GÖTEBORG UNIVERSITY

Department of Statistics

RESEARCH REPORT 1994:6
ISSN 0349-8034

CONSTANT PREDICTIVE VALUE OF AN ALARM

by

Göran Åkermo

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S-411 25 Göteborg
Sweden

CONSTANT PREDICTIVE VALUE OF AN ALARM

By GÖRAN ÅKERMO
Department of Statistics,
Göteborg University, S-41125 Göteborg, Sweden

SUMMARY

One main purpose of statistical surveillance is to detect a change in a process, often expressed as a shift from one level to another. When a sequence of decisions is made, measures, like the number of decisions that have to be taken before an alarm are of interest. In many situations a shift might occur any time after the surveillance was initiated.

Prior knowledge of the probability of a change, the incidence, can become crucial when a method is selected and the parameter values of the method are set. The predictive value of an alarm is a measure of performance that takes this information into consideration and is an important tool for evaluating methods.

Mostly an alarm is useful only if its predictive value is large. The predictive value of an alarm is the probability that a change has occurred given an alarm. In this paper it is demonstrated that the incidence in the first point has to be relatively high, or the alarm limits very wide, in order to achieve a predictive value greater than, say 0.5.

The interpretation of an alarm is difficult to make if the predictive value of an alarm varies with time. For the ordinary Shewhart method and a selection of Moving Average Methods it is demonstrated how the predictive value increases with time if the incidence is constant. The incidence which would give the methods a constant predictive value are determined. The methods are thus demonstrated to give easily interpreted alarms only if the values of the incidence are strongly decreasing with time.

Since in most applications a constant incidence is assumed a modification of the ordinary Shewhart method is suggested. With this modification it is possible to obtain a constant predictive value in the whole range of observations or in some interesting interval.

KEY WORDS: Predictive Value; Shewhart; Moving Average.

CONTENTS

1 INTRODUCTION

This paper deals with the situation where the number of observations is successively increasing and successive decisions are required. The goal is to detect an important change in an underlying process as soon as possible after the change has occurred. The time point when the change occurs is here regarded as a random variable.

The predictive value is a measure of how strong an indication of a critical event an alarm is. A constant predictive value is desired if the same action is supposed to be taken whether the alarm occurs late or early.

In Section 2 the situation is formally described with the notation introduced by Frisén and de Maré (1991)[1]. In Section 3 some measures of performance are discussed and in Section 4 some general aspects of the predictive value are given.

In the following section a selection of statistical standard methods are studied. The conditions necessary for a constant level of the predictive value of an alarm are examined. The methods used represent different ways to invoke the history of the process in the test procedure.

Finally in Section 6 some concluding remarks are made about how the methods discussed in this paper behave in some different situations and also the possibilities of modifications are further discussed.

2 SPECIFICATIONS.

The random process that determines the state of a system is denoted $\mu=\{\mu(u):u\in T\}$. In the following examples there is a shift in the mean value of Gassuian random variables. The shift occurs as a sudden jump from an acceptable value μ^0 to an unacceptable value μ^1 ($\mu^1=1$ in examples). $\mu(u)=\mu^0$ for $u=1,\ldots,\tau-1$ and $\mu(u)=\mu^1$ for $u=\tau,\tau+1,\ldots$. Both μ^0 and μ^1 are assumed to be known values and τ is a generalized random variable with density $P(\tau=k)=\pi_k$ ($k=1,2,\ldots$) and $\Sigma\pi_k=1-\pi_\infty$. The incidence of a shift is $p_k=P(\tau=k|\tau\geq k)$.

At decision time point s the aim is to discriminate between the critical event C(s) and some other set of events, in this case its complement, D(s). C(s) is a set of realizations of the μ-process. The critical event under consideration is $C(s)=\{\tau\leq s\}=\{\mu(s)=\mu^1\}$, the event that a shift occurs at s or earlier with the complement $D(s)=\{\tau>s\}=\{\mu(s)=\mu^0\}$. All changes before s are of interest, and hypotheses involved are such that they change successively.

The test procedures are based on observation $X_s=\{X(u):u\in T,u\leq s\}$ and in the following it is assumed that $X(1)-\mu(1),X(2)-\mu(2),\ldots$ are independent Gaussian random variables with expected value zero and the same and known variance ($\sigma^2=1$ in examples). A(s) is the alarm set given by the method under consideration. It is a set of events with the property that when X_s belongs to A(s) it is an indication that C(s) occurs and a hypothesis stating a stable system is rejected.

The surveillance is active in the sense that the procedure is stopped if A(1) occurs. Otherwise the complement $A^c(1)$ occurs, the procedure continues for s=2,3,... as long as $A^c(s-1)$ occurs.

3 MEASURES OF PERFORMANCE

In most application areas averages from the run length distributions, the time to an alarm, are used both for evaluating methods and comparisons between competing methods. ARL^0, the average run length under the hypothesis of a stable process, is defined as the average number of decisions taken until an alarm occurs when there is no change in the process under surveillance. The average run length under the alternative hypothesis, ARL^1 is the average number of decisions that must be taken to detect a true level change that occurred at the same time as the inspection started.

In this paper the ARL is used as the basis for a comparison between some test procedures. In the following examples all methods are assigned the same ARL^0. The average run length profiles for the test procedures are presented in figures. This measure has the obvious disadvantage of being sensitive to skew run length distributions. And furthermore when this measure is used the test procedure is restricted to taking only a single alternative hypothesis into consideration, which in most cases of continual surveillance is an unrealistic simplification.

The false alarm probability, conditioned on no alarm before s, for each time s is defined as

$$\alpha(s)=P(A(s)\,|\,D(s),A^C_{s-1}),$$

where

$$A^C_{s-1}=A^C(1)\cap\ldots\cap A^C(s-1)$$

The cumulative false alarm probability is here defined as

$$\alpha_s = P(A_s | D_s) =$$

$$1 - P(A^C(1) | D_s, A_0^C) \cdots P(A^C(s) | D_s, A_{s-1}^C)$$

where

$$D_s = D(1) \cap \ldots \cap D(s),$$

which in the present case is D(s). Also the probability of a false alarm at s is used in this paper and it is written as

$$\alpha^{\bullet}(s) = \alpha_s - \alpha_{s-1}$$

For each method discussed, α_s is shown in a figure.

The probability to neglect to stop the process when the critical event occurs is defined as

$$\beta(1) = P(A^c(1) | C(1)), \quad \beta(s) = P(A^c(s) | C(s), A^c_{s-1}) \quad (s \geq 2)$$

The predictive value of an alarm is a measure where the time point of the change is considered as a random variable, this measure will be defined in the next section.

4 GENERAL CHARACTERISTICS OF THE PREDICTIVE VALUE

The measure of performance that this paper is focused on is the predictive value of an alarm. The predictive value is the relative frequency of motivated alarms among all alarms at s,

$$PV(s)=P(C(s)\,|\,A(s),A^{C}_{s-1})=\frac{PMA(s)}{PMA(s)+PFA(s)}$$

where the unconditional probability of a false alarm at t^A is

$$PFA(t^{A})=[\prod_{t=1}^{t^{A}}(1-p_{t})]\;\alpha^{*}(t^{A})$$

where

$$\alpha^{*}(t^{A})=\alpha_{t^{A}}-\alpha_{t^{A}-1}$$

If the incidence of a shift is constant, $p_t=p$, PFA is reduced to

$$PFA(t^{A})=(1-p)^{t^{A}}\alpha^{*}(t^{A}).$$

The probability of a motivated alarm at t^A is defined as

$$PMA(t^{A})=\sum_{t^{C}=1}^{t^{A}}[\prod_{i=1}^{t^{C}}(1-p_{i-1})]p_{t^{C}}P(RL=t^{A}|\,\tau=t^{C}),$$

where $p_0=0$. With a constant incidence of a shift, PMA becomes

$$PMA(t^{A})=\sum_{t^{C}=1}^{t^{A}}(1-p)^{t^{C}-1}p\;P(RL=t^{A}|\,\tau=t^{C})$$

Contrary to passive surveillance, with active surveillance the predictive value typically has an asymptote below one, Frisén (1994)[2], but with a constant incidence the function is not necessarily monotonously increasing for all methods. If it is possible to obtain a constant level of the predictive value of an alarm, this would be a desirable property of a method in those applications where it is considered important to take the same action whenever an alarm occurs.

A constant level of the predicted value implies

$$\frac{PMA(1)}{PFA(1)} = \frac{PMA(t)}{PFA(t)}, \quad t > 1.$$

And furthermore

$$\frac{PV(t)}{1-PV(t)} = \frac{PMA(t)}{PFA(t)} < 1.0 \quad \Leftrightarrow \quad PV(t) < 0.5$$

Also,

$$\frac{p_1}{1-p_1} < \frac{P(A(1)|D(1))}{P(A(1)|C(1))} \qquad (i)$$

$$\Leftrightarrow \; PV(1) < 0.5,$$

since

$$\frac{PV(1)}{1-PV(1)} = \frac{PMA(1)}{PFA(1)} = \frac{p_1 P(A(1)|C(1))}{(1-p_1)P(A(1)|D(1))} < 1.0,$$

With the requirement of a constant level of the predictive value the condition (i) implies PV(t)<0.5.

5 THE PREDICTIVE VALUE OF SOME METHODS.

In this section a selection of standard methods representing different approaches in statistical surveillance are briefly described. The observation X_s, together with some rules for rejecting the hypothesis of a stable process constitutes a method. Some characteristic differences between methods are pointed out, particulary in regard to different structures in shift probabilities. The methods to be mentioned represent different ways to include the history of the process in the test procedure.

The two main approaches discussed in this paper are rules based on the last observation, $X_s=\{X(t):t\in T,t=s\}=X(s)$, that is no consideration is taken to earlier observations. This approach will henceforth be referred to as the Shewhart method.

The second approach is rules based on the history of the process, $X_s=\{X(t):t\in T,t\leq s\}$. Some methods based on moving averages will be discussed.

For methods where it is possible to obtain a constant, or almost constant level of the predictive value, the corresponding shift probabilities are illustrated in figures.

5.1 SHEWHART.

The Shewhart method is widely used in different application areas of statistical surveillance. For a situation formulated by Frisén and de Maré(1992)[1] the method is optimal in an extended Neyman-Pearson sense. This is the situation when the only interesting time point is the present, s=t. It is assumed that no previous observations include any valuable information about the state of the underlying process.

The performance of the Shewhart method in different situations often stands as a standard for comparisons when other methods, maybe competing ones, are evaluated. There are probably several reasons for this. One obvious reason is the fact that the method is a standard norm since it is used in a wide range of areas, i.e. quality control, where the Shewhart method often has the alarm limits 3 σ away from the target value, μ^0, Bergman and Klevsjö (1994)[3], which in the literature concerning quality control is referred to as the natural variation of the process. The illustrations in this paper are based on the Shewhart method with alarm limits closer to μ^0 and ARL^0 equals 15. For comparison with other methods, the corresponding sequence from the Shewhart method is added in each figure.

Another reason is that the method has a simple structure. This is achieved through the restrictive assumptions about the underlying process. If the successive observations are independent, successive values of the statistic used also becomes independent and by that measures of performance can be reached with straightforward calculations.

Also, the Shewhart method turns out to be a special case of many other methods, i.e. EWMA, MA and Hinkley's method. In more complicated test procedures, like Hinkley's method, the Shewhart test is usually one component in that procedure. Another example of the latter is Shewhart control charts with additional warning limits, or/and some other decision rules added to it. Notable is also that at t=1 almost any method has the same properties as the corresponding Shewhart test. Methods excluded are for example, tests based on window techniques with a window size greater than one.

Often the ordinary theory of testing hypotheses is applied to the problem, every time t the hypothesis is challenged by the alternatives it is done with identical probabilities of rejection for each test.

In the one-sided case the method only has one parameter G, the alarm limit. This makes the Shewhart method easy to work with and the interpretation of the parameter setting and the outcome at t becomes straightforward.

With Normally distributed observations it is possible to find distributions of the incidence that gives a constant level of the predicted value of an alarm. To obtain a constant and preferable high level of the predictive value, the probability that a shift in the underlying process has occurred at, or before the first observation, has to be rather high compared to the probability for the following observations. This will be further discussed in Section 5.1.2.

The method is defined by the quantity X(t), which might be the observation itself or some other quantity derived from a set of observations obtained in the choosen interval. The usual choice of quantity is an average calculated over equally spaced intervals and the method is also referred to as a Xbar Chart.

The hypothesis stating a stable process is rejected when, for the first time X(t) exceeds the alarm limit G. Usually G is expressed as

$$L\sigma_{X(t)},$$

where L is a chosen constant, and

$$\sigma_{X(t)},$$

the standard deviation of X(t). With independent observations and a common standard deviation the unconditional probability of a false alarm is $\alpha(t)=\alpha$, and the expected value in the RL distribution is α^{-1} when the hypothesis of no change in the process is true. Considering the sequential procedure the probability of a false alarm at or before a certain time point t, Fig.1, is α_t, where

$$\alpha_t-\alpha_{t-1} = \alpha^*(t) = (1-\alpha)^{t-1}\alpha.$$

In Fig. 1 and the following figures the twosided case is considered where the alarm limits G_u and G_l are located on each side of, and at the same distance from, μ^0.

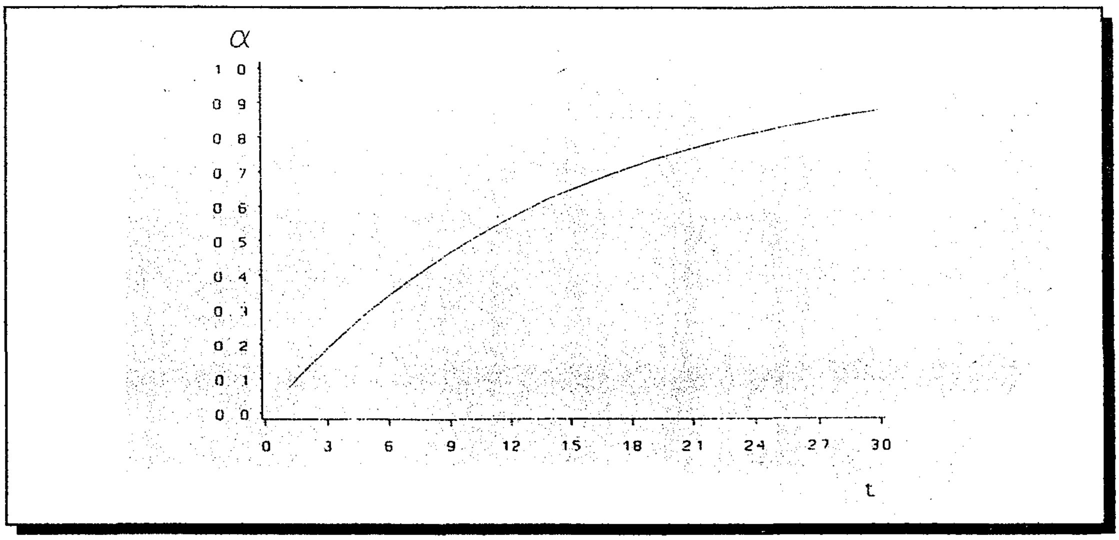


Figure 1. *The Shewhart method. The probability of a false alarm at time point t or earlier.*

In Fig. 1 the constant L is equals 1.83 and σ_X is unity.

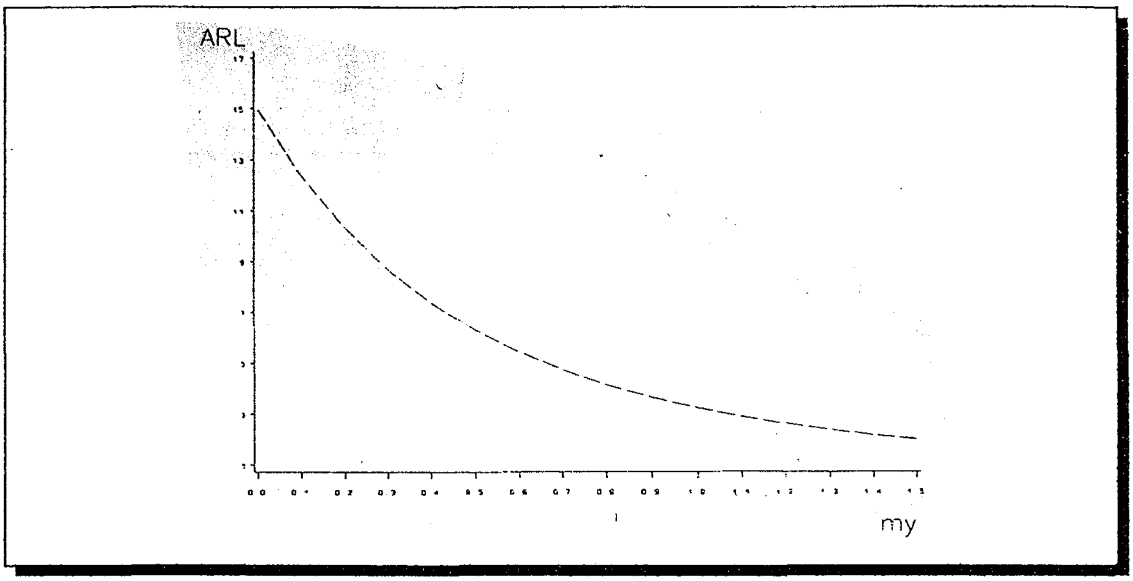


Figure 2 *The average run length profile for the Shewhart method, $ARL^0 = 15$.*

This gives a series, with an expected value of RL equal to 15, if the process is stable. All methods discussed in this paper have an average run length, ARL^0, equal to 15.

In Fig.2 the average run length profile is calculated with the usual assumption that the shift occurred at the same time as the inspection started, $t^C = 1$. If the time point for the shift is described by a stochastic variable, the probability of a false alarm at t^A is

$$PFA(t^A) = [\prod_{t=1}^{t^A} (1-p_t)](1-\alpha)^{t^A-1}\alpha,$$

and with a constant incidence of a shift, $p_t = p$, the probability of a false alarm at t^A becomes,

$$PFA(t^A) = (1-p)^{t^A}(1-\alpha)^{t^A-1}\alpha.$$

In this model the probability of not stopping at t when there has been a shift is $\beta(t)=\beta$, and the probability of a motivated alarm becomes

$$PMA(t^A)=\sum_{t^C=1}^{t^A}\left[\prod_{i=1}^{t^C}(1-p_{i-1})\right]p_{t^C}P(RL=t^A|\,\tau=t^C)$$

where

$$P(RL=t^A|\,\tau=t^C)=(1-\alpha)^{t^C-1}\beta^{t^A-t^C}(1-\beta)$$

With the same incidence of a shift at each time the probability of a motivated alarm becomes

$$PMA(t^A) = \sum_{t^C=1}^{t^A}(1-p)^{t^C-1}p\,P(RL=t^A|\,\tau=t^C)$$

Usually the alarm limits G_u and G_l are located at the same distance from μ^0, and are also considered to be constants. If any assumptions are made about the incidence of a shift it is mostly stated that it is constant with time. In the following section some consequences of this approach will be discussed, particulary when requirements on the predictive value are present.

The Shewhart method always has its highest probability of first detection at the time point of the shift, which is seen in that

$$P(RL=t^C|\,\tau=t^C)\le P(RL=t^C+n|\,\tau=t^C),\quad n=1,2,\ldots$$

$$\Leftrightarrow\quad (1-\beta)\le\beta^n(1-\beta).$$

Actually, given the same probability of a false alarm, α, no other method exists that has a higher detection probability in the present time point than the Shewhart method, Frisén and de Maré (1992)[1]. That is, if, according to the previous notation, s=t then the best choice of method is the Shewhart method.

5.1.1 CONSTANT INCIDENCE.

With the same incidence of a shift at each time t, and if $\alpha(t)=\alpha$, the predictive value can not be the same in the whole range of time points. This is true except for a test where the hypothesis is always rejected at t=1. This is seen in that,

$$PV(t) = PV(t+1) \quad \Rightarrow$$

$$PMA(t+1) = \frac{PFA(t+1)}{PFA(t)} PMA(t), \quad \{t=1,2,\ldots\},$$

$$p\beta^{t}(1-\beta)+(1-p)p(1-\alpha)\beta^{t-1}(1-\beta)+\ldots+(1-p)^{t}p(1-\alpha)^{t}(1-\beta) =$$

$$(1-p)(1-\alpha)[p\beta^{t-1}(1-\beta)+(1-p)p(1-\alpha)\beta^{t-2}(1-\beta)+\ldots+(1-p)^{t-1}p(1-\alpha)^{t-1}(1-\beta)],$$

$$\beta^{t}+(1-p)(1-\alpha)\beta^{t-1}+\ldots+(1-p)^{t}(1-\alpha)^{t} =$$

$$(1-p)(1-\alpha)[\beta^{t-1}+(1-p)(1-\alpha)\beta^{t-2}+\ldots+(1-p)^{t-1}(1-\alpha)^{t-1}] \quad \Leftrightarrow$$

$$\beta^{t} = 0 \quad \Rightarrow \quad \beta = 0.$$

Furthermore, with the same assumptions as above, the predictive value of an alarm is a monotonously increasing function in t. Suppose

$$PV(t) > PV(t+1),$$

$$\frac{PMA(t)}{PFA(t)} > \frac{PMA(t+1)}{PFA(t+1)} \quad \Rightarrow$$

$$\beta^{t}+(1-p)(1-\alpha)\beta^{t-1}+\ldots+(1-p)^{t}(1-\alpha)^{t} <$$

$$(1-p)(1-\alpha)[\beta^{t-1}+(1-p)(1-\alpha)\beta^{t-2}+\ldots+(1-p)^{t-1}(1-\alpha)^{t-1}] \quad \Rightarrow$$

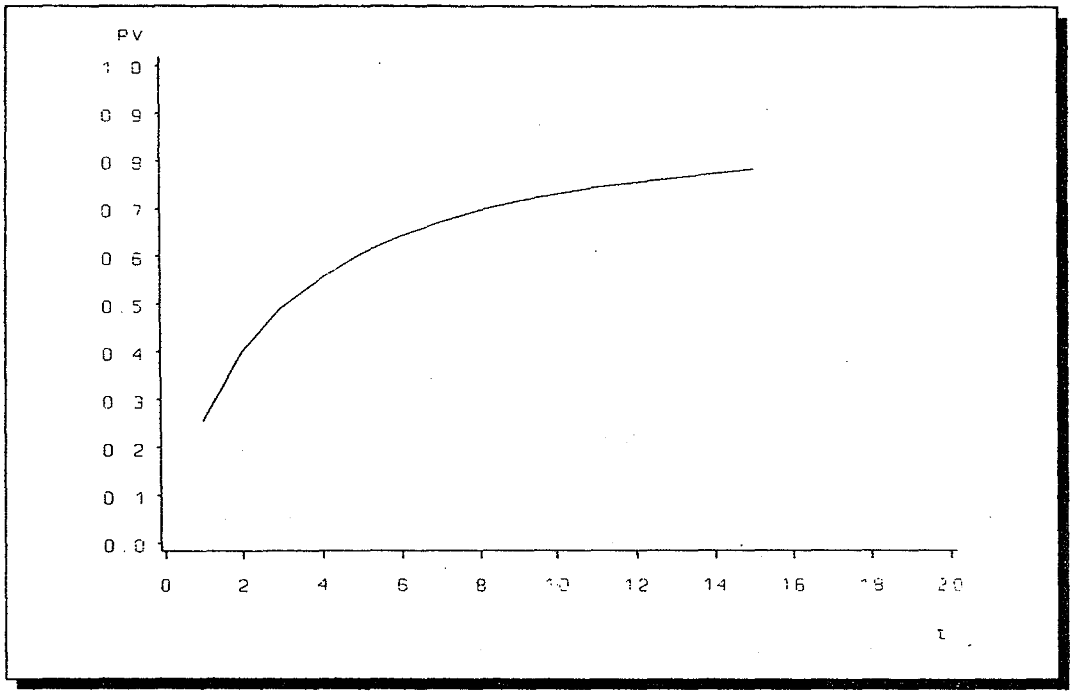
$$\beta^{t} < 0.$$

When t becomes large and the surveillance is active, the predictive value will increase to its limiting value, Frisén(1991)[4],

$$\lim_{t\to\infty} PV_t = \frac{p}{p+\alpha c},$$

where,

$$c = \frac{|(1-p)(1-\alpha)-1|}{(1-\alpha)(1-\beta)} + \frac{1}{1-\alpha}.$$



Figur 3 *Predictive value of an alarm in the case of the same probability of a shift at each time point. inc=0.1*

In Figure 3, p is set to 0.1. The low predictive value of an alarm for early observations is explained by the high false alarm probabilities compared to the probability of a shift. The probability of a false alarm, PFA, is a monotonously decreasing function in t. Suppose

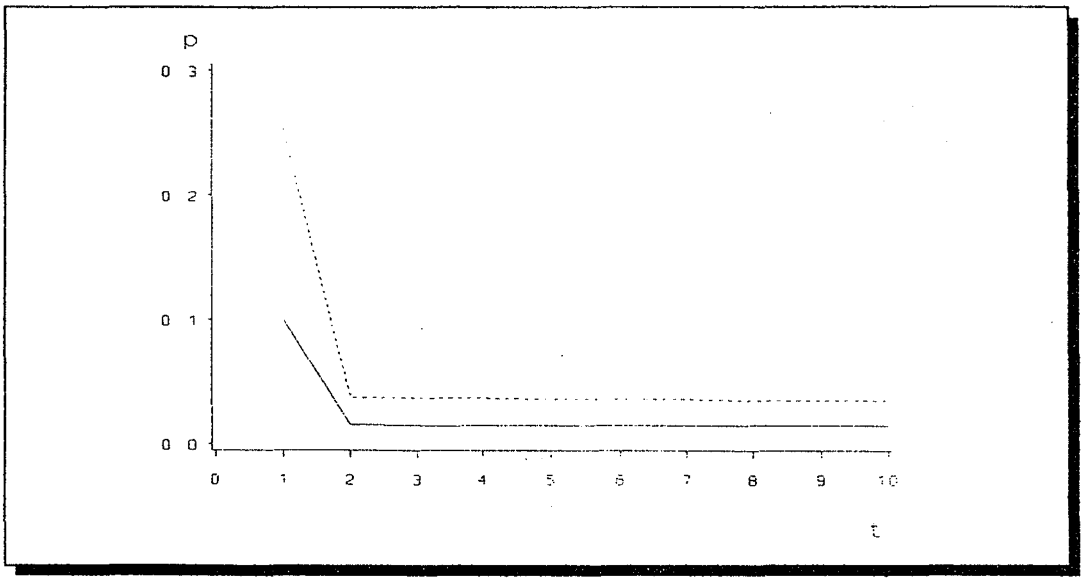
$$PFA(t) \leq PFA(t+1) \quad \Rightarrow$$

$$(1-p)^t(1-\alpha)^{t-1}\alpha \leq (1-p)^{t+1}(1-\alpha)^t\alpha$$

$$1-p \geq \frac{1}{1-\alpha} > 1.$$

5.1.2 VARYING INCIDENCE.

If, as usual, we have the Shewhart method with constant limits, that is $\alpha(t)=\alpha$, it is possible to obtain a constant level of the predictive value for specific series of decreasing incidence.



Figur 4 *The probability structure that gives a constant predictive value level. The dotted curve corresponds to PV=0.50*

In Figure 4, the lower curve, p_1 is set equal to 0.1 and p_t, $i>1$, are chosen in such a way that at a constant predictive value is obtained.

$$\frac{PMA(t)}{PFA(t)} = \frac{PMA(t+1)}{PFA(t+1)}$$

$$p_1 \frac{\beta^t}{(1-\alpha)} =$$

$$(1-p_{t+1})[p_1\beta^{t-1}+(1-p_1)p_2(1-\alpha)\beta^{t-2}+...+(1-p_1)\cdots(1-p)_{t-1})p_t(1-\alpha)^{t-1}]$$

$$-(1-p_1)[p_2\beta^{t-1}+(1-p_2)p_3(1-\alpha)\beta^{t-2}+...+(1-p_2)\cdots(1-p_t)p_{t-1}(1-\alpha)^{t-1}]$$

The required property is satisfied for the first step if $1-\alpha>\beta$ and

$$p_2 = p_1 \frac{(1-\alpha-\beta)}{(1-\alpha)}$$

For greater values of t the formula is more complicated, but it is no problem to compute the values. In figure 4 p_1 is set equal to 0.1 and $\mu^1=1$, the lower curve. In this case the predictive value is just above 0.25 at $t=1$, which means that on average only one out of four alarms at $t=1$ will actually be a motivated one. The curve representing the higher shift probabilities in figure 4 corresponds to the situation where the predictive value of an alarm at t^A is 0.50. Both curves indicate that a constant level of the predictive value for the Shewhart method sometimes requires a rather high shift probability at the first time point compared to the second, and later observations.

5.1.3 CONSTANT INCIDENCE AND VARYING ALARM LIMITS.

By letting the alarm limits of the Shewhart method change over time the predictive value can be made constant.

First, consider the first and second observation from a Normal distribution. If $\alpha(1) < \alpha(2)$ this implies that

$$\frac{\alpha(1)}{1-\beta(1)} < \frac{\alpha(2)}{1-\beta(2)} .$$

With a constant incidence of a shift it is possible to obtain the same predictive value of an alarm for both points.

$$PV(2) = PV(1) \quad \Rightarrow$$

$$\frac{PMA(2)}{PFA(2)} = \frac{PMA(1)}{PFA(1)} \quad \Rightarrow$$

$$\frac{\alpha(2)}{1-\beta(2)} = \frac{\alpha(1)}{1-\beta(1)} \cdot k,$$

where

$$k = \frac{\beta(1)}{(1-p)(1-\alpha(1)} + 1 > 1.$$

This means that, with normally distributed observations, the alarm limits for the second observation should be closer to μ^0 than for the first one if a test procedure with the same predictive value of an alarm is required.

It is possible to choose the limit g(t+1) in such a way that a constant level of the predictive value is obtained.

$$PV(t) = \frac{PMA(t)}{PMA(t)+PFA(t)} \quad \Rightarrow$$

$$PV(t) = (1+\frac{PFA(t)}{PMA(t)})^{-1} \quad \Rightarrow$$

$$PV(t) = \frac{1}{1+\frac{(1-p)^{t}\prod_{i=1}^{t-1}(1-\alpha(i))\alpha(t)}{\sum_{j=1}^{t}P(\tau=j)P(RL=t \mid \tau=j)}}$$

where the sum in the numerator can be written,

$$p(1-\beta(t))[\beta(1)\cdots\beta(t-1)+(1-p)(1-\alpha(1))\beta(2)\cdots\beta(t-1)+...+$$

$$(1-p)^{t-2}(1-\alpha(1))\cdots(1-\alpha(t-2))\beta(t-1)+(1-p)^{t-1}(1-\alpha(t-1))].$$

This implies that

$$\frac{PV(t)}{1-PV(t)} = \frac{1-\beta(t)}{\alpha(t)}\cdot f_{t-1},$$

where f_{t-1} only depends on the characteristics of the surveillance before time point t.

$$\frac{1-\beta(t)}{\alpha(t)} = \frac{1-\Phi(g-\mu^{1})}{1-\Phi(g-\mu^{0})}$$

increase continuously when the limit, g(t), increase.

Choose g(t+1) in such a way that,

$$\frac{1-\beta(t+1)}{\alpha(t+1)} = \frac{f_{t-1}}{f_t} \cdot \frac{1-\beta(t)}{\alpha(t)}.$$

It can be shown that,

$$g(t) \geq g(t+1) \quad \Rightarrow$$

$$PV(t) \leq PV(t+1) \quad \Rightarrow$$

$$f_{t-1} < f_t.$$

A constant level of the predictive value of an alarm requires that g(t)>g(t+1).

For example, a one sided 3σ control chart (Xbar Chart) is used on Normal distributed observations, $\mu^0=0$ and $\sigma_X=1$, and the aim is to detect a shift of size σ. If the incidence is constant and equals 0.1, this test procedure will give a predictive value of an alarm at t=1 that is just above 0.65. To obtain the same predictive value of an alarm at time point two the alarm limit has to be moved to a 2.1σ distance from μ_0.

5.2 MOVING AVERAGES

The results in the following sections are mainly obtained through computer simulations.

Surveillance methods based on moving averages dates back to about forty years ago, and since then, the most investigated and discussed method in this family has been the ones with weights exponentially decreasing in time, Exponentially Weighted Moving Average. The EWMA methods are developments of the Moving Average, MA, method based on k observations, Roberts(1959)[6]. Some consequences of choosing these type of methods, or methods with similar properties, are discussed. In this paper the MA method based on k observations is modified to a Expanding Average, k=s, to illustrate some effects on the ARL and the predictive value properties of a method if the alarm probabilities in the initial face of the series are not constant.

5.2.1 EXPANDING AVERAGE.

Moving Windows are techniques to describe the influence of earlier observations in the process. The quantity are usually averages calculated from the last k observations at decision time point s. Observations obtained before s-k+1 are considered of no interest at all, while the last k observations usually are considered equally important. By using this techniques to modify methods improved predicted value properties can be achieved since this technique demands k-1 auxiliary points before the first test can be made. If the parameter k is one in this model the Shewhart method is obtained.

The Expanding Average is a special case of a Moving Average. If at decision time point s all available observations are considered equally important, k=s, a Moving Average with equal weights, k^{-1}, assigned to the last k observations, is obtained. This model is mainly discussed because of its remarkably good ARL properties and the general effect on the ARL measure if a skewed RL-distribution is present, Fig 5.

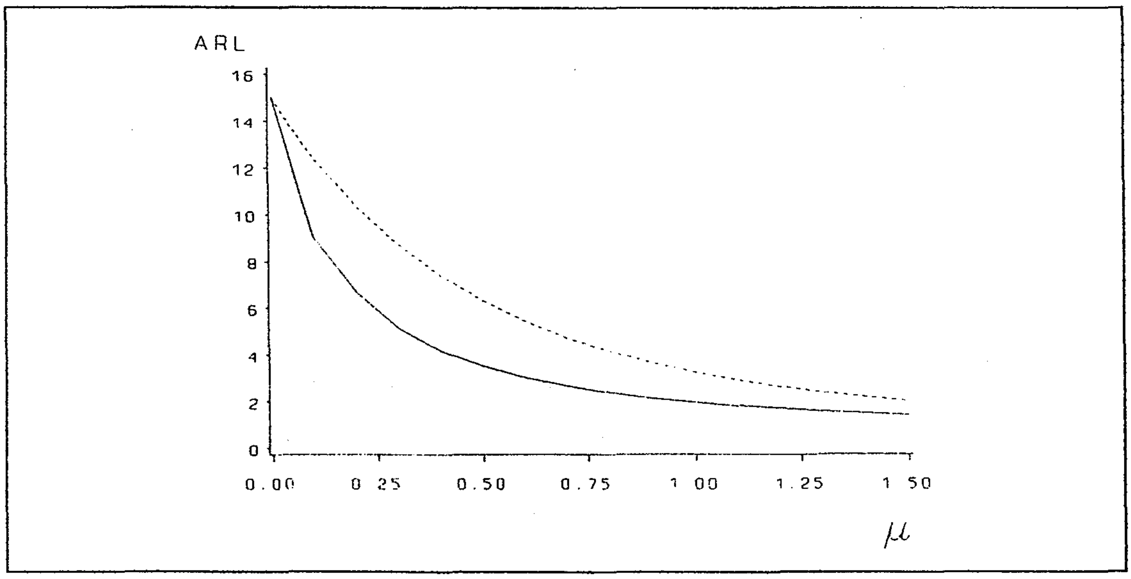


Figure 5. *The ARL profile of the Expanding Average and the Shewhart method (dotted line).*

For each new observation an average is calculated with the actual s as denominator,

$$M_s = \frac{\sum_{k=1}^{k=s} x_k}{s} ,$$

The quantity M_s converges to μ^0 as s becomes large, and the variance of M_s becomes small.

The process under surveillance is considered to be in balance as long as the outcome on M_s stays within the alarm limits

$$g_s = L\sigma_X \frac{1}{\sqrt{s}} ,$$

where σ_X is the same known standard deviation of the observation X_t.

In the Figures 5-8 L is 0.924 and $\sigma_X=1$, this gives an average run length of 15 if $D(t)\equiv D$, the process is in control. In Fig. 6 the probability of a false alarm, α_t, is compared with the Shewhart method.

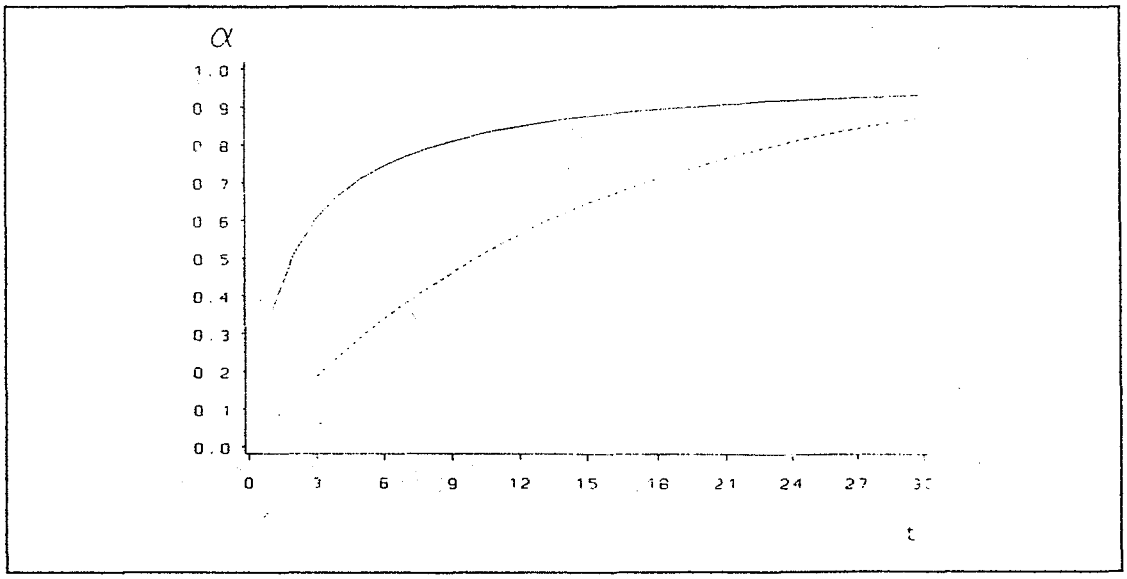


Figure 6. *The probability of a false alarm at or before t for the Expanding Average and the Shewhart method (dotted line).*

This method is characterised by a initially high alarm probability. With this choice of parameter values the false alarm probability at $t=1$ is 0.36. This makes the run length distribution markedly skew to generate an average of 15. Given a shift in the process, this method will most of the times generate an early alarm, but in a few trials it will have difficulties to detect that change. This gives a rather impressive ARL-profile for the method, but it is only suitable in situations where late alarms are not crucial. Given a shift at t^C, the method has its highest probability of detection at t^C at the beginning of the series, just like the Shewhart method. But if there is a late shift the method have its highest probability of detection at some time point later then t^C, and also, the probability of an immediate detection is decreasing with t. For the case illustrated (Fig.5-7) the probability of an immediate detection if $t^C=1$ is 0.55, while if $t^C=20$, the probability falls below 0.1.

In the following sections the predictive value properties of the Expanding Average will briefly be discussed.

5.2.1.1 CONSTANT INCIDENCE.

With this choice of L and a constant incidence of a shift it is not possible to obtain a constant level of the predictive value of an alarm. In the Normal distribution the probability of a motivated alarm, PMA, and the probability of a false alarm, PFA, are decreasing functions in t. The predictive value of an alarm is an increasing function in t which means that an application where this method is suitable has to be such that a low predictive value at the beginning of the sequence is acceptable, Fig. 7.

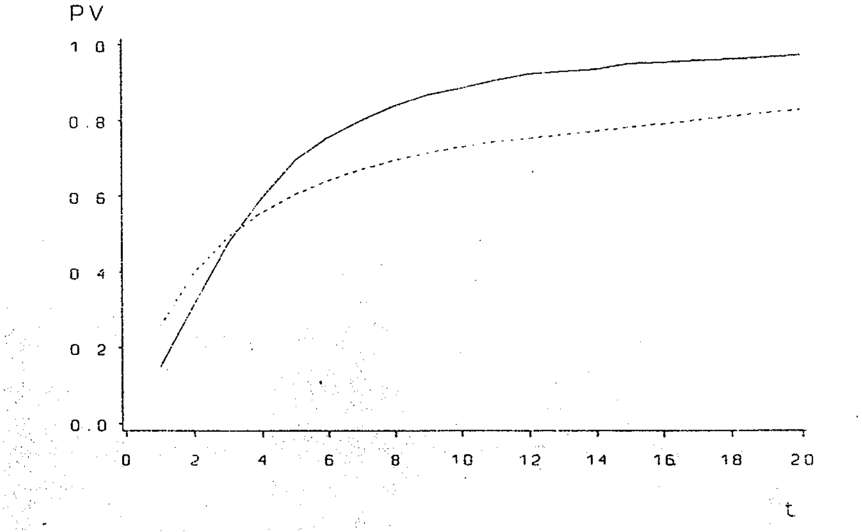


Figure 7. *Predictive value of an alarm for Expanded Average and the Shewhart method, the lower curve. The probability of a shift is 0.1 at each time point.*

5.2.1.2 VARYING INCIDENCE.

In a Normal process it is possible to obtain a constant predictive value of an alarm with this method. Though the incidence of a shift has to be rather high at the first time point relatively to the following points.

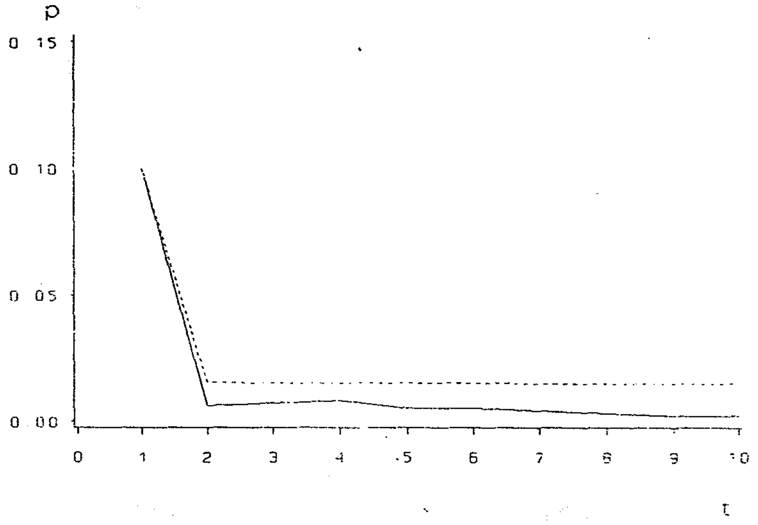


Figure 8. *The incidence of a shift that gives a constant predictive value of an alarm for the Expanding Average method and the Shewhart method (the lower curve).*

In the example the probability of a shift, the incidence, at t=1 is 0.1 and then decreases to give an overall predictive value of an alarm of 0.26 for the Shewhart method and 0.15 for the Expanding Average method. The 50 percent level of the predictive value of an alarm for the Expanding Average corresponds to p_1 equals 0.39.

5.3.2. EXPONENTIALLY WEIGHTED MOVING AVERAGE

The EWMA methods are a group of methods with a particular weight pattern that assigns the history of the underlying process weights in an infinite, geometrically decreasing sequence from the most recent back to the first observation.

If the parameter λ is equal to one in the EWMA model the Shewhart method is obtained, and if λ tends to zero the method degenerates towards a test with the present observation excluded. EWMA(g) methods where the alarm limits, g, are based on the standard deviations of the statistic usually (unimodally distributed statistic) have higher alarm probabilities than the standard version with straight alarm limits, G, when the limiting value of g is G. This generates alarm probabilities, and predictive value properties, closer to the Shewhart method.

If a Fast Initial Response feature, Lucas and Crosier(1990)[6], is added to the standard EWMA(G) the alarm probabilities in the initial points are increased in a way similar to the EWMA(g) methods.

$$Z_0 = \mu^0,$$

$$Z_t = (1-\lambda)Z_{t-1} + \lambda X_t, \; t > 0.$$

The EWMA is sometimes referred to as a geometric moving average since it can be written as a moving average of the current and the past observations,

$$Z_t=\lambda\sum_{j=0}^{t-1}(1-\lambda)^j X_{t-j}+(1-\lambda)^t Z_0 .$$

If the X's are independent and have a common standard deviation σ_X, the standard deviation of Z_i is

$$\sigma_{z_t}=\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2t}\ )}\sigma_x$$

For the first observation σ_Z takes the value $\lambda\sigma_X$, and as i increases σ_Z increases to its limiting value

$$\sigma_z=\sqrt{\frac{\lambda}{(2-\lambda)}}\sigma_x$$

An out of-control alarm is triggered if the estimate $|Z_t|$ falls outside the limits

$$g=L\sigma_{z_t}\ .$$

Usually the limiting value σ_Z rather than σ_{Zt} is used in constructing the warning-limits for EWMA control charts, e.g. Roberts(1959)[5], Robinson and Ho(1978)[7], Crowder(1987)[8] and Lucas and Saccucci(1990)[9]. For a twosided control chart this results in two straight alarm limits, G_u and G_l. They are usually placed at equal distance, on each side, from the nominal level, μ^0. As shown in Fig. 9 the EWMA(G) probability of a false alarm at or before t has a structure similar to the Shewhart method. In this example λ is arbitrarily set to 0.5, any other choice would also have been possible. Actually, for a given shift, μ, the parameter λ and L can be chosen in such a way that both ARL^0 and ARL^1 are exactly the same as for any other method, specified by one or two parameters.

The parameter L is set to 0.924 to satisfy the ARL requirement of 15. The low initial alarm probabilities are consequences of using G_u and G_l as alarm limits. By choosing G_u and G_l as alarm limits the lowest alarm probabilities are obtained at t=1. This fact is sometimes considered as a

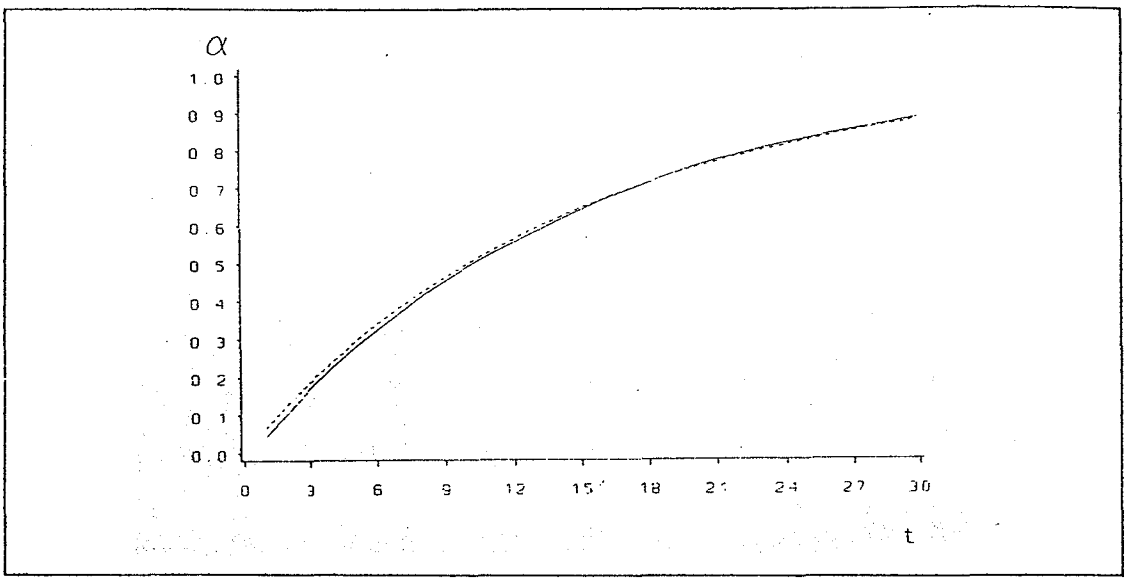


Figure 9 *The probability of a false alarm at time point t or earlier for the EWMA method and the Shewhart method (the dotted line).*

disadvantage of the EWMA(G) method, and other methods with the same structure. To overcome this a FIR feature can be applied to the method. A desirable change of the probability pattern can be accomplished in several different ways. Two different approaches using the same quantity and with the same purpose are discussed and their relation to the EWMA(G) is established.

First, by using σ_{Zt}, alarm limits are created that start at a distance of $L\lambda\sigma_X$ from the nominal level and increase to $L\sigma_Z$. By using the variance of the statistic a EWMA(g_t) is created. This gives faster detection for early shifts, but also higher probabilities of early false alarms are obtained. If the EWMA(G) method, with $G=L\sigma_Z$, is transformed to a EWMA(g_t) in such a way that,

$$\lim_{t\to\infty} L\sigma_{z_t} = G$$

then,

$$\alpha^{EWMA(G)} < \alpha^{EWMA(g_t)} \quad \forall \lambda, t,$$

that is given the same value on $L\sigma_Z$ for both methods, the early alarm probabilities that are obtained after a modification, are higher than they would have been without the modification. The difference decreases as t gets larger. How fast this goes depends on the choice of λ which also determines the magnitude of the difference.

If the FIR technique is applied to the EWMA(G) method a similar false alarm pattern is recognized as for the variance corrected EWMA(g_t). This way to accomplish higher alarm probabilities are of a more general kind than the previous and can be applied to almost any other method. The idea is to assign a starting value, Z_0, that is not equal to the expected value in the process. By this a faster detection is obtained if there is an early shift in the process. If no shift occurs the statistic, Z_t, will find its way back to the expected level. The hypothesis is rejected for the first time z_t exceeds the limit of type g_t,

$$g_t = G - (1-\lambda)^t A,$$

When the constant A is written as,

$$A = L\sigma_x \frac{\sqrt{\frac{\lambda}{2-\lambda}} - \lambda}{1-\lambda},$$

it is the same as if Z_0 is selected in such a way that, at $t=1$, the same probabilities of alarm as for the EWMA(g_t) are obtained, and

$$\lim_{t\to\infty} G - (1-\lambda)^t A = L\sigma_z$$

With the same way of reasoning as above, given the same value on $L\sigma_Z$

$$\alpha_t^{\mathrm{EWMA}(g_t)} \leq \alpha_t^{\mathrm{EWMA}(G)^{FIR}} \qquad \forall \lambda, t$$

with equality at $t=1$. This is seen in that

$$L\sigma_x\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2t}} \geq L\sigma_x-(1-\lambda)^{t-1}L\sigma_x\sqrt{\frac{\lambda}{2-\lambda}-\lambda}$$

for all λ and t. The probability of not rejecting the null hypothesis when one of the alternatives is true is affected in an analogous way for both modification procedures.

A modification of a method might have an influence on all the measures involved in an analysis. But if, like here, the procedures involved are specified by some known functions, the directions of changes are known.

In figure 10 the ARL profile for the EWMA(G) is compared to the Shewhart method.

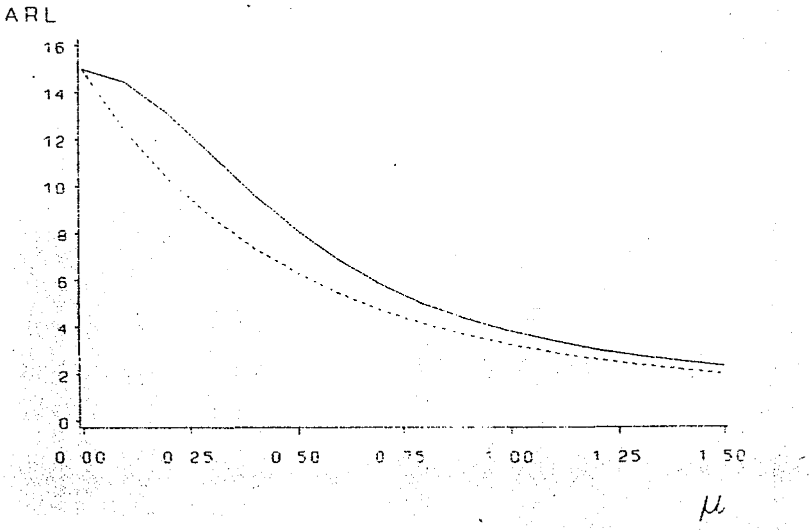


Figure 10 *The ARL profile for the EWMA(G) and the Shewhart method, the lower curve.*

Its relatively poor ARL profile is explained by low alarm probabilities at the beginning of the run, and the way the ARL^1 measure is defined.

If the EWMA(G) is modified according to the previous, both the ARL^0 and the ARL^1 decrease. The only possibility to equality among these methods is if the rejection rule is to always reject at t=1. The magnitudes of the changes are determined by the choice of λ and L. The relationship between the average run length properties given by these examples can be summarized as follows.

Given the same choice of $L\sigma_Z$ for all three methods, and Z_0 in EWMA(FIR) are such that, the probability of a false alarm at the first time point is the same as for EWMA(g_t), then

$$ARL^{0}_{\mathrm{EWMA(G)}^{FIR}} \leq ARL^{0}_{\mathrm{EWMA}(g_t)} < ARL^{0}_{\mathrm{EWMA(G)}} \, ,$$

and

$$ARL^{1}_{\mathrm{EWMA(G)}^{FIR}} < ARL^{1}_{\mathrm{EWMA}(g_t)} < ARL^{1}_{\mathrm{EWMA(G)}} \, .$$

When a method is modified in the way described above its seems reasonable to do so, only if there is a high probability of a shift at the beginning of the series. Using the ARL measure as a guideline might be misleading since crucial information about later observations are lost. The run length distribution at t=1 can become rather extreme compared to distributions obtained at a later shift. A consequence of this is that, in the light of other measures the ARL^1 alone becomes a poor measure of performance when the effect of the FIR features are evaluated. In particular if the assumption of a high initial shift probability is not satisfied.

5.3.2.1 CONSTANT INCIDENCE.

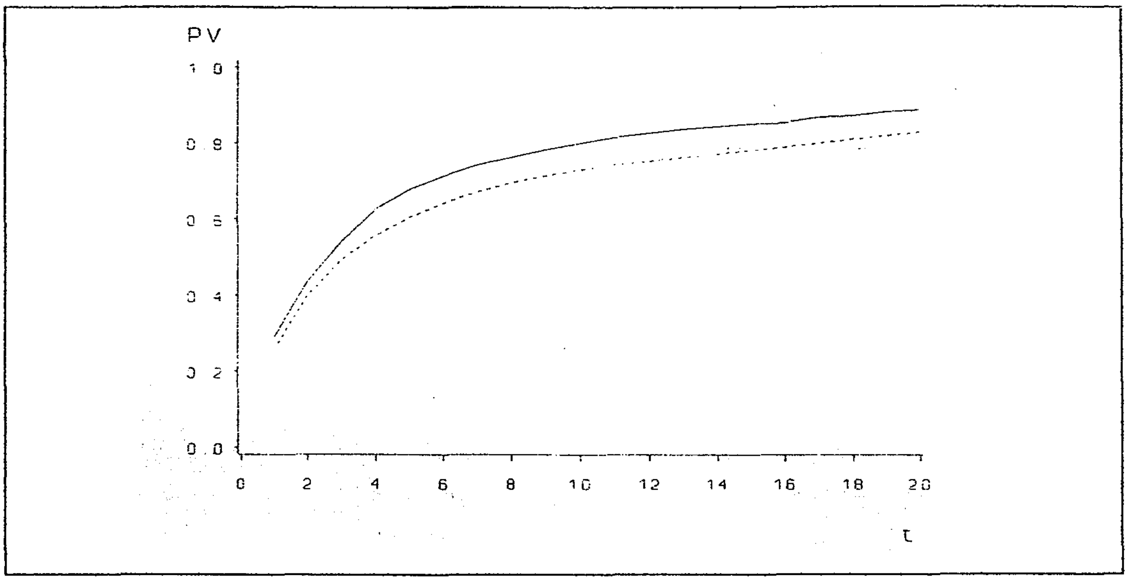


Figure 11 *The predictive value of an alarm with a constant probability of a shift. The EWMA(G), the upper curve, and the Shewhart method.*

The predictive value of an alarm curve for the EWMA(G) has a shape similar to the Shewhart tests, which was indicated by the α-structure, Fig. 6. Consider the case where a Shewhart test is replaced by a EWMA(G) in such a way that $G=L\sigma_Z$. With this restriction the EWMA(G) has a higher PV than the Shewhart, figure 10, and equality only for λ equals one. The differences due to the choice of statistic is mainly stated at the behaviour at the first two observations after a shift. The Shewhart method always has its highest detection probability at the time point of the shift, t^C. With this choice of λ, the detection probability for the EWMA(G) has its peak at t^C+1, and then decreasing faster than for the Shewhart method. In case of a shift, the EWMA(G) is most sensitive to that shift at t^C+1 or higher, depending on the choice of λ. The EWMA(G) might be a candidate if $s>t$, and in particular if a specific time point after the shift is of main interest.

5.3.2.2 VARYING INCIDENCE.

With the parameter setting given in Section 5.3.2 it is not possible to obtain an overall constant level of the predictive value of an alarm.

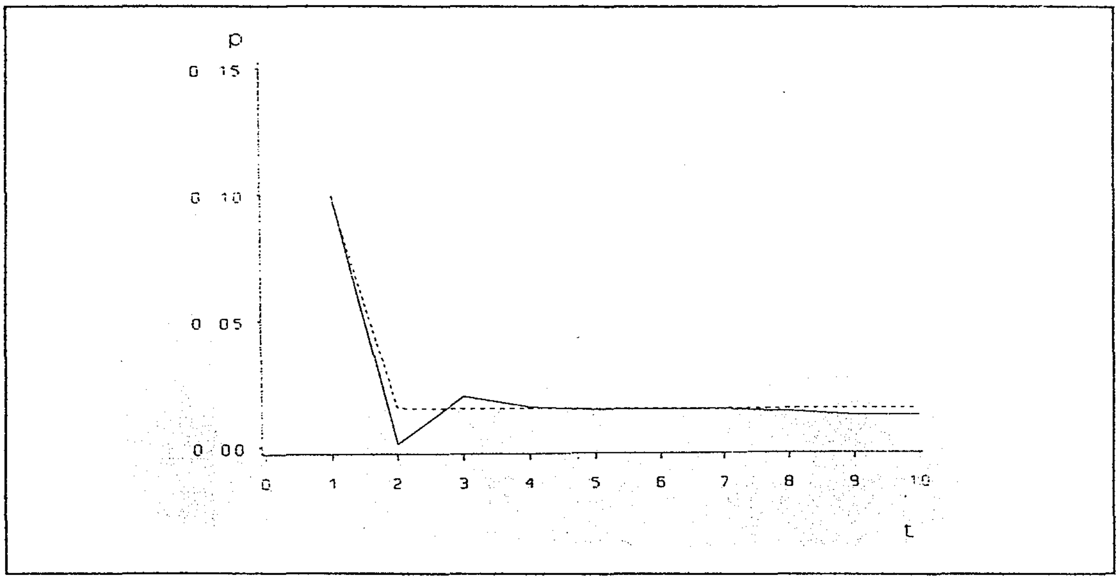


Figure 12. *The probabilities of a shift that generate a constant level of the predictive value of an alarm. The EWMA(G) and the Shewhart method.*

The shift probability for the EWMA(G) has to be negative at t=2 to obtain the desired property.

6 CONCLUDING REMARKS.

A constant level of the predictive value is a reasonable requirement in a situation where an alarm is considered equally important regardless of when the alarm occurs. That is, the surveillance model has to be designed in such a way that the proportionality between the probability of a false alarm, PFA and the probability of a motivated alarm, PMA, is constant and preferably in such a way that the predictive value is greater than 0.5.

If the predictive value is used as a design criterion when a surveillance system is designed knowledge or assumptions about the shift probability structure is crucial, both with regard to the choice of method as well as for the parameter setting. Since all the methods considered in this paper have the properties of the ordinary Shewhart method at $t=1$ and the predictive value functions are increasing it is possible to calculate the minimum predictive value for each method. The rather low values of the predicted value of an alarm is partially explained by the choice of two-sided tests.

If a constant incidence of a shift is present the predictive value can not be constant for the ordinary Shewhart method. If the method is modified in such a way that the alarm limits are moved away from μ^0 for early observations it is possible to obtain a constant level of the predictive value in the whole range of observations. Generally, for a method characterized by a monotonously increasing predictive value function the method has to be modified in such a way that the probabilities of an alarm are decreasing for early observations to reach a constant predictive value. This also means that methods designed to detect early shifts can be applied when initially there is a rather high probability of a shift. If, for instance , the incidence of a shift is constant, the system under surveillance has to be such that the consequences of a false alarm is not crucial.

ACKNOWLEDGEMENT.

This work has been supported by the Swedish Council for Research in the Humanities and Social Sciences. Marianne Frisén, my supervisor, suggested the topic and has encouraged me in every possible way through this work. I am deeply grateful for this. I also thank my colleague Martin Gellerstedt and other colleagues at the Department of Statistics at the Göteborg University, for many enjoyable (and some times constructive) discussions. I am also grateful to Göran Broström and Peter Wessman who have commented upon an earlier version of this paper.

7. REFERENCES.

1 Frisén, M and de Maré, J. (1991), "Optimal surveillance," *Biometrika*, 78, 271-280.

2 Frisén, M. (1994), "Characterization of methods for surveillance by optimality," *Research Report 1994:2*. Department of Statistics, Göteborg.

3 Bergman, B and Klevsjö, . (1994), "Quality from customer needs to customer satisfaction," Studentlitteratur.

4 Frisén, M. (1992), "Evaluations of methods for statistical surveillance," *Statistics in Medicine*, 11, 1489-1502.

5 Roberts, S. W. (1959), "Control Charts Tests Based on Geometric Moving Averages," *Technometrics*, 1, 239-250.

6 Lucas, J. M and Crosier, R. B. (1982), "Fast initial response for cusum quality schemes: give your cusum a head start," *Technometrics*, 24, 199-205.

7 Robinson, P. B and Ho, T. Y. (1978), "Average Run Length of Geometric Moving Average Charts by Numerical Methods" *Technometrics*, 20, 85-93.

8 Crowder, S. V. (1987), "A simple method for studying run-length distribution of exponentially weighted moving average charts," *Technometrics*, 29, 401-407.

9 Lucas, J. M and Saccucci, M. S. (1990), "Exponentially weighted moving average control schemes: properties and enhancements," *Technometrics*, 32, 1-12.

1993:1	Frisén, M & Åkermo, G.	Comparison between two methods of surveillance: exponentially weighted moving average vs cusum
1993:2	Jonsson, R.	Exact properties of McNemar's test in small samples.
1993:3	Gellerstedt, M.	Resampling procedures in linear models.
1994:1	Frisén, M.	Statistical surveillance of business cycles.
1994:2	Frisén, M.	Characterization of methods for surveillance by optimality.
1994:3	Frisén, M. & Cassel, C.	Visual evaluation of statistical surveillance.
1994:4	Ekman, C.	A comparison of two designs for estimating a second order surface with a known maximum.
1994:5	Palaszewski,B.	Comparing power and multiple significance level for step up and step fown multiple test procedures for correlated estimates.