EWMA Control Charts with Time-Varying Control Limits and Fast Initial Response

STEFAN H. STEINER

University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

The control limits of an exponentially weighted moving average (EWMA) control chart should vary with time, approaching asymptotic limits as time increases. However, previous analyses of EWMA charts consider only asymptotic control limits. In this article, the run length properties of EWMA charts with time-varying control limits are approximated using non-homogeneous Markov chains. Comparing the average run lengths (ARL's) of EWMA charts with time-varying control limits and results previously obtained for asymptotic EWMA charts shows that using time-varying control limits is akin to the fast initial response (FIR) feature suggested for cumulative sum charts. The ARL of the EWMA scheme with time-varying limits is substantially more sensitive to early process shifts, especially when the EWMA weight is small. An additional improvement in FIR performance can be achieved by further narrowing the control limits for the first twenty observations. The methodology is illustrated assuming a normal process with known standard deviation where we wish to detect shifts in the mean.

Introduction

EXPONENTIALLY weighted moving average (EWMA) control charts and other sequential approaches, like cumulative sum (CUSUM) charts, are an alternative to Shewhart control charts and are especially effective in detecting small persistent process shifts (Montgomery (1991)). First introduced by Roberts (1959), EWMA charts have a fairly long history, but only recently have been evaluated analytically (Crowder (1987) and Lucas and Saccucci (1990)). The EWMA chart also is known to have optimal properties in some forecasting and control applications (Box, Jenkins, and MacGregor (1974)). In this article, we focus on the quality monitoring applications. For monitoring the process mean, the EWMA control chart consists of plotting

$$z_t = \lambda \bar{x}_t + (1-\lambda)\, z_{t-1}, \quad 0 < \lambda \leq 1 \qquad (1)$$

versus time, t, where λ is a constant and the starting value, z_0, is set equal to an estimate of the process mean, often given as $\bar{\bar{x}}$ and calculated from previous data. In this definition, $\bar{x}_t$ is the sample mean from time period t, z_t is the plotted test statistic, and λ is the weight assigned to the current observation. The definition of the EWMA test statistic given in Equation (1) can be adapted to monitor any process parameter of interest.

By writing out the recursion in Equation (1), the EWMA test statistic is shown to be an exponentially weighted average of all previous observations. In quality monitoring applications, typical values for λ are between 0.05 and 0.25, although larger values may be used in forecasting and control applications. In the limiting case of λ=1, the EWMA chart is the same as a Shewhart $\bar{X}$ control chart. Using an EWMA chart, the process is considered out of control whenever z_t falls outside the range of the control limits. EWMA control limits are discussed in detail in the next section.

As shown in Montgomery (1991), the control limits for EWMA charts should be time-varying since the variance of z_t depends on t and because the effect of the starting constant, z_0, decreases as t increases. However, all past studies of the properties of the EWMA chart have used fixed (asymptotic) control limits to make analysis easier. This article presents a methodology for determining the expected value and standard deviation of the run length of the EWMA chart with time-varying control limits. Numerical results are given for monitoring the mean of a normal distribution. EWMA control charts with time-varying control limits are useful because pro-

Dr. Steiner is an Assistant Professor in the Department of Statistics and Actuarial Sciences. He is a Member of ASQ. His email address is shsteine@uwaterloo.ca.

cesses are likely different from the target value when a control scheme is initiated due to start-up problems or because of ineffective control action after the previous out-of-control signal. In addition, after a process change or adjustment we often wish to quickly confirm that the change had the desired effect.

Using time-varying control limits has an effect similar to the fast initial response (FIR) feature recommended by Lucas and Crosier (1982) for CUSUM charts, since it helps detect problems with the start-up quality. For CUSUM's, the FIR feature substantially decreases the ARL for an out-of-control process, while only slightly decreasing the ARL of an in-control process. For EWMA charts, Lucas and Saccucci (1990) suggested the simultaneous use of two one-sided EWMA charts with initial states different than zero as an implementation of the FIR feature. One EWMA chart monitors for increases in the process parameter, while the other chart monitors for decreases. Rhoads, Montgomery, and Mastrangelo (1996) adapt the Lucas and Saccucci approach by allowing the one-sided EWMA chart to have time-varying control limits as given by Equation (2) and discussed in the second section of this paper. Rhoads, Montgomery, and Mastrangelo (1996) compare the run length properties determined through simulation. Both these implementations of FIR-EWMA charts require the use of two EWMA charts to monitor a process for two-sided shifts.

This article shows that the use of time-varying control limits makes an EWMA chart more sensitive to start-up quality problems than the traditional asymptotic limits. If additional protection to start-up quality problems is desired, then the further narrowing of the time-varying control limits according to an exponential weighting scheme mimics the FIR feature. The derivation of time-varying control limits for an EWMA chart is presented in the second section, and the effect of time-varying control limits is illustrated for a simple example. The third section uses numerical results to contrast and compare EWMA charts with time-varying control limits and EWMA charts with asymptotic limits. The fourth section introduces a FIR feature for two-sided EWMA charts and shows that this approach is superior to methods suggested previously by Lucas and Saccucci (1990) and Rhoads, Montgomery, and Mastrangelo (1996). The Appendix shows that the run length properties of an EWMA chart with time-varying control limits can be approximated using a non-homogenous Markov chain.

EWMA Control Charts with Time-Varying Control Limits

The mean value and variance of z_t are easily derived from Equation (1) (Montgomery (1991)). Assuming the $\bar{x}_i$'s are independent random variables with mean, μ_x, and variance, σ_x^2/n, where n is the sample size used at each time interval to calculate $\bar{x}_i$, we obtain

$$\mu_{z_t} = \mu_x \quad \text{and}$$
$$\sigma_{z_t}^2 = \frac{\sigma_x^2}{n}\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2t}\right]. \qquad (2)$$

Notice that the variance of z_t is a function of time. This should be expected since the number of observations used to derive z_t varies with time and the influence of the initial fixed value z_0 slowly decreases.

Control limits for an EWMA control chart are typically derived based on $\pm L$ sigma limits, where L is usually equal to 3, as in the design of Shewhart control chart limits. Thus, the time-varying upper and lower EWMA control limits, $UCL(t)$ and $LCL(t)$, respectively, are given by

$$UCL(t) = \mu_x + L\sigma_x\sqrt{\frac{\lambda\left[1-(1-\lambda)^{2t}\right]}{(2-\lambda)n}} \quad \text{and}$$
$$LCL(t) = \mu_x - L\sigma_x\sqrt{\frac{\lambda\left[1-(1-\lambda)^{2t}\right]}{(2-\lambda)n}}, \qquad (3)$$

where, in applications, μ_x and σ_x are typically estimated from preliminary data as the sample mean and sample standard deviation. As t increases, $UCL(t)$ and $LCL(t)$ converge to the asymptotic control limits, UCL and LCL, which are given by $\mu_x \pm L\sigma_x\sqrt{\lambda/(2-\lambda)\,n}$. The rate of convergence to the asymptotic values depends critically on λ, with the convergence being much slower for small λ.

To illustrate the effect of time-varying limits, consider the following example used by Lucas and Crosier (1982) to show the effect of the FIR feature on a CUSUM chart. For the example we shall assume $\mu_x = 0$, $\sigma_x = 1$, and $L = 3$. The raw data are given by $\bar{x}_t$ in Table 1 and represents an initial out-of-control situation. Table 1 also gives z_t derived from Equation (1) and the time-varying control limits derived from Equation (3), with $\lambda = 0.1$.

Figure 1 shows the resulting EWMA charts for different values of λ. The time-varying upper control limit $UCL(t)$ is shown as a solid line, whereas the asymptotic control limit UCL is shown as a dashed line. Figure 1 shows only the upper control limits

TABLE 1. Simple EWMA Example with $\lambda = 0.1$

Sample Number	$\bar{x}_t$	z_t	$UCL(t)$
0	–	0.00	0.00
1	0.8	0.08	0.30
2	1.9	0.26	0.40
3	1.4	0.38	0.47
4	2.0	0.54	0.52
5	1.1	0.59	0.56
6	0.7	0.61	0.58
7	2.6	0.80	0.60
8	0.5	0.77	0.62
9	1.2	0.82	0.63

to aid display; normally both upper and lower control limits are shown. The number of observations needed to generate an out-of-control signal depends on both the value of λ and whether time-varying control limits are used. When λ equals .05, .1, or .25, an EWMA chart with time-varying control limits signals after only four observations, whereas an EWMA chart with asymptotic limits will not generate a signal until observation seven for $\lambda = .1$ and .25 or until observation nine for $\lambda = .05$. When $\lambda = 0.5$, the time-varying control limit quickly converges to the asymptotic value (and thus has little effect) and a signal occurs after seven observations using either $UCL(t)$ or UCL as the control limit.

As can be seen in Figure 1, using asymptotic control limits rather than time-varying limits makes the EWMA chart much less sensitive to process shifts in the first few observations. This could be a significant problem if a large shift occurs early or if the process is not properly reset after an out-of-control condition.

Run Length Properties of EWMA Charts with Time-varying Control Limits

In this section, the run length properties of EWMA charts with time-varying control limits, such as the ARL, are compared with the run length properties of EWMA charts with asymptotic control limits. As will be shown, while the process is in-control, the ARL's of EWMA control charts with time-varying control limits are nearly identical to the ARL's of traditional EWMA charts with asymptotic control limits. However, when the initial process level is out of control, the ARL of the two charts may differ substantially depending on the value of λ.

It is important to quantify the effect of using time-varying control limits since EWMA control charts are usually designed to have given ARL's under certain operating conditions. For an EWMA chart, the design parameters include λ and L. However, since the

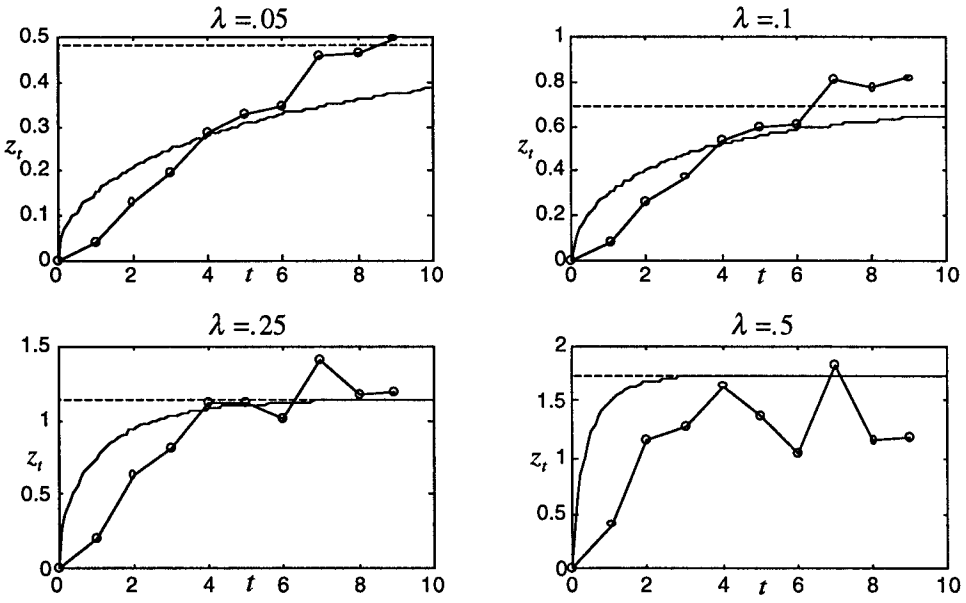

FIGURE 1. Plot of EWMA Control Charts with Time-Varying Control Limits. Dashed Lines Show the Asymptotic Control Limits, Solid Lines Show the Time-Varying Control Limits Generated by Equation (3), and Circles Represent the EWMA Values.

TABLE 2. ARL's for Two-Sided EWMA Charts

	Zero States, $L = 3.0$							
	Asymptotic Control Limits				Time-Varying Control Limits			
$\mu_x/\sigma_{\bar{x}}$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$
0.00	398.0	503.0	842.0	1379.0	382.0	500.0	828.0	1353.0
0.25	209.0	171.0	145.0	135.0	207.0	170.0	140.0	127.0
0.50	75.4	48.5	37.4	37.4	74.5	47.6	34.5	32.5
0.75	31.5	20.2	17.9	20.0	30.8	19.5	15.3	15.6
1.00	15.7	11.2	11.4	13.5	15.2	10.2	9.1	9.0
1.50	6.1	5.5	6.6	8.3	5.7	4.7	4.5	4.5
2.00	3.5	3.6	4.7	6.0	3.2	2.9	2.8	2.8
2.50	2.4	2.8	3.7	4.8	2.2	2.1	2.0	2.0
3.00	1.9	2.3	3.1	4.0	1.6	1.6	1.6	1.6
3.50	1.5	2.0	2.6	3.4	1.3	1.3	1.3	1.3
4.00	1.3	1.7	2.3	3.0	1.2	1.2	1.2	1.1

time-varying control limits converge to the constant asymptotic values as time increases, for process shifts that occur later in time the two charts will have similar run length properties. As a result, EWMA control charts with time-varying control limits can be designed in the same manner as EWMA charts with asymptotic limits. See Crowder (1987) for guidelines.

The run length properties of EWMA control charts with asymptotic control limits were determined by Crowder (1987) using an integral equation approach. Unfortunately, this integral equation solution approach is not applicable for EWMA charts with time-varying control limits. However, the run length properties of the EWMA chart with time-varying control limits can be approximated using a non-homogeneous discrete Markov chain. Using a Markov chain, the feasible state space is approximated by dividing it into distinct, discrete states, and the probability of moving from any one state to any other state for each time period is determined. By using a greater number of distinct states, the approximation of the run length properties can be made more precise. A detailed explanation of the solution procedure is given in the Appendix.

The effect of time-varying control limits on the ARL is illustrated in Table 2 and Figure 2. The results were derived using $L = 3.0$ as the control limit constant and, without loss of generality, assuming an in-control mean of zero and a standard deviation of unity. Table 2 reports process shifts in units of $\mu_x/\sigma_{\bar{x}}$. The Appendix outlines how these values were obtained. In Figure 2, the horizontal axis gives the initial true process mean in $\sigma_{\bar{x}}$ units, the standard deviation of the sample mean. The results are given only for positive shifts, but since the problem is symmetric the same pattern is observed for negative shifts. ARL values for the asymptotic case are taken from Crowder (1987), while ARL results for EWMA charts with time-varying control limits are determined using the methodology presented in the Appendix. Figure 2 shows that the effect of using time-varying control limits on the ARL of the EWMA chart is substantial when the process is not initially in control, especially when λ is small. Figure 2 uses log(ARL) to improve the visual comparison.

For example, assume that the initial process mean value is $2.0\sigma_{\bar{x}}$ units greater than the in-control value

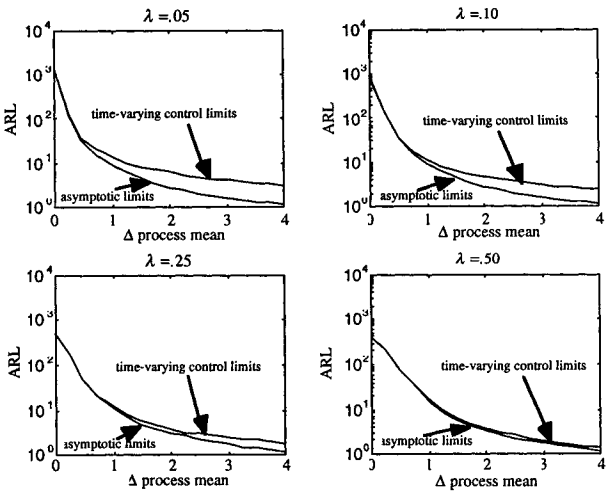

FIGURE 2. Plot of the ARL's for EWMA Charts with Time-Varying and Asymptotic Control Limits.

TABLE 3. Standard Deviation of the Run Length for Two-Sided EWMA Charts

	Zero States, $L = 3.0$							
	Asymptotic Control Limits				Time-Varying Control Limits			
$\mu_x/\sigma_{\bar{x}}$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$
0.00	396.0	499.0	833.0	1363.0	396.0	499.0	834.0	1364.0
0.25	207.0	167.0	133.0	113.0	207.0	167.0	133.0	113.0
0.50	73.2	43.8	27.6	22.0	73.2	43.8	28.0	23.0
0.75	29.3	15.9	10.2	8.8	29.2	16.0	10.6	9.7
1.00	13.6	7.5	5.3	4.9	13.6	7.4	5.7	5.5
1.50	4.3	2.7	2.3	2.3	4.2	2.8	2.5	2.6
2.00	1.9	1.4	1.3	1.4	1.9	1.5	1.5	1.5
2.50	1.1	0.9	0.9	1.0	1.1	1.0	1.0	1.0
3.00	0.8	0.6	0.7	0.8	0.8	0.7	0.7	0.7
3.50	0.6	0.5	0.6	0.6	0.5	0.5	0.5	0.5
4.00	0.5	0.5	0.5	0.5	0.4	0.4	0.4	0.4

used to set up the EWMA chart. Then using $\lambda = .05$, the ARL of the EWMA chart with time-varying control limits is 2.8 which is much shorter than the ARL of 6.0 required for an EWMA chart using asymptotic control limits. The effect of the time-varying control limits, however, has very little influence on the in-control run length as shown in Figure 2 and by the $\mu_x/\sigma_{\bar{x}} = 0.0$ row in Table 2. Time-varying control limits are recommended for all EWMA charts, since their performance will be substantially better than asymptotic limit EWMA charts when the process is fairly likely to start out of control.

Standard deviation values for the asymptotic EWMA control charts are also given in Crowder (1987). Table 3 reproduces the Crowder results and gives the standard deviation values for the time-varying case also calculated using the time non-homogenous Markov chain methodology presented in the Appendix. Table 3 points out that Crowder's value for $\mu_x/\sigma_{\bar{x}}$ and $\lambda = .05$ is incorrectly given as 1,623.50 when the true value is 1,363.0. Table 3 shows that the standard deviation of the run lengths are nearly identical for the asymptotic EWMA control chart and the EWMA control chart with time-varying control limits.

It is also of interest to examine how the distribution of the run length of an EWMA control chart changes when time-varying control limits are adopted. The run length distribution can be determined using Equation (A1) given in the Appendix. Figures 3 and 4 show the run length distributions for EWMA charts with time-varying control limits and asymptotic control limits when the initial process is in control and shifted one $\sigma_{\bar{x}}$ unit.

Figure 3 shows an initial spike in the run length probability density for the EWMA chart with time-varying control limits, with the two probability densities nearly converging for long run lengths. This greater probability of a short run length is undesirable since the initial process state is in control and since we would like the run length to be very long. However, since the probabilities involved are still very small, this spike has a corresponding small influence on the ARL. The size of this initial spike in the run length probability density function depends on L, with smaller L leading to larger spikes. In Figure 4, by contrast, the bulk of the probability density for the two cases is quite different, and the ARL under the time-varying control limits will be substantially shorter. Of course, given an initial out-of-control state, a short ARL is desirable.

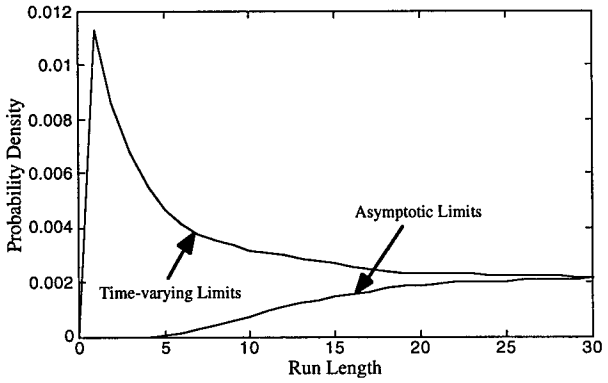

FIGURE 3. In-Control Run Length Distribution of EWMA Charts with Time-Varying and Asymptotic Control Limits with $\lambda = 0.05$ and $L = 2.587$.

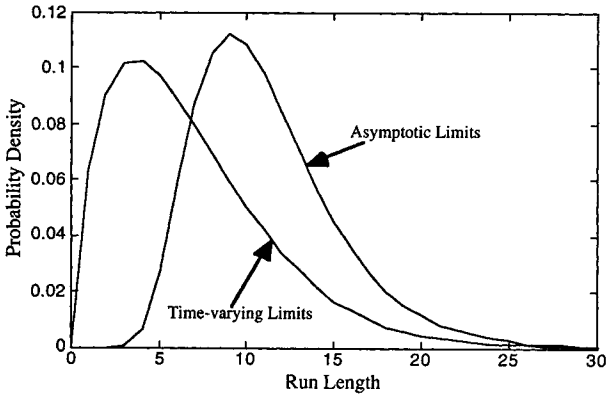

FIGURE 4. Out-of-Control Run Length Distributions of EWMA Charts with Time-Varying and Asymptotic Control Limits with $\lambda = 0.05$, $L = 2.587$, and Initial Mean Shift of One Standard Deviation Unit.

Comparing the run length distribution plots shown in Figures 3 and 4 with similar plots for CUSUM and FIR-CUSUM in Lucas and Crosier (1982) and for FIR-EWMA charts in Lucas and Saccucci (1990) suggests that the effect of the time-varying limits is similar to that achieved with the FIR feature. The effect of the time-varying limits appears less pronounced than the FIR-CUSUM, which suggests that an additional narrowing of the time-varying control limits for small values of t may be appropriate to make the EWMA chart even more sensitive to start-up quality problems.

EWMA Control Charts with FIR

EWMA charts with time-varying control limits were shown in the previous section to have properties similar to the FIR feature when compared with asymptotic EWMA charts. However, using time-varying control limits is not the same as the FIR feature for CUSUM charts since the adjustment of the control limits only corrects the control limits to take into account the time-dependent nature of the EWMA statistic given by Equation (1).

A few authors have suggested adaptations to the EWMA scheme to build in a true FIR feature. As discussed in the introduction, Lucas and Saccucci (1990) suggested the use of two one-sided EWMA charts with initial states different than zero to create a two-sided EWMA chart that reacts quickly. Rhoads, Montgomery, and Mastrangelo (1996) adapt the Lucas and Saccucci (1990) approach by allowing each one-sided chart to have time-varying control limits. Both these methods have the desired effect of making the chart more sensitive to start-up quality problems, but are rather awkward since they require the simultaneous use of two EWMA charts to accomplish the task previously achieved with just one chart.

Here, we suggest an approach that retains the simplicity of a single control chart. To give EWMA

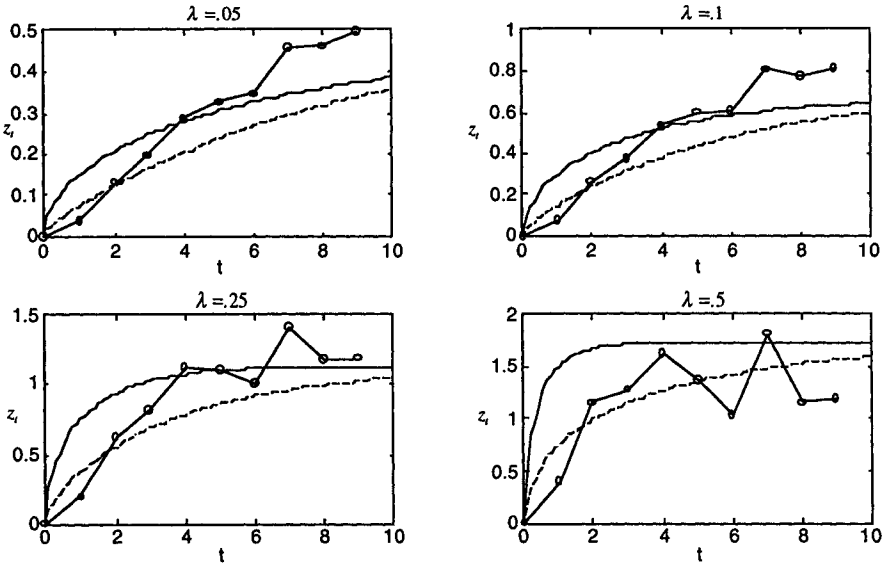

FIGURE 5. EWMA Charts with Time-Varying Control Limits. Dashed Lines Show the FIR Time-Varying Control Limits From Equation (4), Solid Lines Show the Time-Varying Control Limits Generated by Equation (3), and Circles Represent the EWMA Values.

TABLE 4. ARL Results for Different FIR Proportions

			$\lambda = 0.25, L = 3$			
$\mu_x/\sigma_{\bar{x}}$	$f = 0.4$	$f = 0.5$	$f = 0.6$	$f = 0.7$	$f = 0.8$	$f = 1$
0.00	349.4	368.0	429.8	464.5	480.7	500.5
0.25	111.7	118.4	139.7	153.1	159.7	169.8
0.50	26.8	29.0	35.0	39.8	42.3	47.6
0.75	9.2	10.1	12.5	14.8	16.1	19.5
1.00	4.4	4.8	6.1	7.5	8.2	10.6
1.50	2.0	2.1	2.7	3.2	3.7	5.0
2.00	1.4	1.4	1.7	2.0	2.3	3.1
3.00	1.1	1.1	1.1	1.2	1.4	1.7
4.00	1.0	1.0	1.0	1.0	1.1	1.2
			$\lambda = 0.10, L = 3$			
$\mu_x/\sigma_{\bar{x}}$	$f = 0.4$	$f = 0.5$	$f = 0.6$	$f = 0.7$	$f = 0.8$	$f = 1$
0.00	515.6	613.8	737.2	795.0	805.9	832.1
0.25	83.1	99.2	120.8	132.1	133.7	141.1
0.50	18.5	22.1	27.6	31.2	31.6	35.2
0.75	7.3	8.8	11.2	13.3	13.6	16.0
1.00	3.8	4.6	6.1	7.5	7.8	9.6
1.50	1.7	2.1	2.7	3.4	3.8	4.8
2.00	1.3	1.4	1.8	2.1	2.4	3.0
3.00	1.0	1.0	1.2	1.3	1.4	1.7
4.00	1.0	1.0	1.0	1.1	1.1	1.2

charts with time-varying control limits a FIR feature, the control limits are narrowed further for the first few sample points. This approach is easily implemented since the control limits are already time-varying. Since the time-varying control limits exponentially approach the asymptotic limits, it is reasonable to use an exponentially decreasing adjustment to further narrow the limits. Let

$$FIR_{adj} = 1 - (1 - f)^{1+a(t-1)}. \quad (4)$$

With this setup, the FIR adjustment makes the control limits for the first sample point ($t = 1$) a proportion, f, of the original distance from the starting value. The effect of the FIR adjustment decreases with time to ensure that the long-term run length properties of the EWMA chart will be virtually unchanged. A reasonable setup would be to set the adjustment parameter, a, so that the FIR adjustment has very little effect after observation 20, say, FIR_{adj} at observation 20 is .99. This should be sufficient to allow the detection of quality problems in the start-up. This idea implies that we should set $a = (-2/\log(1 - f) - 1)/19$.

For example, using $f = 0.5$ yields $a = 0.3$. Using this adjustment factor and Equation (3), the FIR-EWMA control limits are

$$\mu_x \pm L\sigma_x \left(1 - (1 - f)^{1+a(t-1)}\right) \sqrt{\frac{\lambda[1 - (1 - \lambda)^{2t}]}{(2 - \lambda)n}}. \quad (5)$$

The control limits given by Equation (5) are time-varying; thus, the run length properties of the proposed FIR-EWMA chart can also be determined using the non-homogeneous Markov chain methodology presented in the Appendix.

Figure 5 shows the effect of using the limits in Equation (5) with $f = 0.5$ and $a = 0.3$ in the example initially discussed in the second section of this paper and illustrated in Figure 1. In Figure 5, the advantage of the additional narrowing of the control limits in detecting start-up quality problems is clearly demonstrated. For all the different values of λ, the FIR-EWMA control chart signals in just two observations. This is a substantial improvement over the run lengths obtained with only the time-varying control limits, especially for large values of λ.

To explore the effect of different levels of FIR, Table 4 gives ARL results for different levels of f. From

TABLE 5. ARL Comparison of EWMA Charts with FIR

$\mu_x/\sigma_{\bar{x}}$	$\lambda = 0.25$			$\lambda = 0.10$		
	LFIR $L = 2.81$	RFIR $L = 3.00$	FIR $L = 3.07$	LFIR $L = 2.81$	RFIR $L = 3.00$	FIR $L = 2.91$
0.0	483.0	452.0	468.0	463.0	466.0	459.0
0.5	42.1	39.3	33.5	24.2	22.2	19.6
1.0	8.5	7.6	5.2	6.9	5.4	4.5
1.5	3.9	3.2	2.3	3.7	2.4	2.1
2.0	2.5	1.9	1.5	2.7	1.6	1.4
3.0	1.5	1.1	1.1	1.8	1.1	1.1
4.0	1.1	1.0	1.0	1.3	1.0	1.0
$\mu_x/\sigma_{\bar{x}}$	$\lambda = 0.05$			$\lambda = 0.03$		
	LFIR $L = 2.62$	RFIR $L = 2.72$	FIR $L = 2.69$	LFIR $L = 2.44$	RFIR $L = 2.54$	FIR $L = 2.55$
0.0	421.0	417.0	419.0	383.0	384.0	391.0
0.5	19.7	17.0	16.5	18.6	14.9	13.8
1.0	7.0	4.4	4.2	7.4	3.9	3.6
1.5	4.1	2.2	2.0	4.6	2.0	1.8
2.0	3.1	1.5	1.4	3.4	1.4	1.3
3.0	2.1	1.1	1.1	2.4	1.1	1.0
4.0	1.7	1.0	1.0	1.9	1.0	1.0

these results it is clear that to derive a substantial benefit from the FIR feature, the narrowing of the control limits should also be substantial, say, corresponding to $f = 0.5$. Using $f = 0.5$ corresponds to adjusting the time-varying limits by a factor of one-half for the first time period, as shown in Figure 5, and is an attractive choice because it mimics the 50% head start typically suggested for FIR-CUSUM charts.

Table 5 compares the ARL's of the Lucas and Saccucci (1990) FIR-EWMA chart, denoted LFIR; the Rhoads, Montgomery, and Mastrangelo (1996) FIR-EWMA chart, denoted RFIR; and a FIR-EWMA chart with adjusted time-varying control limits given by Equation (4). The results for the LFIR and the RFIR are taken from simulation results published in Rhoads, Montgomery, and Mastrangelo (1996). The run length results for the proposed FIR-EWMA chart were approximated using the methodology described in the Appendix. For all the FIR-EWMA charts, L has been adjusted so that, in control, all methods have approximately the same ARL.

The results in Table 5 suggest that the proposed FIR-EWMA chart is superior to the previous approaches. For example, with $\lambda = 0.1$ and a mean shift of one standard deviation unit, the proposed FIR-EWMA chart requires on average only 4.5 observations to signal, while the LFIR-EWMA chart and the RFIR-EWMA chart require 6.9 and 5.4 observations, respectively. The reduction in out-of-control ARL's appears to be greatest when λ is not small. In addition to the benefit of better run length properties, the EWMA charts with time-varying control limits also provide two-sided protection from start-up quality problems through only a single control chart. This is a major advantage from an implementation perspective. It should be noted that the FIR-EWMA chart requires larger values of L than the traditional EWMA chart. As a result, if the process shift does not occur near start-up, the FIR-EWMA chart will actually have a slightly longer ARL than the traditional EWMA chart.

Summary

This article derives the run length properties for EWMA control charts with time-varying control limits. Since the variance of the EWMA test statistic is a function of time, time-varying control limits result in improved process shift detection capabilities if the process is initially out of control or if it goes out of control quickly. The magnitude of the benefit of using time-varying control limits over traditional asymptotic limits depends on the EWMA constant

λ and the size of the initial process shift. Results are presented that quantify the difference for an EWMA chart designed to monitor the process mean. In general, time-varying control limits are useful if λ is small, say, less than 0.3.

Not surprisingly, the results show that EWMA charts with time-varying control limits have shorter ARL's than EWMA charts with asymptotic control limits for start-up quality problems. The effect for out-of-control mean values is more pronounced than for the in-control case, especially for large process shifts. As a result, EWMA control charts with time-varying control limits are appropriate in all situations where the initial quality level is suspect.

In situations where at the start of process monitoring there is a good chance the process is out of control, further narrowing of the time-varying control limits is shown to provide an additional FIR benefit. Adjusting the control limits to start at half the regular value and then exponentially approach the regular time-varying limits for 20 observations is shown to be a better approach to creating a FIR-EWMA chart than previously suggested approaches. The proposed approach has the additional benefit of retaining the EWMA chart's ability to allow two-sided detection of problems with a single chart.

Appendix

In this Appendix, approximations for the distribution, expected value, and variance of the run length of EWMA charts with time-varying control limits are derived. The solution procedure utilizes a nonhomogenous Markov chain with g distinct states. In the solution, the state space between the control limits is divided into $g-1$ distinct discrete states, and the out-of-control condition corresponds to the g^{th} state. The different states are defined as

$$\begin{aligned}\mathbf{s} &= (s_1, s_2, \ldots, s_{g-1}) \\ &= (LCL + w, LCL + 2w, \\ &\quad \ldots, UCL - 2w, UCL - w)\ ,\end{aligned}$$

where $w = (UCL - LCL)/g$ and UCL and LCL are the asymptotic control limits as given by setting $t = \infty$ in Equation (3). As g increases the approximation improves.

Assume that the transition probability matrix for time period t is given by

$$\begin{aligned}\mathbf{P}_t &= \begin{bmatrix} {}_tp_{11} & {}_tp_{12} & \cdots & {}_tp_{1g} \\ {}_tp_{21} & \cdots & \cdots & {}_tp_{2g} \\ \vdots & \vdots & \vdots & \vdots \\ {}_tp_{g1} & \cdots & \cdots & {}_tp_{gg} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{R}_t & (\mathbf{I} - \mathbf{R}_t)\mathbf{1} \\ \mathbf{0} & 1 \end{bmatrix},\end{aligned}$$

where $\mathbf{I}$ is the $g \times g$ identity matrix, $\mathbf{1}$ is a $g \times 1$ column vector of ones, and ${}_tp_{ij}$ equals the transition probability from state s_i to state s_j for time period t. The last row and column correspond to the absorbing state that represents an out-of-control signal. The $\mathbf{R}_t$ matrix equals the transition probability matrix with the row and column that correspond to the absorbing (out-of-control) state deleted. $\mathbf{R}_t$ will be used to derive the run length properties of the EWMA control chart with time-varying control limits.

Since the time-varying control limits in Equation (3) asymptotically approach constant values, the ${}_tp_{ij}$'s converge to ${}_\infty p_{ij}$'s, and $\mathbf{R}_t$ converges to the infinite time transition matrix, $\mathbf{R}_\infty$, as $t \to \infty$. The values for ${}_\infty p_{ij}$ can be determined by making some process assumptions. Assuming a normal model with $X_i \sim N\left(\mu_x, \sigma_x^2\right)$ and given the current EWMA value, the distribution of the future EWMA value z_{t+1} is $N(\lambda\mu_x + (1-\lambda)z_t, \lambda^2\sigma_x^2)$. Thus, the infinite time transition probabilities are

$$\begin{aligned}{}_\infty p_{ij} &= \Pr\left(s_j - \frac{w}{2} < z < s_j + \frac{w}{2}\right) \\ &\qquad \text{for } j = 1, 2, \ldots, g-1 \quad \text{and} \\ {}_\infty p_{ig} &= \Pr\left(z > s_{g-1} + \frac{w}{2}\right) + \Pr\left(z < s_1 - \frac{w}{2}\right),\end{aligned}$$

where $z \sim N(\lambda\mu + (1-\lambda)s_i, \lambda^2\sigma^2)$. These values can be easily calculated to determine $\mathbf{P}_\infty$ and $\mathbf{R}_\infty$.

The time-dependent transition matrices $\mathbf{R}_t$ can be determined from $\mathbf{R}_\infty$ by changing the transition probabilities that lead to an earlier signal. For both $\mathbf{R}_t$ and $\mathbf{R}_\infty$, the rows represent the starting values and the columns represent the ending values for each transition probability. For each value of t, the appropriate rows and columns are identified by comparing the time-varying control limits with the states in the state space. In other words, to determine $\mathbf{R}_t$, the first $f_1(t)$ and last $f_2(t)$ rows and columns of $\mathbf{R}_\infty$ are set to zero vectors, where $f_1(t)$ equals the largest integer for which $s_{f_1} - w/2 \le LCL(t)$ and where $f_2(t)$ is the smallest integer for which $s_{f_2} + w/2 \ge UCL(t)$. In an attempt to consistently yield run length values less than the true value, any state whose transition probability is at all affected by the changing control

limit is set to zero. A state s_i is affected if the time-varying control limit is either closer to zero than s_i or within $w/2$ of s_i. Using this procedure, estimates for $\mathbf{R}_1$, $\mathbf{R}_2$, ... are obtained.

Determining the expected run length and the variance of the run length can now proceed using the matrices $\mathbf{R}_t$. Letting RL equal the run length of the EWMA chart conditional on the starting state, s_i, we have

$$\Pr(RL \le t \mid s_i) = \left(\mathbf{I} - \prod_{i=1}^{t} \mathbf{R}_i\right)\mathbf{1} \text{ and}$$
$$\Pr(RL = t \mid s_i) = \left(\prod_{i=1}^{t-1} \mathbf{R}_i - \prod_{i=1}^{t} \mathbf{R}_i\right)\mathbf{1} \text{ for } t \ge 1. \tag{A1}$$

Thus,

$$E(RL \mid s_i) = \sum_{t=1}^{\infty} t \Pr(RL = t) = \sum_{t=1}^{\infty} \left(\prod_{s=1}^{t} \mathbf{R}_s \mathbf{1}\right). \tag{A2}$$

Similarly, the variance of the run length is

$$\mathrm{Var}(RL \mid s_i) = \mathbf{1} + \sum_{t=1}^{\infty} \left[(2t+1)\left(\prod_{s=1}^{t} \mathbf{R}_s \mathbf{1}\right)\right]. \tag{A3}$$

Equations (A2) and (A3) yield $g \times 1$ vectors that correspond to the ARL and variance, respectively, from initial state, s_i. The values that correspond to the starting value with $z_0 = \overline{\overline{X}}$ are easily found. Assuming that the control limits are symmetric about $\overline{\overline{X}}$, the corresponding state is $s_{g/2}$.

Equations (A2) and (A3) give the moments of the run length in terms of infinite sums that converge for large t. These expressions can be simplified in this case since the control limits converge asymptotically; thus, the transition probability matrices $\mathbf{R}_t$ also converge to $\mathbf{R}_\infty$ as t increases. Replacing all $\mathbf{R}_t$ matrices for large t values with $\mathbf{R}_\infty$, the infinite sums in Equations (A2) and (A3) can be written as

$$E(RL \mid s_i) = \sum_{t=1}^{t_{\max}-1} \left(\prod_{s=1}^{t} \mathbf{R}_s \mathbf{1}\right) + \left(\prod_{s=1}^{t} \mathbf{R}_s\right)\left[\mathbf{I} - \mathbf{R}_\infty\right]^{-1}\mathbf{1} \tag{A4}$$

and

$$\mathrm{Var}(RL \mid s_i) = \mathbf{1} + \sum_{t=1}^{t_{\max}-1} \left[(2t+1)\left(\prod_{s=1}^{t} \mathbf{R}_s \mathbf{1}\right)\right] + (2t_{\max}+1)\left(\prod_{s=1}^{t_{\max}} \mathbf{R}_s\right)(\mathbf{I} - \mathbf{R}_\infty)^{-1}\mathbf{1} + 2\left(\prod_{s=1}^{t_{\max}} \mathbf{R}_s\right)\mathbf{R}_\infty(\mathbf{I} - \mathbf{R}_\infty)^{-2}\mathbf{1}, \tag{A5}$$

where $t_{\max}$ equals the number of time periods for which different transition probability matrices are used. For the computations, $t_{\max}$ was chosen based on λ and g so that the matrix $\mathbf{R}_{t_{\max}}$ is indistinguishable from $\mathbf{R}_\infty$. In this way, increasing $t_{\max}$ further will have no influence on the solution accuracy. The minimum value for $t_{\max}$ is derived by realizing that if the time-varying control limits at time $t_{\max}$ differ from the asymptotic limits by less than $w/2$, then the matrix $\mathbf{R}_{t_{\max}}$ is the same as $\mathbf{R}_\infty$. Solving $UCL - UCL(t) \le w/2$ and $LCL - LCL(t) \le w/2$ for the minimum t value yields $t_{\max}$ as the smallest integer larger than

$$\frac{\log\left(\dfrac{12nw(2-\lambda)\sigma\sqrt{\lambda/n(2-\lambda)} - w}{36\lambda\sigma^2}\right)}{2\log(1-\lambda)}.$$

For computational efficiency and accuracy, $E(RL \mid s_i)$ and $\mathrm{Var}(RL \mid s_i)$ are determined using Gaussian elimination rather than by finding the matrix inverse directly as suggested by Equations (A4) and (A5).

In general, as g increases the $E(RL \mid s_i)$ and $\mathrm{Var}(RL \mid s_i)$ values obtained through Equations (A4) and (A5) increase and more closely approximate the true values. The values increase because the procedure always underestimates the true run length. The run lengths are underestimated for two reasons: first, the absorbing boundaries for $\mathbf{R}_\infty$ are narrower than the control limits since they are set at $LCL + w/2$ and $UCL - w/2$; and second, for $\mathbf{R}_t$ the absorbing probabilities are conservatively calculated since all states even marginally affected by the control limit are assumed to lead to absorption.

The advantage of consistently underestimating the run lengths of the EWMA chart are that we can use the rate of increase to estimate the true values. The values shown in the Tables 2, 3, and A1 were derived by estimating the true value $E(RL \mid s_i)_{g=\infty}$ based on fitting the model $E(RL \mid s_i) = E(RL \mid s_i)_{g=\infty} + B/g + C/g^2$ derived using the results gen-

TABLE A1. ARL's for Time-Varying Control Limits for EWMA Charts

Zero States

	$L = 3.50$				$L = 3.25$			
$\mu_x/\sigma_{\bar{x}}$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$
0.00	2226.0	2638.0	4101.0	6442.0	910.6	1112.0	1789.0	2852.0
0.25	950.3	624.5	382.3	270.7	431.0	315.3	225.9	180.6
0.50	266.9	122.6	62.8	48.8	137.0	74.3	46.5	39.4
0.75	88.3	38.1	23.7	21.3	50.9	26.8	19.4	17.9
1.00	35.6	17.2	13.2	12.3	22.9	13.4	11.3	10.6
1.50	10.0	6.8	6.3	6.0	7.6	5.8	5.5	5.2
2.00	4.7	4.0	3.9	3.5	3.9	3.5	3.4	3.0
3.00	2.1	2.1	2.0	1.6	1.9	1.9	1.8	1.5
4.00	1.4	1.4	1.3	1.1	1.3	1.3	1.2	1.1
	$L = 3.00$				$L = 2.75$			
$\mu_x/\sigma_{\bar{x}}$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$
0.00	396.6	500.5	832.1	1341.0	184.0	240.0	410.0	664.0
0.25	208.0	169.8	141.1	125.0	107.0	96.9	92.3	88.7
0.50	75.0	47.6	35.2	31.8	43.6	32.0	27.1	25.6
0.75	31.1	19.5	16.0	15.1	20.1	14.7	13.2	12.6
1.00	15.5	10.6	9.6	9.1	10.9	8.5	8.1	7.7
1.50	5.9	5.0	4.8	4.5	4.7	4.3	4.2	3.8
2.00	3.3	3.1	3.0	2.6	2.8	2.7	2.6	2.3
3.00	1.7	1.7	1.7	1.4	1.5	1.5	1.5	1.3
4.00	1.2	1.2	1.2	1.0	1.1	1.1	1.1	1.0
	$L = 2.50$				$L = 2.25$			
$\mu_x/\sigma_{\bar{x}}$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$	$\lambda = 0.50$	$\lambda = 0.25$	$\lambda = 0.10$	$\lambda = 0.05$
.00	90.5	122.0	213.0	343.0	47.2	65.2	115.2	182.1
.25	57.8	58.2	62.3	63.7	33.0	36.4	42.6	45.8
.50	26.8	22.4	1.1	20.5	17.3	16.1	16.4	16.2
.75	13.7	11.3	10.8	10.4	9.7	8.8	8.8	8.5
1.00	8.0	6.9	6.8	6.4	6.1	5.6	5.7	5.3
1.50	3.9	3.6	3.6	3.2	3.2	3.1	3.1	2.7
2.00	2.4	2.4	2.3	2.0	2.1	2.1	2.0	1.7
3.00	1.4	1.4	1.4	1.2	1.3	1.3	1.3	1.1
4.00	1.1	1.1	1.1	1.0	1.1	1.1	1.0	1.0

erated with $g = 50$, 100, and 150. Verification of this approach using simulation suggests that our results differ from the true value by less than 1%, except for very large process shifts when the ARL is near unity. For very large shifts, the values in $\mathbf{R}_t$ become smaller, and calculations required to derive $E(RL \mid s_i)$ become more prone to rounding error. As a result, for large shifts the $E(RL \mid s_i)$ estimate may not increase as g increases. If this occurs, we use the largest obtained $E(RL \mid s_i)$ as an estimate of the true $E(RL \mid s_i)_{g=\infty}$, noting that the estimate may be off by as much as 10%. A similar problem is also reported in Lucas and Crosier (1982). However, in our case, for comparison purposes, the results are adequate.

To provide more details, Table A1 gives the ARL values for EWMA charts with time-varying control limits for some different values of L. Results are derived for the two-sided case, but the methodology can be easily adapted for one-sided case EWMA charts defined as $z_t = \max(\lambda\, \overline{x_t} + (1-\lambda) z_{t-1}, z_0)$. In addition, the examples provided assume the distribution of the observed process parameter is normal. However, similar results are easily derived for other underlying distributions.

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Manufacturing Research Council of Ontario. The author would also like to thank two anonymous referees for their helpful comments.

References

BOX, G. E. P.; JENKINS, G. M.; and MACGREGOR, J. F. (1974). "Some Recent Advances in Forecasting and Control". *Applied Statistics* 23, pp. 158–179.

CROWDER, S. V. (1987). "Run-Length Distributions of EWMA Charts". *Technometrics* 29, pp. 401–407.

LUCAS, J. M. and CROSIER, R. B. (1982). "Fast Initial Response for CUSUM Quality Control Schemes". *Technometrics* 24, pp. 199–205.

LUCAS, J. M. and SACCUCCI, M. S. (1990). "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements" (with discussion). *Technometrics* 32, pp. 1–29.

MONTGOMERY, D. C. (1991). *Introduction to Statistical Quality Control,* 2nd ed. John Wiley & Sons, New York, NY.

ROBERTS, S. W. (1959). "Control Chart Tests Based on Geometric Moving Averages". *Technometrics* 1, pp. 239–250.

RHOADS, T. R.; MONTGOMERY, D. C.; and MASTRANGELO, C. M. (1996). "Fast Initial Response Scheme for the Exponentially Weighted Moving Average Control Chart". *Quality Engineering* 9, pp. 317–327.

Key Words: *Average Run Length, Cumulative Sum, Exponentially Weighted Moving Average, Fast Initial Response, Markov Chains.*