Research Report
Statistical Research Unit
Göteborg University
Sweden

Aspects on the control of false alarms in statistical surveillance and the impact on the return of financial decision systems

David Bock

Research Report 2004:2
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Statistical Research Unit	Nat: 031-773 12 74	Nat: 031-773 10 00	http://www.stat.gu.se/stat
P.O. Box 660	Int: +46 31 773 12 74	Int: +46 31 773 10 00	
SE 405 30 Göteborg			
Sweden			

ASPECTS ON THE CONTROL OF FALSE ALARMS IN STATISTICAL SURVEILLANCE AND THE IMPACT ON THE RETURN OF FINANCIAL DECISION SYSTEMS

By David Bock

Statistical Research Unit, Göteborg University

ABSTRACT

Systems for on-line detection of regime shifts are important, e.g. for making timely financial transactions. For daily data, it means that we make a new decision each day, based on the data available, and when there is enough evidence of a regime shift, an alarm is called. There is always the risk of a false alarm and here two principally different ways of controlling the false alarms are compared: systems with a fixed average run length until the first false alarm, and systems with a fixed probability (<1) of any false alarm (fixed size). The effects of 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 rapidly 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 are used in the investigation, expressing the different costs for the gain of a motivated alarm and the loss of a false alarm. Drawbacks and advantages of the two approaches are investigated. How the choice of the best approach can be guided by the parameters of the process and the relation between the cost of a too early or too late alarm is demonstrated. The technique is illustrated by application to transactions of the Hang Seng Index.

Key words: *monitoring, surveillance, repeated decisions, moving average; Shewhart method.*

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 et al. (2004)), changes in volatility in financial asset returns (Severin and Schmid (1998), Severin and Schmid (1999), Schipper and Schmid (2001a), Schipper and Schmid (2001b)), e.g. for timely trading of financial assets (Bock et al. (2003)). 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)). In quality control, if a manufacturing process produces contaminated products, we want to detect it early (Wetherhill and Brown (1991)).

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 this type of on-line detection have been developed in different areas (e.g. econometrics and quality control). Much of the work has emerged from the pioneering work of Shewhart (1931) and it 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), Wetherhill and Brown (1991), Srivastava and Wu (1993), Lai (1995), Frisén and Wessman (1999) and Frisén (2003).

On-line detection problems are receiving increasing attention in econometric literature. In the papers by Chu et al. (1996), Carsoule and Franses (1999), Leisch et al. (2000), Carsoule and Franses (2003), Bock et al. (2004) and Zeileis et al. (2004) 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. Chu et al. (1996) implemented tests based on cumulative sums of observations in a prospective setting to monitor the stability of regression parameters in a model for the conditional mean. Leisch et al. (2000) and Zeileis et al. (2004) used statistics based on a moving window of observations. Carsoule and Franses (1999) and Bock et al. (2004) considered prospective tests of stability of the variance and Carsoule and Franses (2003) joint tests for changes in the autoregressive and variance parameters of an autoregression.

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.

In on-line detection it is not only important that the change is detected, it is also important that it is made quickly without having too many false alarms, i.e. the timeliness of alarms is relevant. Therefore, the behavior is investigated in terms of the timeliness of motivated alarms and 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 $t=\{1, 2, \ldots\}$. The observation X may be some summary statistic such as an average or the residual of an estimated time series model. We consider the case of a regime shift that occurs in the expected value μ of the process

$$X(t) = \mu(t) + \varepsilon(t) \tag{1}$$

from an acceptable level μ_0 to an unacceptable level μ_1 where $\mu_1 > \mu_0$ and $\varepsilon(t)$~iid $N[0, \sigma^2]$, t=1, 2, ... The assumption in (1) is in general too simple for economic time series and extensions could be motivated by the features of the data or economic theory. Model (1), however, is used here to emphasize the inferential issues. Without loss of generality we impose $\mu_0 = 0$ and standard deviation $\sigma = 1$. Hence the size of the shift is specified by μ_1. The shift occurs at an unknown time point, denoted by τ such that

$$\mu(t)=\begin{cases}\mu_0, & t < \tau \\ \mu_1, & t \geq \tau\end{cases}.$$

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. The parameters μ_0 and μ_1 are regarded as known and τ is a random variable with intensity

$$\nu_t = P(\tau = t | \tau \geq t).$$

In this paper we treat the case of a constant intensity ν that is τ has a Geometric distribution.

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 the process 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

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

where $C(s)=\left\{\bigcup_{t=1}^{s} C_t\right\}$ and where $C_t=\{\tau = t\}$, t={1, 2, ..., s}, are disjoint. Thus for a change in μ we have

$$C_t=\{\mu(1)=\mu(2)= \dots =\mu(t-1)= \mu_0 \text{ and } \mu(t)= \dots =\mu(s) = \mu_1\}$$
$$D(s)=\{\mu(1)=\mu(2)= \dots =\mu(s) = \mu_0\}.$$

When the monitoring is done from a repeated hypothesis testing angle, then at each time s 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. $\mu(1)=\mu(2)= \dots =\mu(s) = \mu_0$. 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. $\mu(1)=\mu(2)= \dots =\mu(t-1)= \mu_0$ and $\mu(t)= \dots =\mu(s) = \mu_1$. 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 $X_s \in A(s)$ 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

In the traditional hypothesis testing framework the behavior of the procedure under the null hypothesis is usually characterized by the error probability P(reject H_0| H_0 true), referred to as the size. The evaluation under the alternative hypothesis is made using the power, P(reject H_0| H_A true). Since this paper concerns on-line monitoring, we want an evaluation with a timeliness aspect. There is 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. In the classical hypothesis testing situation, we only make one decision: can the null hypothesis be rejected or not? The monitoring situation on the other hand is characterized by repeated decisions as well as not having fixed hypotheses and an increasing sample size.

Many methods that were originally developed for testing one hypothesis are actually used for on-line detection, which involves repeated decisions. With repeated decisions it is important to consider the timeliness aspect in the evaluation. A natural evaluation measure in an on-line situation is 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, on-line 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 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 tends to infinity, is hereafter referred to as the asymptotic size or α. It is defined as

$$\lim_{i \to \infty} \alpha(i) = \alpha \qquad (3)$$

where $\alpha(i)=P(t_A \leq i| H_0)$ and H_0 is defined in (2). The $\alpha(i)$ is therefore equal to $P(t_A \leq i|\tau > i)$. The alarm limit is constructed so that we have a sequence of alarm sets, resulting in $\alpha < 1$ and hence it is 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, summarizes the information,

$$ARL^0 = E[t_A \mid \tau = \infty].$$

A similar measure is the median run length conditional of no change, $MRL^0 = \text{Median}[t_A \mid \tau = \infty]$. Another summarizing measure of the false alarm distribution is the probability of a false alarm (PFA),

$$PFA = P(t_A < \tau) = \sum_{t=1}^{\infty} P(\tau = t) \cdot P(t_A < \tau \mid \tau = t). \tag{4}$$

As mentioned above there are several measures which reflect the timeliness of a motivated alarm. In some applications, such as medical intensive care (Petzold et al. (2004) and Frisén (1992)) and turning point detection in business cycles (Andersson et al. (2004)), 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

$$PSD(t, d) = P\left(t_A - t < d \mid t_A \geq \tau, \tau = t\right) \tag{5}$$

where $d \geq 1$. Another aspect of the timeliness can be reflected by the delay of a motivated alarm, here presented as the conditional expected delay

$$CED(t) = E[t_A - t \mid t_A \geq \tau, \tau = t]. \tag{6}$$

An evaluation measure that is often used is the average run length

$$ARL^1 = E[t_A \mid \tau = 1],$$

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 an industrial manufacturing process where one considers various start-up problems. There are however some drawbacks with this optimality criterion. First, this criterion only considers changes that occur at the start of the monitoring, which is not realistic in e.g. economics. Second, the average run length provides us with limited information about the behaviour of methods, especially since the run length distributions are often skewed (Frisén (1992)). A third reason is that degenerated methods which would never be used in practice satisfies this criterion (Frisén (2003)).

Apart from the delay of an alarm, 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

$$PV(i) = P(C(i) \mid t_A = i)$$

has been suggested as a criterion of evaluation by Frisén (1992).

Optimality criterions that reflect timeliness is treated by Girshick and Rubin (1952) and Shiryaev (1963), where the utility is specified in the following way: 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

$$u(t_A, \tau) = \begin{cases} h(t_A - \tau) & , t_A < \tau \\ a_1 \cdot (t_A - \tau) + a_2, & t_A \geq \tau \end{cases} \tag{7}$$

where the function $h(t_A\text{-}\tau)$ is an arbitrary function and a_1 is typically negative. In a situation where the intensity of a change, v_t, is constant, the full likelihood ratio method (LR, described in section 4.1) maximizes the expected value of the utility (see Frisén and de Maré (1991)). If the function $h(t_A\text{-}\tau)$ is specified as a constant b, the expected utility is

$$E[u(t_A, \tau)] = b \cdot PFA + a_1 \cdot ED + a_2,$$

where ED is the expected delay, defined as

$$ED = \sum_{t=1}^{\infty} P(\tau = t) \cdot ED(t),$$

where $ED(t) = E[\max(0, t_A\text{-}t)|\tau=t] = CED(t) \cdot P(t_A \geq \tau)$. When PFA is fixed, the expected utility is maximized for a minimal ED (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 D and C is based on the likelihood ratio (LR) between C(s) and D(s), and an alarm is given when

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

where $L(s, t) = f_{Xs}(x_s|\tau=t)/f_{Xs}(x_s|D)$ 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 L(s, s) 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

$$X(s) - \mu_0 > g, \tag{8}$$

where g is a constant and μ_0 is the expected value of the X given the process is in-control.

When $C(s)=\{\tau=s\text{-}p+1\}$ 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

$$\sum_{i=s-p+1}^{s} (x(i)-\mu_0) > g, \tag{9}$$

where p is the window width and g is a constant. The methods in (8) and (9) 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 evaluation in section 5, a window width of two observations (p=2) is used for MASur.

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 (8) and (9), 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 (regime). The cumulative probability of a false alarm no later than at time i from the start, $P(t_A \leq i | \tau > i)$, is

$$1-P\left(\bigcap_{j=1}^{i} X(j) < g\right) = 1 - \Phi(g)^i \tag{10}$$

and

$$1-P\left[\bigcap_{j=1}^{i}\left(\sum_{t=j-p+1}^{j} X(t) < g\right)\right], i = 1, 2, \ldots,$$

for the ShewSur and MASur, respectively, where $\Phi(\cdot)$ is the standard Normal probability distribution function. Both the probability expressions above tend to 1 as $i \to \infty$, which means that 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, nor should the limit grow at too slow a rate.

Leisch et al. (2000) suggested the following alarm limit for decision time s

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

where $e \approx 2.718$ and $c > \sqrt{2}$ is a constant to be determined. The methods giving an alarm as soon as

$$\sum_{i=s-p+1}^{s} (x(i)-\mu_0) > g(s)$$

are for p=1 (only the last observation) and p≥2 hereafter referred to as ShewTest and MATest, respectively. In the evaluation in section 5, a window width of two observations (p=2) is used for MATest.

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

Proof: According to theorem 4.1 in Frisén and de Maré (1991), it holds that $\alpha<1$ if and only if $P(t_A=s|\tau>s,\ t_A\geq s)<1$ for all s and $\sum_{s=1}^{\infty} P(t_A = s \mid t_A \geq s,\ \tau > s) < \infty$.
We have that $P(t_A=s|\tau>s,\ t_A\geq s)=1-\Phi(g(s))<1$ since $\Phi(g(s))>0$ for all s.

$$\sum_{s=1}^{\infty} P(t_A = s \mid t_A \geq s,\ \tau > s) = \sum_{s=1}^{\infty}(1-\Phi(g(s))) =$$

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

$$\sum_{s=1}^{\infty}(2\cdot\pi)^{-1/2}\cdot\exp\{-g^2(s)/2\}/g(s) =$$

$(2/\pi)^{1/2}\cdot c^{-1}\cdot e^{-c^2/2} + \sum_{s=3}^{\infty}(2\cdot\pi)^{-1/2}\cdot\left(c\cdot\sqrt{\ln s}\right)^{-1}\cdot s^{-c^2/2}$. 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 $c\neq 0$, $s>1$ and $\sum_{s=1}^{\infty} s^{-c^2/2}$ is convergent for $c>\sqrt{2}$. Therefore $\alpha<1$ for $c>\sqrt{2}$.

Leisch (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 asymptotic size α and a fixed ARL^0. We will demonstrate the consequences of these two approaches in terms of the timeliness of alarms. The in-control 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, where the technique of using a utility function to determine which method to choose is illustrated by a case of trading a stock index.

Chu et al. (1996) argues, having applications to economic time series in mind, that sampling under the null hypothesis is costless, whereas resetting the monitoring system after a false alarm does create a large cost. Thus false alarms are severe and from this point of view we should set the asymptotic size to a small value, e.g. 10 % (α=0.10), since the cost of a false alarm is high.

In a situation where the cost of a false alarm is low, we can instead set the asymptotic size to a large value, e.g. 90 %.

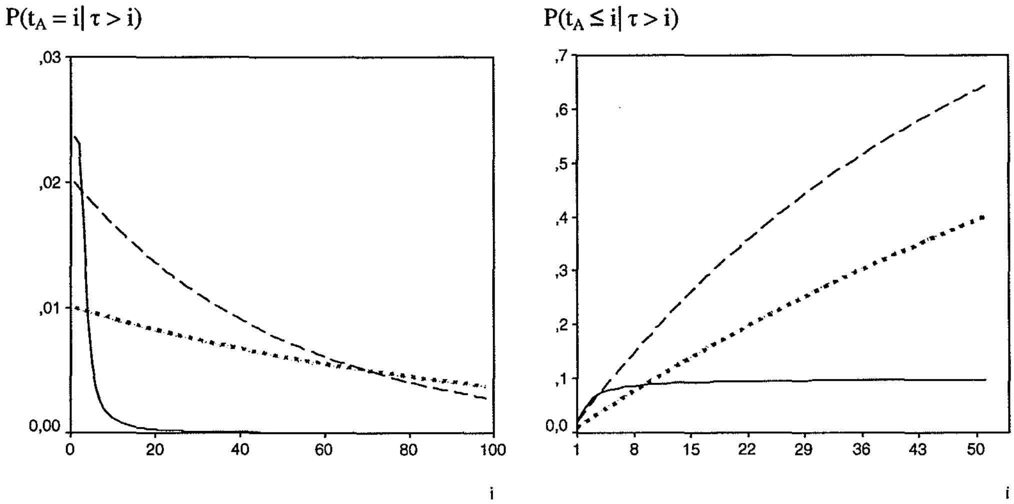
The respective alarm limits of ShewTest and MATest are adjusted to give α = {0.10, 0.90}, and ShewSur is adjusted to give ARL^0 = {50, 100, 250}. The limit of MASur is adjusted to give ARL^0 ={50, 100}. For MASur and MATest, a window width of two observations (p=2) is considered and simulations determine the alarm limits. 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

We assess the in-control properties of the methods by the run length distributions of the false alarms and associated summarizing measures. In Fig 1 below, the false alarm probability and the cumulative false alarm probability for ShewSur and the ShewTest are shown. For ShewSur and ShewTest, the cumulative false alarm probability $P(t_A \leq i | \tau > i)$ is given by (10) and

$$1-\prod_{j=1}^{i}\Phi(g(j)),$$

respectively.



a b

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

Fig 1, panel a, shows that the probability of a false alarm for the ShewTest becomes small very fast, so that almost all alarms are located at early time points. For ShewSur the alarm probabilities decrease more slowly. Panel b shows that ShewTest(0.10) quickly reach the size level 0.10.

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). A result of the skewness is that the detection power is highly concentrated to early time points. The tendency to give early alarms for the test approach is an important difference to the surveillance approach. Consequences of the allocation of the false alarms on the ability to detect changes will be considered in the next section.

The PFA in (4) summarizes the false alarm distribution in Fig 1, panel b by weights with the distribution of τ. When the Geometric distribution with intensity ν is used, the PFA for ShewSur equals

$$1-\nu/\left(1-(1-\nu)\cdot\Phi(g)\right).$$

PFA is illustrated as a function of the intensity ν in Fig 2 below.

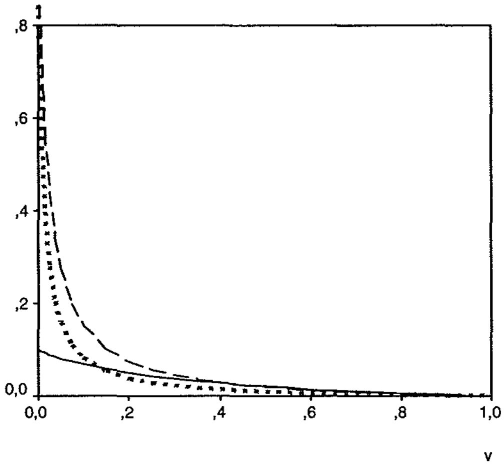
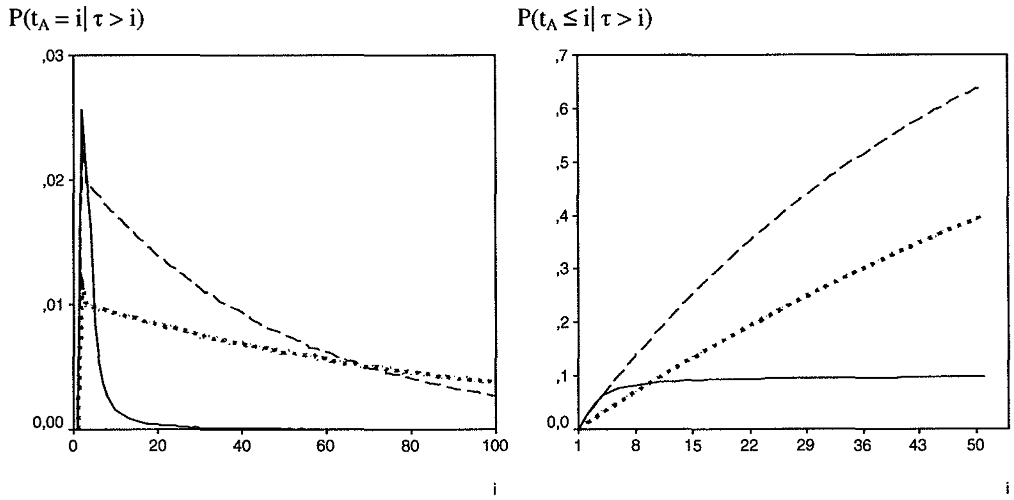


Fig 2. The probability of a false alarm, as a function of the intensity ν. ShewTest(0.10) (——), ShewSur(50) (– –), ShewSur(100) (■ ■ ■).

The difference in level between the two surveillance methods (ShewSur) is due to the difference in the value of ARL^0. When ν tends to 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 which means that the probability of an early change is large. Therefore 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 the intensity ν tends to zero, the density of τ tends to a uniform distribution, which means that the regime shift is equally likely to occur early as very late. Most of the alarms are therefore false. When ν tends to zero, PFA for the test methods tends to the fixed size (e.g. 0.10), whereas PFA for the surveillance methods tends to 1, as is seen in Figure 2.

The above investigation was for the situation when the alarm statistic was based on only the last observation, the Shewhart approach. We also briefely investigate the moving average with p=2. 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). In Fig 3 below the false alarm probability and the cumulative false probability for MASur and the MATest are shown.



a b

Fig 3. Panel a: False alarm density. Panel b: False alarm distribution. MATest(0.10) (——), MASur(50) (– –), MASur(100) (▪ ▪ ▪).

Apart from that the alarm probability by construction is zero at the first time point an d very high at the first decision time 2, the shapes of the curves in Fig 3 are similar to those of the Shewhart approaches (Fig 1).

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 $\alpha<1$ also have a low probability of a late false alarm (a false alarm long after the monitoring has started):

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

$$\Rightarrow \lim_{j\to\infty} P(t_A = j \mid t_A \geq j) = 0$$

since

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

This explains the shapes of the false alarm probability of the test approaches in Fig 1 and 3 in the previous section (the probabilites tend to zero).

That the false alarm probability is low might at a first glance seem like a good property. But the probability of a false alarm at time j gives an indication of the overall alarm probability at that time. If the false alarm probability, $P(t_A = j \mid t_A \geq j)$, tends to zero, the 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 a false alarm probability that tends to zero, implies that the alarm limit tends to infinity as $j \to \infty$. Therefore also $\lim_{j\to\infty} P(t_A = j \mid t_A \geq j, \tau = j) = 0$. Consequences of this will be illustrated below.

The case where the change occurs at the same time as the surveillance was started (τ=1) is the most widely considered case for evaluation in literature. The run length density when the change occurred immediately, $P(t_A=i \mid \tau=1)$, is shown in Fig 4 for μ_1=1. For ShewSur it is calculated as

$$\left(1-\Phi(g - \mu_1)\right)\cdot\Phi(g - \mu_1)^{i-1}$$

and for the ShewTest as

$$\left(1-\Phi(g(i) - \mu_1)\right)\cdot\prod_{j=1}^{i-1}\Phi(g(j) - \mu_1).$$

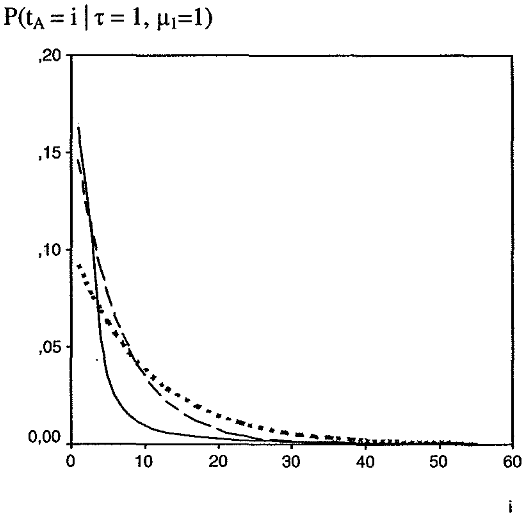


Fig 4. The density of the time of an alarm, when the change occurred immediately and μ_1=1. ShewTest(0.10) (——), ShewSur(50) (– –), ShewSur(100) (■ ■ ■).

The probability of an alarm at the very first time points is highest for the ShewTest.

ARL^1 is the average run length, given a change at the start of the monitoring. This corresponds to τ = 1 and τ = 2 for the Shewhart and MA approach with p=2, respectively. ARL^1, for μ_1=3, is presented in Table 1.

Table 1. Values of ARL^1 when μ_1=3

ARL^0	ShewSur	MASur
50	1.208	2.454
100	1.334	2.586
250	1.5722	-
α	ShewTest	MATest
0.10	1.185	2.443
0.90	1.049	2.212

For ShewSur, $ARL^1 = 1/(1-\Phi(g-\mu))$. 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 emphasized when α = 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 (5), tends to zero as the time of the change tends to infinity. We have that

$$PSD(t, d) = \sum_{j=0}^{d-1} P(t_A = t+j \mid t_A \geq t, \tau = t)$$

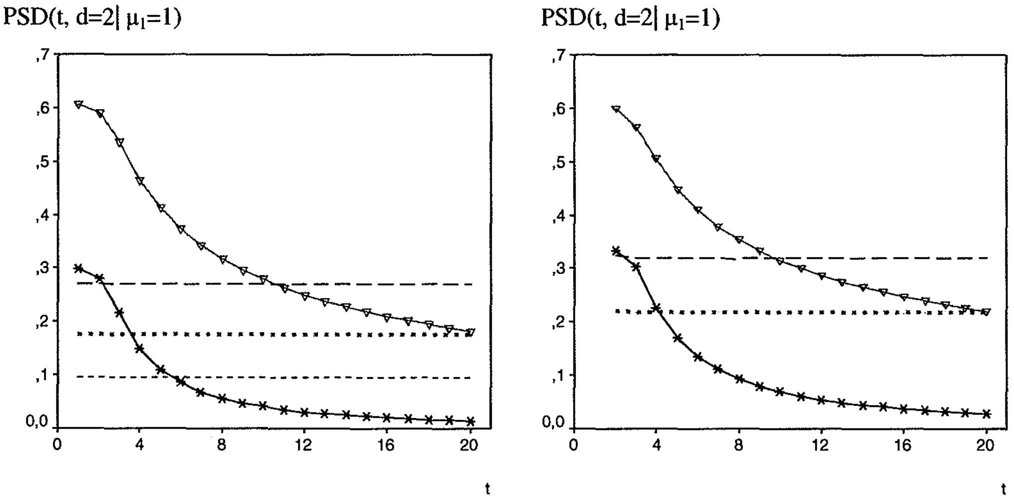
and for the test methods $\lim_{t\to\infty} P(t_A = t+j \mid t_A \geq t, \tau = t) = 0$, j = {0, 1, ...}. Therefore $\lim_{t\to\infty} PSD(t, d) = 0$ for any $d \geq 1$. For the ShewSur and the ShewTest the PSD(t, d) is equals to

$$1 - \Phi(g - \mu_1)^d$$

and

$$1 - \prod_{j=0}^{d-1} \Phi(g(t+j) - \mu_1)$$

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) functions are shown in Fig 5, panel a and b, respectively, when d=2 and μ_1=1.



a b

Fig 5. The probability of successful detection PSD(t, d=2) for different values of the time of the change when μ_1=1. Panel a: ShewTest(0.10) (—— *) ShewTest(0.90) (——∇), ShewSur(50) (– –), ShewSur(100) (∗ ∗ ∗), ShewSur(250) (----).Panel b: MATest(0.10) (—— *), MATest(0.90) (——∇), MASur(50) (– –), MASur(100) (∗ ∗ ∗).

We showed above that PSD tends to zero for the test methods, which means that if the regime shift occurs late, these methods have very little chance of detecting it. This drawback can not be overcome by changing α. As seen in Fig 5, the behavior is the same for $\alpha = 0.10$ and $\alpha = 0.90$ 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) and Bock et al. (2004) from simulation experiments. However, it was not recognized as a direct consequence of the way the false alarms are controlled but as a consequence of the way the alarm limit changed with time.

Finding alarm limits that increase the detection power at later time points has been discussed by Leisch et al. (2000) and Zeileis et al. (2004). Among other things, alarm limits that depend on a specified prior distribution for τ has been briefly suggested. This means that the intensity of a regime shift is considered, i.e. how often we can expect a regime shift. In the modeling of economic series, there are some results regarding different approaches for the intensity. The intensity can be allowed to change with the time spent in the state, which is referred to as duration dependence, see Durland and McCurdy (1994) and Zuehlke (2003). Another approach is to let the intensity depend on explanatory variables, see e.g. Filardo (1994) and Filardo and Gordon (1998). In a surveillance context, Andersson (2004) used the official turning points times in the Swedish business cycle to construct and use an empirical distribution of τ in several surveillance systems. The disadvantage of using an empirical prior distribution of τ is that the ability to detect a change that takes place at an unexpected time point is poor. In many surveillance approaches to economic and financial turning point detection, τ has a geometric distribution, i.e. the intensity is constant as in Hamilton (1989), Koskinen and Öller (2003), Marsh (2000), Dewachter (2001). Though the alarm probability can be increased for certain time points in this way, the ability to detect late changes will still be low.

Chu et al. (1996) motivated using a fixed asymptotic size in terms of the cost of false alarms, but the cost of the delay of motivated alarms was not taken into account. Here the delay is summarized by the conditional expected delay (6). For ShewSur

$$CED(t) = ARL^1\text{-}1$$

for all t and for ShewTest,

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

The CED functions are shown in Fig 6 and 7 for μ_1=3.

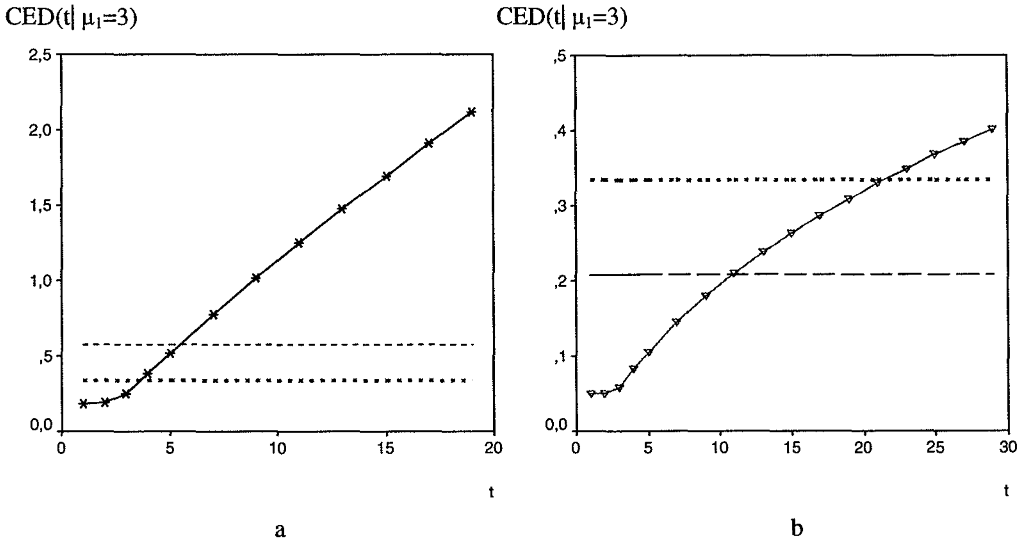


Fig 6. The conditional expected delay CED(t) for different values of t (the time of the change) when μ_1=3. Panel a: ShewTest(0.10) (—— *), ShewSur(100) (▪ ▪ ▪), ShewSur(250) (- - - -). Panel b: ShewTest(0.90) (——▽), ShewSur(50) (— —), ShewSur(100) (▪ ▪ ▪).

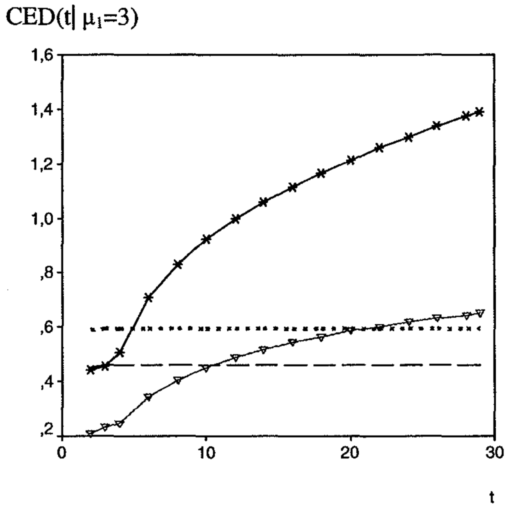


Fig 7. The conditional expected delay CED(t) for different values of the time of the change when μ_1=3. MATest(0.10)(—— *), MATest(0.90) (——▽), MASur(50) (— —), MASur(100) (▪ ▪ ▪).

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

$$\sum_{d=1}^{\infty} \left(1\text{-PSD}(t,\, d)\right).$$

For the ShewTest we have that PSD(t, d) ≥ PSD(t+1, d) for all t and d, and then it follows that CED(t) ≤ CED(t-1), i.e. CED(t) is increasing with t. Since $\lim_{t\to\infty} \mathrm{PSD}(t, d) = 0$ 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. This drawback can not be helped by choosing a large asymptotic size and 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

An important thing to consider in the monitoring is what to do if an alarm is given. The problem of what kind of action to take can be seen in the light of the trust that you have in an alarm and the utility of an action given an alarm. The predictive value at time i, PV(i), reflects the trust of an alarm at that time and can be expressed as

$$\mathrm{PV(i)} = \frac{\mathrm{PMA(i)}}{\mathrm{PMA(i)+PFA(i)}}$$

where $\mathrm{PFA(i)} = P(t_A = i \mid i < \tau)\cdot P(\tau > i)$ is the probability of a false alarm at time i and and $\mathrm{PMA(i)} = \sum_{j=1}^{i} P(\tau = j)\cdot P(t_A = i \mid \tau = j)$ is the probability of a motivated alarm at time i. For ShewSur and the ShewTest, the PFA(i) is equal to $(1\text{-}\nu)^i\cdot(1\text{-}\Phi(g))\cdot\Phi(g)^{i-1}$ and $(1\text{-}\nu)^i\cdot(1\text{-}\Phi(g(i)))\cdot\prod_{j=1}^{i-1}\Phi(g(j))$, respectively, whereas PMA(i) is equal to

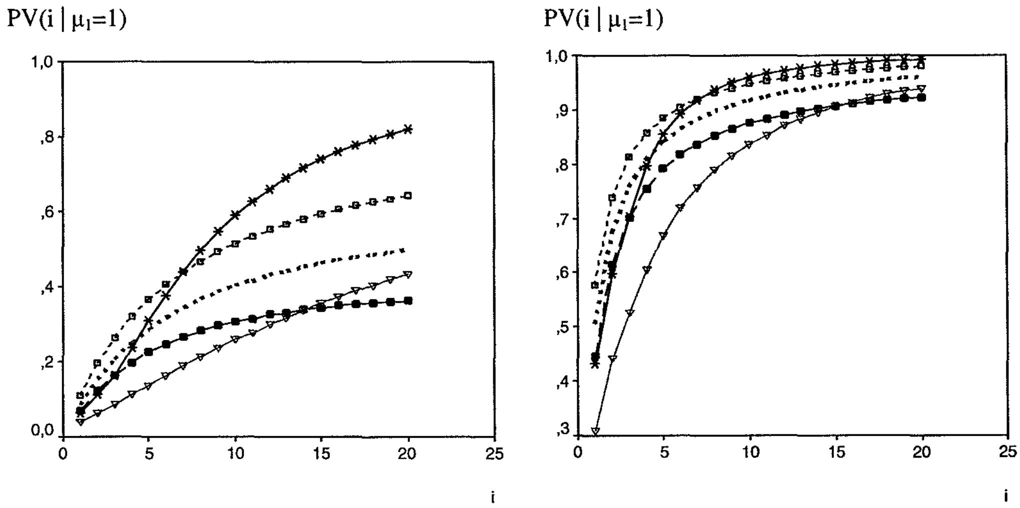
$$(1\text{-}\Phi(g\text{-}\mu_1))\cdot\sum_{j=1}^{i}\nu\cdot(1\text{-}\nu)^{j-1}\cdot\Phi(g)^{j-1}\cdot\Phi(g\text{-}\mu_1)^{i-j} =$$

$$(1\text{-}\Phi(g\text{-}\mu_1))\cdot\nu\cdot\left\{\frac{\Phi(g\text{-}\mu_1)^i - ((1\text{-}\nu)\cdot\Phi(g))^i}{\Phi(g\text{-}\mu_1)\text{-}(1\text{-}\nu)\cdot\Phi(g)}\right\}$$

and

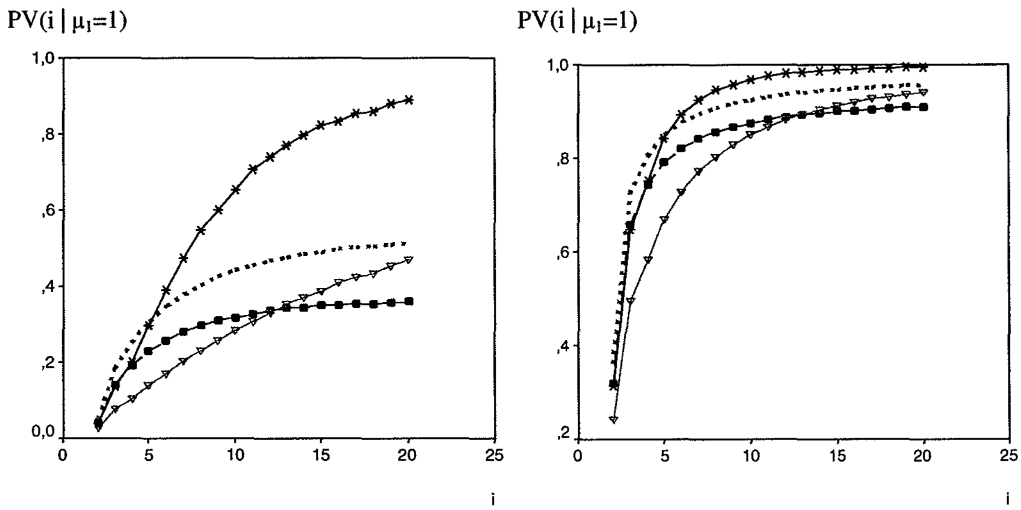
$$\sum_{j=1}^{i}\nu\cdot(1\text{-}\nu)^{j-1}\cdot\prod_{t=1}^{j-1}\Phi(g(t))\cdot\prod_{t=j}^{i-1}\Phi(g(t)\text{-}\mu_1)\cdot(1\text{-}\Phi(g(i)\text{-}\mu_1))$$

for these two methods. For ShewSur, the PV(i) has the asymptote $\nu/(\nu+(1\text{-}\Phi(g))\cdot c)$ when $i\to\infty$, where $c = ((1\text{-}\nu)\cdot\Phi(g)\text{-}1)/(\Phi(g)\cdot(1\text{-}\Phi(g\text{-}\mu_1)) + 1/\Phi(g)$. In Fig 8 and 9 the predictive value is shown as a function of the time of the alarm.



a b

Fig 8. The predictive value as a function of i when μ_1=1. Panel a: ν =0.01, Panel b: ν =0.1. ShewTest(0.10) (—— *) ShewTest(0.90) (——▽), ShewSur(50) (– – ●), ShewSur(100) (▪ ▪ ▪), ShewSur(250) (----□).



a b

Fig 9. The predictive value as a function of i when μ_1=1. Panel a: ν =0.01, Panel b: ν =0.1. MATest(0.10) (—— *) MATest(0.90) (——▽), MASur(50) (– – ●), MASur(100) (▪ ▪ ▪).

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 method that is optimal, the LR method, the predictive value was found by Frisén and Wessman (1998) to be relatively constant.

5.4 *Utility*

Timeliness can be measured indirectly, as the amount gained by an action after an alarm is given at the "right" time. In the specification of utility (7) in section 3 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,

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

In a situation where the intensity of a change is constant, the LR method maximizes the expected value of the utility, $E[u(t_A, \tau)]$. 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 that we 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 trading of the Hang Seng Index (HSI). For a timely trading of assets in the financial markets, for example shares of a stock index fund or currency, it is important to determine rules which give maximal utility. Bock et al. (2003) and Lam and Yam (1997) considered trading closing HSI using different surveillance systems. HSI is a marked-value weighted index of the stock prices of the 33 largest companies on the Hong Kong stock market. The weight each stock is assigned in the index is related to the price of the stock. HSI can thus be seen as the price of a portfolio of stock.

The aim was to timely detect turning points and buy or sell units of HSI as soon as an alarm was given that a trough or and peak had occurred, respectively. An assumption made in Lam and Yam (1997) and Bock et al. (2003) 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 ($\mu_0 \leq 0$, $\mu_1 > 0$) 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

Different specifications of the utility function (7) are possible. The $E[u(t_A, \tau)]$ depends on the in- and out-of-control properties, characterized by the false alarm behavior and the delay properties of motivated alarms. Depending on which function is chosen for $h(t_A-\tau)$, the $E[u(t_A, \tau)]$ will be influenced by the false alarms in different ways, e.g. solely through the PFA (4) when $h(t_A-\tau)$is a constant b.

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

$$u(t_A, \tau) = E[r(t_A) \mid t_A, \tau]. \tag{11}$$

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. (2003) where for a cyclical stock price process in which the expected return increases as t_A approaches the time of the change τ a piecewise linear utility function was suggested. If the X function in the return expression above can be modelled by a piecewise linear trend, then (11) can be written as

$$u(t_A, \tau) = c + \begin{cases} \mu_0 \cdot t_A, & t_A < \tau \\ \mu_0 \cdot (\tau\text{-}1) + \mu_1 \cdot (t_A - \tau + 1), & t_A \geq \tau \end{cases}. \tag{12}$$

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, (12)is a reasonable specification of the utility function. The expected value of (12) depends on the behavior of both false and motivated alarms and given τ=t,

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

For false alarms ($t_A < \tau$),

$$EFA(t)=E[t_A|\, t_A < \tau, \tau=t] \cdot P(t_A < \tau)=\sum_{i=1}^{t-1} i \cdot P(t_A = i \mid t_A < \tau, \tau = t) \cdot P(t_A < \tau)$$

is the expected time of a false alarm given τ=t, which is summarized with respect to the distribution of τ by EFA = E[EFA(t)]. For motivated alarms ($t_A \geq \tau$), recall from section 3 that $CED(t)=E[t_A-t|t_A \geq \tau, \tau = t]= E[t_A-t+1|t_A \geq \tau, \tau = t]-1$ and

$$ED(t) = E[\max(0, t_A-t)|\tau =t]=CED(t) \cdot P(t_A \geq \tau).$$

Since $E[t_A-t+1|t_A \geq \tau, \tau=t] \cdot P(t_A \geq \tau)$ equals $ED(t)+P(t_A \geq \tau)$, summarizing $E[t_A-t+1|t_A \geq \tau, \tau=t] \cdot P(t_A \geq \tau)$ with respect to the distribution of τ yields $ED+E[P(t_A \geq \tau)]$. When we summarize the whole utility with respect to the distribution of τ, we get

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

When τ follows a geometric distribution with intensity ν such that $P(\tau=t)=\nu \cdot (1-\nu)^{t-1}$, exact expressions can be found for ShewSur: $E[\tau \cdot P(t_A \geq \tau)]=\nu/\{1-(1-\nu) \cdot \Phi(g)\}$, $ED=(\nu \cdot (ARL^1-1))/\{1-(1-\nu) \cdot \Phi(g)\}$, $EFA(t)=\{1-\Phi(g)^{t-1} \cdot (\Phi(g)+t-t \cdot \Phi(g))\}/(1-\Phi(g))$ and $EFA=\sum_{t=1}^{\infty} P(\tau = t) \cdot EFA(t)$, i.e.

$$\text{EFA}=\frac{\nu}{(1-\Phi(g))}\cdot\left\{1-\frac{\nu\cdot\Phi(g)}{1-(1-\nu)\cdot\Phi(g)}-\frac{\nu}{(1-(1-\nu)\cdot\Phi(g))^2}+\frac{\nu\cdot\Phi(g)}{(1-(1-\nu)\cdot\Phi(g))^2}\right\}.$$

For ShewTest the values of $E[P(t_A \geq \tau)]$, $E[\tau \cdot P(t_A \geq \tau)]$ and EFA can be numerically approximated. ED for ShewTest is

$$\text{ED}=\sum_{t=1}^{\infty} P(\tau=t)\cdot\sum_{i=t}^{\infty}(i-t)\cdot P(t_A=i \mid t_A \geq \tau, \tau=t)\cdot P(t_A \geq \tau).$$

A lower boundary for ED is

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

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) motivated using a monitoring system with a fixed asymptotic size in terms of the cost of false alarms but 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 (7). If $a_1=0$, then the maximization of the expected utility $E[u(t_A, \tau)]$ 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. The cost often depends on what price and quantity the asset is being traded at and can therefore be non-constant. If there is no transaction cost then $c = 0$ in (12). We will use that value in the utility illustration below.

The period February 10 to May 28, 1999 for HSI (analyzed by Bock et al. (2003)) 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 μ is set to the average of the absolute values of the two slopes, which yields $\mu= 0.0069$. Then

$$E[u(t_A, \tau)]=\mu\cdot\{\text{EFA}-\text{ED}+E[\tau\cdot P(t_A \geq \tau)]-2\cdot E[P(t_A \geq \tau)]\},$$

which is maximized for a minimal $E[|t_A-(\tau-1)|]$. That is because

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

which is equals to $-\{E[u(t_A, \tau)]-\max_{tA}\{E[u(t_A, \tau)]\}\}/\mu$ where $\max_{tA}\{E[u(t_A, \tau)]\}= E[(\tau-1)]$. Since for a given ν, $E[(\tau -1)]$ 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

The parameters of the X process influence the relation between the utilities for the two approaches (test and sur). 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 (μ_1), 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 and therefore the differences in false alarm properties will not be important. Thus for small shifts the surveillance approach will be preferred since the delay is shorter, except possibly for very large values of ν because then the delay for the test approach is small.

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 including a peak in the logarithm of the price, the corresponding standardized (μ_0=0 and σ^2=1) downward shift (negative μ_1) in the differences had an estimated size of 0.82, by Bock et al. (2003).

For a shift of such size, the level of the CED curve for the Shewhart approaches will be substantially higher than in Fig 6 where μ_1=3, so it is reasonable to say that much concentration is on the delay and the false alarm properties are not important. Then the surveillance approach will be preferred except possibly for very large values of ν. For HSI in the above mentioned period, ν was estimated to 0.018, which is not very large.

As an illustration of the technique of expected utility, we calculate it for the estimated parameters of the period of HSI. With the costs and parameters discussed above we have $E[u(t_A, \tau)]$ equal to 0.149 and 0.174 for ShewSur with ARL^0 equal to 50 and 100, respectively. The utility in (11) 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%. When $E[u(t_A, \tau)]$ is calculated for ShewTest(0.10), the delay is very long. This results in a highly negative value for the utility (the value is less than -0.442), even though the figure is conservative because of the truncation when calculating ED (see section 5.4.2). 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. A process (e.g. price of stocks) is monitored and when there is enough evidence that a regime shift has occurred an alarm is called. Sometimes this alarm can be false. The two 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). The approaches are denoted test methods and surveillance methods respectively.

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. Chu et al. (1996) argue in favor of this when sampling under the null hypothesis is costless but resetting the monitoring system after a false alarm does create a large cost.

One argument against monitoring methods controlled by a fixed size is that ordinary statements for hypothesis testing do not consider the timeliness of alarms. For example, the power of a test does not give any information about the time of the alarm. The use of a fixed size gives the result that the probability of making an alarm is very low, if a change occurs a long time after the monitoring has started. Hence the ability to timely detect a change that occurs late will be low. Consequences of this were illustrated here by the delay of a motivated alarm and the probability of detection within two time units. The expected delay was shown to have a limiting value of infinity, i.e. the delay gets longer when the regime shift occurs later. The probability of successful detection has a limiting value of zero, i.e. no detection ability for late occurring regime shifts. These drawbacks can not be adjusted by choosing a large asymptotic size (e.g. 0.90). The limited ability to detect changes that occur late 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. A consequence of the many early alarms is that the predictive values are lower for test methods, compared to surveillance methods, at early time points. The predictive value of the test methods is higher at late alarms. Therefore, the alarms given early by the test methods are less reliable compared to those of the surveillance methods. The better predicted value of late alarms has no practical importance since they are rare and tend to be given with great delay.

In order to compare different methods, a utility function can be used. In a situation of on-line detection, the utility often consists of two parts: one concerning the false alarms and the other concerning the delay of motivated alarms. In terms of the utility of a false or motivated alarm, methods controlled by a fixed size consider that 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.

The smaller the size of the shift, the larger the delay and the larger the impact of the delay on the utility as compared to the impact of false alarms. 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. 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. 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

Andersson, E. (2004) The impact of intensity in surveillance of cyclical processes. *Communications in Statistics. Simulations and Computations,* **to appear**.

Andersson, E., Bock, D. and Frisén, M. (2004) Detection of turning points in business cycles. *Journal of business cycle measurement and analysis,* **1,** 93-108.

Bock, D., Andersson, E. and Frisén, M. (2003) The relation between statistical surveillance and certain decision rules in finance. *Submitted.*

Bock, D., van Dijk, D. and Franses, P. H. (2004) Monitoring macroeconomic volatility. Research Report, Statistical Research Unit, Göteborg University, Sweden, 2004:1

Carsoule, F. and Franses, P. H. (1999) Monitoring structural change in variance, with an application to European nominal exchange rate volatility. Research Report, Econometric Institute, Erasmus University Rotterdam, Rotterdam EI-9925/A

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

Chu, C.-S. J., Stinchcombe, M. and White, H. (1996) Monitoring structural change. *Econometrica,* **64,** 1045-1065.

Dewachter, H. (2001) Can Markov switching models replicate chartist profits in the foreign exchange market? *Journal of International Money and Finance,* **20,** 25-41.

Durland, J. M. and McCurdy, T. H. (1994) Duration-dependent transitions in a Markov model of U.S. GNP growth. *Journal of Business and Economic Statistics,* **12,** 279-287.

Filardo, A. J. (1994) Business-Cycle Phases and Their Transitional Dynamics. *Journal of Business & Economic Statistics,* **12,** 299-308.

Filardo, A. J. and Gordon, S. F. (1998) Business cycle durations. *Journal of Econometrics,* **85,** 99-123.

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

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

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

Frisén, M. and de Maré, J. (1991) Optimal Surveillance. *Biometrika,* **78,** 271-80.

Frisén, M. and Wessman, P. (1998) Quality improvement by likelihood ratio methods for surveillance. In *Quality Improvement Through Statistical Methods*(Ed, Abraham, B.) Birkhauser, Boston.

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

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

Hamilton, J. D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica,* **57,** 357-384.

Koskinen, L. and Öller, L.-E. (2003) A Classifying Procedure for Signalling Turning Points. *Journal of Forecasting,* **23,** 197-214.

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

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

Leisch, F., Hornik, K. and Kuan, C.-M. (2000) Monitoring structural changes with the generalized fluctation test. *Econometric Theory,* **16,** 835-854.

Marsh, I. W. (2000) High-frequency Markov Switching Models in the Foreign Exchange Market. *Journal of Forecasting,* **19,** 123-134.

Petzold, M., Sonesson, C., Bergman, E. and Kieler, H. (2004) Surveillance of longitudinal models. Detection of intrauterine growth retardation. *Biometrics,* **to appear**.

Pollak, M. and Siegmund, D. (1975) Approximations to the Expected Sample Size of Certain Sequential Tests. *Annals of Statistics,* **3,** 1267-1282.

Royston, P. (1991) Identifying the fertile phase of the human menstrual cycle. *Statistics in Medicine,* **9.**

Ryan, T. P. (2000) *Statistical methods for quality improvement,* John Wiley & Sons, New York.

Schipper, S. and Schmid, W. (2001a) Control charts for GARCH processes. *Nonlinear Analysis,* **47,** 2049-2060.

Schipper, S. and Schmid, W. (2001b) Sequential Methods for Detecting Changes in the Variance of Economic Time Series. *Sequential Analysis,* **20,** 235-262.

Severin, T. and Schmid, W. (1998) Statistical process control and its application in finance. In *Risk measurement, econometrics and neural networks*(Eds, Bol, G., Nakhaeizadeh, G. and Vollmer, C.-H.) Physica Verlag, Heidelberg, pp. 83-104.

Severin, T. and Schmid, W. (1999) Monitoring changes in GARCH processes. *Allgemeines Statistisches Archiv,* **83,** 281-307.

Shewhart, W. A. (1931) *Economic Control of Quality of Manufactured Product,* MacMillan and Co., London.

Shiryaev, A. N. (1963) On optimum methods in quickest detection problems. *Theory of Probability and its Applications.,* **8,** 22-46.

Smith, A. F. and West, M. (1983) Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter. *Biometrics,* **39,** 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,** 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,** 645-670.

Wetherhill, G. B. and Brown, D. W. (1991) *Statistical Process Control: Theory and Practice,* Chapman and Hill, London.

Zeileis, A., Leisch, F., Kleiber, C. and Hornik, K. (2004) Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics,* **to appear**.

Zuehlke, T. W. (2003) Business cycle duration dependence reconsidered. *Journal of Business and Economic Statistics,* **21,** 564-569.

Research Report

2003:1	Holgersson, T. & Shukur, G.:	Testing for multivariate heteroscedasticity.
2003:2	Holgersson, T.:	Testing for multivariate autocorrelation.
2003:3	Petzold, M. & Sonesson, C.:	Detection of intrauterine growth restriction.
2003:4	Bock, D.:	Similarities and differences between statistical surveillance and certain decision rules in finance.
2003:5	Holgersson, T. & Lindström, F.:	A comparison of conditioned versus unconditioned forecasts of the VAR(1) process.
2003:6	Bock, D.:	Early warnings for turns in business cycles and finance.
2003:7	Lindström, F.:	On prediction accuracy of the first order vector auto regressive process.
2003:8	Petzold, M. & Jonsson, R.:	Maximum Likelihood Ratio based small-sample tests for random coefficients in linear regression.
2003:9	Petzold, M.:	Preliminary testing in a class of simple non-linear mixed models to improve estimation accuracy.
2003:10	Frisén, M. & Gottlow, M.:	Graphical evaluation of statistical surveillance.
2003:11	Frisén, M.:	Statistical measures for evaluation of methods for syndromic surveillance.
2003:12	Andersson, E.:	A monitoring system for detecting starts and declines of influenza epidemics.
2004:1	Bock, D., van Dijk, D. & Franses, P.H.:	Monitoring macroeconomic volatility.