— ✳), Test(0.90) (——▽), Sur(50) (– – ●), Sur(100) (▪ ▪ ▪),Sur(250) (- - - - □). Panel a: ShewSur and Shewtest, ν=0.01, Panel b: ShewSur and Shewtest, ν=0.1.

The test approaches have predictive values that are lower than the surveillance approaches at early time points. Alarms given early by the test approach are therefore not reliable. The opposite relation appears at late time points. However the probability to get a late alarm with the test approach is very low. Thus the better predicted value in this case has no practical importance.

For all approaches under consideration, especially the test approach, the predictive value varies substantially with time. A constant predictive value with respect to time can be a good property as it simplifies matters if the same action can be used whenever an alarm occurs. For the LR method, the predictive value was found by Frisén and Wessman (1999) to be relatively constant.

5.4. Utility

Timeliness can be measured indirectly as the utility of an action after an alarm is given. In a situation where the intensity of a change is constant, the LR method maximizes the expected value of the utility function (3), $E[u(t_A, \tau)]$. The LR does not have a fixed size below one. Methods which have a fixed size will, as pointed out by Frisén (1994), not be optimal in the sense they maximize $E[u(t_A, \tau)]$. Now we discuss some factors influencing $E[u(t_A, \tau)]$ and illustrate the calculation of it.

5.4.1. Example: Trading Hang Seng Index

The techniques of using a utility function to determine which method to choose will now be illustrated by (a slightly simplified version of) the problem of timely trading of the Hang Seng Index (HSI). HSI is a marked-value weighted index of the stock prices of the 33 largest companies on the Hong Kong stock market.

Bock et al. (2005) and Lam and Yam (1997) considered trading closing HSI using different surveillance systems. The aim was to timely detect turning points and trade units of HSI as soon as an alarm was given that a turn had occurred. An assumption made was that the logarithm of the price in Hong Kong dollar had a piecewise linear trend around the turn (a linear regression on time, where the slope changes sign at the turn). A turn then implies a shift from one constant mean level to another of the differentiated series. The case of a peak corresponds to a change from a positive to a negative level ($\mu_0 \geq 0$, $\mu_1 < 0$) or vice versa in case of a trough. One of the methods considered by Lam and Yam (1997) further relied on the assumption that the slope of the linear trend is equally steep before and after the turn, which in the case of a trough implies that $\mu_0=-\mu_1$, $\mu_1 > 0$, for the differentiated series.

5.4.2. Utility and return

The $E[u(t_A, \tau)]$ depends on the false alarm behavior and the delay properties of motivated alarms. Depending on which function is chosen for $h(t_A-\tau)$ in (3), the $E[u(t_A, \tau)]$ will be influenced by the false alarms in different ways.

What are reasonable specifications of the utility function? One measure of the gain of an action is the return earned by timely trading financial assets. It is often measured along the log-price scale. If the asset is bought at $t=0$ and sold at $t=t_A$, the return (r) can be defined as

$$r(t_A)=c+x(t_A)-x(0)$$

where X is the logarithm of the price and $c \leq 0$ would depend on e.g. the transaction cost. The utility function can be defined as the expected return, i.e.

$$u(t_A, \tau) = E[r(t_A) \mid t_A, \tau]. \qquad (6)$$

Hence forward we exemplify the turning point with a peak and in that situation, $E[r(t_A)|t_A, \tau]$ is maximized at the peak, i.e. when $t_A=\tau-1$.

The return does not explicitly take the timeliness of an alarm into account. However, when the return is a known function of time, return and timeliness are related as discussed in Bock et al. (2005). If the X function in the return expression above can be modelled by a piecewise linear trend, then (6) can be written as

$$u(t_A, \tau)= c + \begin{cases} \mu_0 \cdot t_A, & t_A < \tau \\ \mu_0 \cdot (\tau\text{-}1) + \mu_1 \cdot (t_A - \tau + 1), & t_A \geq \tau \end{cases}. \quad (7)$$

where τ is the first time after the peak and μ_0 and μ_1 are the pre-peak slope and post-peak slope, respectively. In some cases, e.g. that of trading HSI, (7) is a reasonable specification of the utility function. The expected value of (7) depends on the behavior of both false and motivated alarms and given $\tau=t$, $E[u(t_A, \tau)|\tau=t]$ equals to

$$c+\mu_0 \cdot E[t_A|t_A<\tau, \tau=t] \cdot P(t_A<\tau) + \{\mu_0 \cdot (\tau-1) + \mu_1 \cdot E[t_A - t + 1 | t_A \geq \tau, \tau=t]\} \cdot P(t_A \geq \tau).$$

When we summarize the whole utility with respect to the distribution of τ, we get

$$E[u(t_A, \tau)] = c + \mu_0 \cdot \{EFA + E_\tau[\tau \cdot P(t_A \geq \tau)] - E_\tau[P(t_A \geq \tau)]\} + \mu_1 \cdot \{ED + E_\tau[P(t_A \geq \tau)]\}$$

where $EFA=E_\tau[E[t_A|t_A<\tau, \tau=t] \cdot P(t_A<\tau)]$ and ED is the expected delay.

For a constant intensity ν, $E[u(t_A, \tau)]$ of ShewSur equals to

$$E\left[u(t_A, \tau)\right] = c + \mu_0 \cdot \left\{ \left(1-\Phi(g)\right)^{-1} \cdot \left(1 - \frac{\nu}{1-(1-\nu)\cdot\Phi(g)} \right) \right\} + \mu_1 \cdot \left(\frac{\nu \cdot ARL^1}{1-(1-\nu)\cdot\Phi(g)} \right).$$

For ShewTest the values of the components can be numerically approximated. For ShewTest $ED = \sum_{t=1}^{\infty} P(\tau = t) \cdot \sum_{i=t}^{\infty} (i\text{-}t) \cdot P(t_A = i \mid t_A \geq \tau, \tau = t) \cdot P(t_A \geq \tau)$. A lower boundary for ED is

$$\sum_{t=1}^{T'} P(\tau = t) \cdot \left\{ \sum_{i=t}^{T} (i\text{-}t) \cdot P(t_A = i | t_A \geq \tau, \tau = t) + (T + 1\text{-}t) \cdot P(t_A > T | t_A \geq \tau, \tau = t) \right\} \cdot P(t_A \geq \tau).$$

In the calculations of ED in section 5.4.4 below, the lower bound is calculated using $T= 200$ and $T'= 100$. T represents the number of t_A-points and T' the values of τ used in the calculation. Basing $u(t_A, \tau)$ on values of t_A up to 200 and τ up to 100 is reasonable in view of the situation at hand with the length of a cycle (trough to trough) of approximately 100 days (see section 5.4.3 and 5.4.4).

5.4.3. The costs of different errors

The relation between the costs for false alarms and the delay of a motivated alarm determines the relative importance of the false alarm distribution and the delay properties in influencing $E[u(t_A, \tau)]$ and that determines which of the surveillance and test approaches that gives the best utility.

Chu et al. (1996) did not take the cost of the delay of motivated alarms into account. This means that the gain of an action caused by a motivated alarm does not

depend on the delay, i.e. $a_1 = 0$ in (3). If $a_1=0$, then the maximization of $E[u(t_A, \tau)]$ would imply a method which never gives an alarm. A less extreme case is when the loss of a false alarm is relatively large compared to the gain of a motivated alarm. Then the false alarm properties would still dominate the utility.

The transaction cost differs between types of investors and can sometimes be negligible. We will use no transaction cost (c=0 in (7)) in the utility illustration below.

The period February 10 to May 28, 1999 for HSI (analyzed by Bock et al. (2005)) including a peak is used to estimate reasonable values for the parameters in the utility expression. The pre-peak slope is slightly steeper than the post-peak slope (the ratio between the slopes is 1.09). In the illustration below, a symmetric peak is considered a reasonable approximation, i.e. $\mu_0=-\mu_1=\mu > 0$ where $\mu=(|\mu_0|+|\mu_1|)/2=0.0069$. Then $E[u(t_A, \tau)]=\mu\cdot\{EFA–ED+E[\tau\cdot P(t_A \geq \tau)]-2\cdot E[P(t_A\geq \tau)]\}$, which is maximized for a minimal $E[|t_A-(\tau-1)|]$. That is because

$$E[|t_A-(\tau-1)|]= -\{EFA-ED-E_\tau[\tau\cdot P(t_A< \tau)]+E_\tau[P(t_A< \tau)]-E_\tau[P(t_A\geq \tau)]\}$$

which is equals to $-\{E[u(t_A, \tau)]-\max_{tA}\{E[u(t_A, \tau)]\}\}/\mu$ where $\max_{tA}\{E[u(t_A, \tau)]\}= E_\tau[(\tau-1)]$. Since for a given ν, $E_\tau[(\tau-1)]$ is a known constant, the minimization of $E[|t_A-(\tau-1)|]$ is the same as maximizing the utility.

5.4.4. The influence of the parameters of the process

In what ways do the intensity ν and the shift size μ_1 influence $E[u(t_A, \tau)]$? The false alarm distribution depends on ν and the delay properties depend on both ν and μ_1. The smaller the size of the shift, the larger the delay and the larger the impact of ED on $E[u(t_A, \tau)]$ as compared to the impact of EFA. Thus for very small shifts, the utility is dominated by the delay and the cost of it. Thus for small shifts the surveillance approach will be preferred since the delay is shorter.

If on the other hand the size of the shift tends to infinity, the delay is small and the false alarm distribution and the cost of false alarms are instead of major importance.

Reasonable values of the shift size μ_1 vary in different practical situations. For the above mentioned period of HSI the standardized ($\mu_0=0$ and $\sigma^2=1$) downward shift (negative μ_1) in the differences had an by Bock et al. (2005) estimated size of 0.82. For a shift of such size, the level of the CED curve for the Shewhart approaches will be substantially higher than in Fig. 4 where $\mu_1=3$, so it is reasonable to say that much concentration is on the delay. Then the surveillance approach will be preferred except possibly for very large values of ν. In the above mentioned period of HSI, ν was estimated to 0.018, which is not very large.

As an illustration we calculate ED, PFA and $E[u(t_A, \tau)]$ for the estimated parameters of the period of HSI. With the costs and parameter discussed above we have the results in Table 1.

Table 1. Results from the illustration on HSI

	ShewSur(50)	ShewSur(100)	ShewTest(0.10)
ED	3.994	9.191	112.164
PFA	0.519	0.351	0.0913
$E[u(t_A, \tau)]$	0.149	0.174	-0.442

The utility in (6) depends on the return, r, which is a function of the price at time t, p(t). If we approximate $E[p(t_A)/p(0)]$ by $\exp\{E[u(t_A,\tau)]\}$, the price at which the HSI is sold is, on the average, 16% higher than it was bought for, for ShewSur(50). The

corresponding figure for ShewSur(100) is 19%. Due to the truncation when calculating ED for ShewTest(0.10) (see section 5.4.2) the value is less than -0.442 for the utility. The price at which the HSI is sold is hence on average less than 64% of the price it was bought for. The ShewTest will here yield such large delays that an alarm will be of no practical value. This illustrates that in the current setting the test approach is not a reasonable method.

6. Discussion and concluding remarks

The properties of two approaches for monitoring have been investigated. The approaches that are compared here differ with respect to how the false alarms are controlled: by a fixed asymptotic size (below 1) or by a fixed measure reflecting the timeliness of the false alarms (e.g. ARL^0).

To use a monitoring method with a fixed size (a test method) is convenient in the sense that ordinary statements of hypothesis testing can be made. One argument against controlling by a fixed size is that ordinary statements for hypothesis testing do not consider the timeliness of alarms.

The use of a fixed size gives the result that the probability of making an alarm long after the monitoring has started is very low. A consequence of this is a limited ability to detect late occurring regime shifts. This drawback can not be adjusted by choosing a large asymptotic size. It remains at any level of the size. Though different alarm limits can increase the detection power at later time points, the probability of a motivated alarm will still tend to zero.

The methods under study that are controlled by a fixed size yield many early but few late alarms compared to the surveillance methods, where the timeliness of false alarms are controlled. The alarms given early by the test methods are less reliable, as measured by the predictive value, compared to those of the surveillance methods. The predictive value of the test methods is higher at late alarms but has no practical importance since the alarms are rare and tend to be given with great delay.

In order to compare different monitoring methods, a utility function can be used. The utility often consists of two parts: one concerning the false alarms and the other the delay of motivated alarms. Chu et al. (1996) argue in favor of a fixed size when sampling under the null hypothesis is costless but resetting the monitoring system after a false alarm does create a large cost. In terms of the utility this means the cost of the delay of an alarm is ignorable, compared to the cost of a false alarm.

Which of the two approaches that is best in terms of utility depends on the specification of the utility function and the relation between the costs of an alarm that is given too early or too late. Also the parameters of the process have an influence as they affect the false alarm and delay properties. If the size of the shift tends to infinity, the delay is small and the false alarm distribution and the cost of false alarms are instead of major importance. As the false alarms are fewer for the test approaches, these might then be preferred.

When the aim is on-line detection, and not hypothesis testing, methods for surveillance are suitable, as they have high probability to detect regime shifts at early as well as late time points.

Acknowledgements

This paper is a part of a research project on statistical surveillance at Göteborg University, supported by the Swedish Council for Research in the Humanities and

Social Sciences and by the Bank of Sweden Tercentenary Foundation. I would like to thank my supervisor, Professor Marianne Frisén and my assistant supervisor Dr Eva Andersson for guidance and support during the work of this report. The work on this report has also been supported by Kungliga and Hvitdfeldtska Stiftelsen and Wilhelm and Martina Lundgrens Vetenskapsfond 1.

References

ABUJIYA, M. R. and MUTTLAK, H. A. (2004) Quality Control Chart for the Mean using Double Ranked Set Sampling, **Journal of Applied Statistics**, 31, (10) pp. 1185 - 1201.

ANDERSSON, E. (2002) Monitoring Cyclical Processes. A Non-parametric Approach, **Journal of Applied Statistics**, 29, (7) pp. 973-990.

ANDERSSON, E., BOCK, D. and FRISÉN, M. (2006) Some statistical aspects on methods for detection of turning points in business cycles., **Journal of Applied Statistics**, to appear.

BOCK, D. (2006) Online testing for switching volatility. **In:** M. WLADYSLAW and P. WDOWINSKI (Ed.), **To appear in Advances in Financial markets Analysis: Principles of Modelling, Forecasting, and Decision-Making.** 2.(Lodz University Press).

BOCK, D., ANDERSSON, E. and FRISÉN, M. (2005) The relation between statistical surveillance and certain decision rules in finance, **Submitted**.

CARSOULE, F. and FRANSES, P. H. (2003) A Note on Monitoring Time-Varying Parameters in an Autoregression, **Metrika**, 57, (1) pp. 51-62.

CHU, C.-S. J., STINCHCOMBE, M. and WHITE, H. (1996) Monitoring structural change, **Econometrica**, 64, (5) pp. 1045-1065.

COSTA, A. F. B. and RAHIM, M. A. (2004) Monitoring Process Mean and Variability with One Non-central Chi-square Chart, **Journal of Applied Statistics**, 31, (10) pp. 1171 - 1183.

FRISÉN, M. (1992) Evaluations of Methods for Statistical Surveillance, **Statistics in Medicine**, 11, pp. 1489-1502.

FRISÉN, M. (1994) Statistical Surveillance of Business Cycles, **Research report** 1994:1 (Revised 2000), Department of Statistics, Göteborg University, Sweden.

FRISÉN, M. (2003) Statistical Surveillance. Optimality and Methods, **International Statistical Review**, 71, (2) pp. 403-434.

FRISÉN, M. and DE MARÉ, J. (1991) Optimal Surveillance, **Biometrika**, 78, pp.271-80.

FRISÉN, M. and WESSMAN, P. (1999) Evaluations of likelihood ratio methods for surveillance. Differences and robustness., **Communications in Statistics. Simulations and Computations**, 28, (3) pp. 597-622.

GIRSHICK, M. A. and RUBIN, H. (1952) A Bayes approach to a quality control model., **Annals of Mathematical Statistics**, 23, pp. 114-125.

KALGONDA, A. A. and KULKARNI, S. R. (2004) Multivariate quality control chart for autocorrelated processes, **Journal of Applied Statistics**, 31, (3) pp. 317-327.

LAI, T. L. (1995) Sequential changepoint detection in quality control and dynamic systems, **Journal of the Royal Statistical Society B**, 57, pp. 613-658.

LAM, K. and YAM, H. C. (1997) CUSUM Techniques for Technical Trading in Financial Markets, **Financial Engineering and the Japanese Markets**, 4, (3) pp. 257-74.

LEISCH, F., HORNIK, K. and KUAN, C.-M. (2000) Monitoring structural changes with the generalized fluctation test, **Econometric Theory**, 16, (6) pp. 835-854.

PAN, X. (2005) An alternative approach to multivariate EWMA control chart, **Journal of Applied Statistics**, 32, (7) pp. 695-705.
POLLAK, M. and SIEGMUND, D. (1975) Approximations to the Expected Sample Size of Certain Sequential Tests, **Annals of Statistics**, 3, (6) pp. 1267-1282.
RADAELLI, G. (1992) Using the Cuscore technique in the surveillance of rare health events, **Journal of Applied Statistics**, 19, pp. 75-81.
ROYSTON, P. (1991) Identifying the fertile phase of the human menstrual cycle, **Statistics in Medicine**, 10, pp. 221-240.
RYAN, T. P. (2000) **Statistical methods for quality improvement** (New York, John Wiley & Sons).
SCHIPPER, S. and SCHMID, W. (2001) Sequential Methods for Detecting Changes in the Variance of Economic Time Series, **Sequential Analysis**, 20, (4) pp. 235-262.
SHEWHART, W. A. (1931) **Economic Control of Quality of Manufactured Product** *(*London, MacMillan and Co.).
SHIRYAEV, A. N. (1963) On optimum methods in quickest detection problems., **Theory of Probability and its Applications.**, 8, pp. 22-46.
SMITH, A. F. and WEST, M. (1983) Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter, **Biometrics**, 39, pp. 867-878.
SONESSON, C. and BOCK, D. (2003) A review and discussion of prospective statistical surveillance in public health, **Journal of the Royal Statistical Society A**, 166, (1) pp. 5-21.
SRIVASTAVA, M. S. and WU, Y. (1993) Comparison of EWMA, CUSUM and Shiryayev-Roberts Procedures for Detecting a Shift in the Mean., **Annals of Statistics**, 21, (2) pp. 645-670.
VANBRACKLE, L. and WILLIAMSON, G. D. (1999) A study of the average run length characteristics of the National Notifiable Diseases Surveillance System, **Statistics in Medicine**, 18, (23) pp. 3309-3319.
WETHERHILL, G. B. and BROWN, D. W. (1991) **Statistical Process Control: Theory and Practice** *(*London, Chapman and Hill).
WONG, H. B., GAN, F. F. and CHANG, T. C. (2004) Designs of moving average control chart, **Journal of Statistical Computation and Simulation**, 74, (1) pp. 47 - 62.
WU, Z., TIAN, Y. and ZHANG, S. (2005) Adjusted-loss-function charts with variable sample sizes and sampling intervals, **Journal of Applied Statistics**, 32, (3) pp.221-242.
YEH, A. B., LIN, D. K. J., ZHOU, H. and VENKATARAMANI, C. (2003) A multivariate exponentially weighted moving average control chart for monitoring process variability, **Journal of Applied Statistics**, 30, (5) pp. 507-536.
YU, F. J. and CHEN, Y.-S. (2005) Economic Design of Moving Average Control Charts, **Quality Engeneering**, 17, (3) pp. 391-397.
ZEILEIS, A., LEISCH, F., KLEIBER, C. and HORNIK, K. (2004) Monitoring structural change in dynamic econometric models, **Journal of Applied Econometrics**, 20, (1) pp. 99-121.

Research Report

2007:1	Andersson, E.:	Effect of dependency in systems for multivariate surveillance.
2007:2	Frisén, M.:	Optimal Sequential Surveillance for Finance, Public Health and other areas.