The Exponentially Weighted Moving Variance

J. F. MACGREGOR

McMaster University, Hamilton, Ontario L8S 4L7

T. J. HARRIS

Queen's University, Kingston, Ontario K7L 3N6

Exponentially weighted moving variance (EWMV) and exponentially weighted mean squared deviation (EWMS) charts are proposed as ways of monitoring various types of continuous process variation. They are particularly useful for augmenting control charts on individual observations where no estimate of variability is available from replicates and for providing measures of process variance when the observations are autocorrelated.

Introduction

STATISTICAL process control charts such as the Shewhart, CUSUM, and EWMA have been used extensively to monitor product quality and detect special events that may be indicators of out-of-control situations. In the discrete parts manufacturing industries the most common charts are the $\overline{X}$ and R charts. In these charts a random sample of m items is repeatedly taken and either the sample average ($\overline{X}$) or the range (R) of one or more properties are sequentially plotted together with appropriate control limits (see, e.g., Montgomery (1991) or John (1990)). Alternatively, a CUSUM or EWMA chart on $\overline{X}$ could replace the Shewhart chart.

However, since in the continuous process industries the sampling and analysis costs are often quite high, it is extremely common to plot charts on individual observations ("individual charts"). Furthermore, the data values are usually autocorrelated. As a result, there are considerable incentives for achieving better control by taking one observation on a more frequent basis than taking m observations less frequently. An apparent disadvantage of individual observation charts is that one loses the information on the process variance that can be computed from the variability among the m samples. Therefore, it is common to try to estimate the process variability by the moving range (see Montgomery (1991)) or the mean square of successive differences (see Nelson (1980)). In this paper, exponentially weighted mean squared deviation (EWMS) and exponentially weighted moving variance (EWMV) charts are proposed as ways of monitoring various types of continuous process variation.

The use of an exponentially weighted mean square to monitor variations was suggested by Wortham and Ringer (1971) and Wortham (1972). Sweet (1986) proposed the exponentially weighted mean absolute deviation and the exponentially weighted moving variance. Control limits for them were derived, in the case of independent observations, by Wortham (1972) and Sweet (1986). Others, such as Montgomery and Mastrangelo (1991), have also advocated the use of exponentially weighted moving variance type statistics. Bauer and Hackl (1978, 1980) investigated moving sums of squares statistics, again for the case of independent observations. Exponentially weighted mean square deviation charts are also a special case of the omnibus EWMA schemes studied by Domangue and Patch (1991). Ng and Case (1989) investigated some average run length (ARL) properties for the various exponentially weighted moving averages and moving ranges, and Crowder and Hamilton (1992) used exponential weighting to smooth $\ell n(s^2)$ obtained from the range statistic in the conventional $\overline{X}$ and R chart.

In this paper, we investigate the ability of the EWMS and EWMV to monitor processes with chang-

Dr. MacGregor is a Professor in the Chemical Engineering Department.

Dr. Harris is currently a Professor and Head of Chemical Engineering.

ing means and variances. Control limits for the EWMS and the EWMV are developed for both the situation where the observations are independent and where they are autocorrelated. Also, different ways of using these statistics to monitor processes with autocorrelated observations are investigated.

A number of articles have been published on testing for variance changes and monitoring variability in different contexts. Hsu (1977) investigated tests for a variance shift in a given set of N independent normal random deviates with a constant mean. Wichern et al. (1976) and Tsay (1988) proposed methods for detecting variance changes in time series models, but they assumed a given length of time series data, and their methods were non-recursive in nature. Trigg and Leach (1967) developed a now widely used approach for adaptive EWMA forecasting based on monitoring the mean absolute deviations of the forecast errors, and Smith (1976) compared several of these approaches.

The Exponentially Weighted Mean Square (EWMS)

Consider the case of normally distributed random variables (Y_k with a constant mean η and variance σ^2) coming available at equispaced time intervals $k = 0, 1, 2, \ldots, n$. Define the exponentially weighted mean squared error at the current time n as

$$S_n^2 = \sum_{k=1}^{n} r(1-r)^{n-k}[Y_k - \eta]^2 + (1-r)^n S_0^2$$
$$= (1-r)S_{n-1}^2 + r[Y_n - \eta]^2 \qquad (1)$$

where r is a weight ($0 < r \leq 1$) that controls the rate of exponential discounting of past data and S_0^2 is an initial estimate of the mean squared error (usually taken to be the historical in-control value). Note that the sum of the weights is given by

$$r\sum_{k=1}^{n}(1-r)^{n-k} + (1-r)^n = 1.$$

The quantity S_n^2/σ^2 is a weighted sum of χ^2 random variables and, as shown in Appendix A, it is approximately distributed as $\chi^2(\nu)/\nu$ where the number of degrees of freedom ν depends upon the exponential weighting parameter r, the correlation structure of the Y_k's, and the degrees of freedom associated with S_0^2. Since $E[\chi^2(\nu)] = \nu$ it follows that $E\left[S_n^2\right] = \sigma^2$ and S_n^2 is an unbiased estimator of σ^2. For independent observations the degrees of freedom are asymptotically given as $\nu = (2-r)/r$. Reasonable choices for the exponential weighting parameter r and the corresponding asymptotic degrees of freedom ν in the estimated mean square are shown in Table 1.

Consider the use of the EWMS (1) to monitor a process where the observations are independent and normally distributed. Under the hypothesis that the process mean is on target (τ) and that the variance is σ_0^2 (where σ_0^2 is obtained from historical data when the process was in control), the upper and lower control limits on S_n^2 are given by the upper $(\alpha/2)100\%$ and $(1-\alpha/2)100\%$ percentiles of the $\sigma_0^2\chi^2(\nu)/\nu$ distribution. In what follows, we will usually plot S_n, the exponentially weighted root mean square (EWRMS) using the appropriate control limits $(C_3\sigma_0, C_4\sigma_0)$ where

$$C_3 = \sqrt{\frac{\chi^2_{1-\alpha/2}(\nu)}{\nu}} \qquad (2a)$$

$$C_4 = \sqrt{\frac{\chi^2_{\alpha/2}(\nu)}{\nu}}\,. \qquad (2b)$$

Values of C_3 and C_4 are given in Table 1 for probability levels $\alpha = 0.01$ and $\alpha = 0.05$.

More accurate control limits than the above two-moment chi squared approximations can be obtained using the four-moment Johnson curves (Johnson

TABLE 1. Control Limit Constants C_3 and C_4 for the EWRMS (S_n) with Exponential Weighting Parameter (r), Corresponding Degrees of Freedom (ν), and Independent Observations

EWRMS parameter	r	0.01	0.02	0.05	0.10	0.20	0.33
degrees of freedom	ν	199	99	39	19	9	5
Box $\alpha = 0.05$	C_3	0.90	0.86	0.78	0.67	0.55	0.41
	C_4	1.10	1.14	1.22	1.32	1.45	1.60
Box $\alpha = 0.01$	C_3	0.86	0.82	0.72	0.66	0.44	0.28
	C_4	1.12	1.18	1.29	1.42	1.62	1.83
Johnson $\alpha = 0.05$	C_3	0.90	0.87	0.79	0.71	0.59	0.48
	C_4	1.10	1.14	1.27	1.33	1.47	1.62
Johnson $\alpha = 0.01$	C_3	0.86	0.83	0.73	0.63	0.50	0.37
	C_4	1.12	1.19	1.31	1.45	1.66	1.89

(1949)). These are discussed in Appendix A. The calculation of the control limits is much more involved. Results for the control limits from this approach are also shown in Table 1.

Figure 1 shows a simulation of this case where the observations are independent and normally distributed about zero with variance $\sigma_0^2 = 1.0$. At time $k = 90$ the variance of the process increases to 2.0. The top plot (a) in Figure 1 is a Shewhart chart of the data with $\pm 3\sigma$ limits. The second plot (b) is an exponentially weighted moving average (EWMA) of the data given by

$$Z_n = (1 - \lambda)Z_{n-1} + \lambda Y_n \qquad (3)$$

where λ is the EWMA weight ($0 < \lambda \leq 1$), taken as $\lambda = 0.2$ in the plot. Its control limits are given by $\pm 3\sigma_0\sqrt{\lambda/(2-\lambda)}$ for independent observations (see Hunter (1986)). The third plot (c) is the EWRMS with initial condition $S_0^2 = 1.0$ and exponential weight $r = 0.05$. This value of r gives $\nu = 39$ degrees of freedom, and using Table 1 for $\alpha = 0.01$, the control limits have been placed at (0.72, 1.29). Note that the EWRMS has picked up the change in variance very quickly and indicates that its new value is approximately 1.4 (or $\sqrt{2}$).

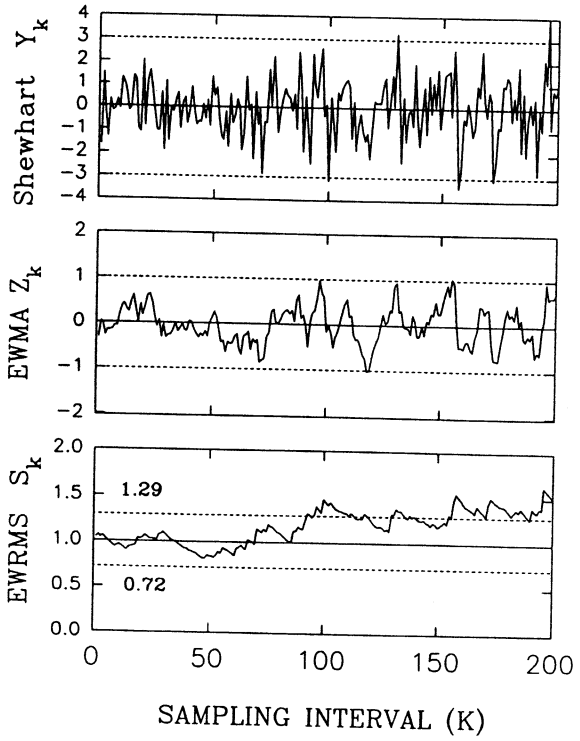

FIGURE 1. Independent Observations with Variance Change from $\sigma^2 = 1.0$ to 2.0 at $k = 90$ with (a) Shewhart Chart on Individuals; (b) EWMA with $\lambda = 0.2$; (c) EWRMS with $r = 0.05$ and $S_0 = 1.0$.

The above control limits were derived based upon the distribution of S_n at time n. The α-risk used in construction of the warning limits is the Type I error if one considers only the current value of the EWRMS statistic, that is, if one performs one test at a time. It is quite common to set the limits for control charts on this basis (see, e.g., John (1990), Bauer and Hackl (1980), or Montgomery (1991)). However, since the test statistics $S_n, S_{n-1}, \ldots$ are not independent, the overall α-risk associated with applying n tests from time $k = 1$ to n will not, in general, be equal to the product of the n instantaneous α-risks. Determining the overall α-risk by finding the joint distribution of all the test statistics $S_n, S_{n-1}, \ldots$ is a difficult task. Alternatively, one could evaluate the overall α-risk by Monte Carlo simulations; a common approach used for designing CUSUM and EWMA charts (see, e.g., Lucas (1976) or Bauer and Hackl (1980)). This correponds to determining the null average run length, ARL(0), for the procedure. The corresponding β-risks for sequential tests are difficult to evaluate since they depend upon the nature of the change to be detected and the point in time at which it occurs. The latter consideration becomes important when one has autocorrelated data. It is common, therefore, to evaluate average run lengths for specific types of changes (usually a step), and for the case of independent observations only.

Consider the situation where independent observations are taken from a process where the mean is on target ($\eta = \tau$) but the variance σ_Y^2 differs by a constant amount from the historical σ_0^2 from time $k = 0$ onwards. The ARL's for the EWRMS corresponding to the control limits in Table 1 are shown in Table 2 for various EWRMS parameter values r and various values of the variance ratio σ_Y^2/σ_0^2. They were computed using simulations based on 4000 replications of the test procedure. The random normal errors were generated using the IMSL library, and, as shown by Harris and Ross (1991), the ARL's determined by simulation with 4000 replicates were within several percent of the ARL's calculated analytically for the case of independent observations. It can be seen from Table 2 that the null average run length (i.e., $\sigma_Y^2/\sigma_0^2 = 1.0$) is extremely sensitive to small changes in the control limits. It is also apparent that for the chosen control limits the EWRMS detects increases in the variance more rapidly than decreases. More complete ARL tables for the EWRMS and comparisons of its performance with other statistics for the

TABLE 2. Average Run Lengths (ARL's) for the EWRMS for Various Values of r and σ_Y^2/σ_0^2, Using the Control Limit Constants in Table 1

Exponential Weighting Parameter (r)	σ_Y^2/σ_0^2	Johnson $\alpha = 0.01$	Johnson $\alpha = 0.05$	Box $\alpha = 0.01$	Box $\alpha = 0.05$
0.33	2.00	18	9	15	9
	1.50	41	16	34	16
	1.25	80	24	65	26
	1.10	127	30	112	35
	1.00	170	34	180	45
	0.91	190	36	280	54
	0.80	177	35	296	62
	0.67	113	29	245	58
	0.50	54	18	178	34
0.10	2.00	20	13	16	12
	1.50	47	24	39	23
	1.25	109	44	82	43
	1.10	207	63	147	69
	1.00	299	74	189	93
	0.91	282	73	176	107
	0.80	167	57	112	88
	0.67	78	35	59	52
	0.50	35	20	20	26
0.05	2.00	22	18	21	14
	1.50	52	40	49	28
	1.25	130	86	114	56
	1.10	299	156	259	97
	1.00	497	179	436	131
	0.91	391	134	464	133
	0.80	177	80	236	86
	0.67	78	43	95	48
	0.50	37	25	42	26

situation of independent observations can be found in Domangue and Patch (1991) by noting that the EWRMS is the special case of their omnibus EWMA with exponent equal to 2.

The Exponentially Weighted Moving Variance (EWMV)

The EWMS will respond to both changes in the mean and changes in the variance. Therefore, it is sometimes more useful to compute the exponentially weighted moving variance about some estimate of the process mean (see Sweet (1986)). The EWMV is obtained from equation (1) by replacing η with a mean estimate η_n at each point in time, that is

$$s_n^2 = (1-r)s_{n-1}^2 + r[Y_n - \eta_n]^2 \ . \qquad (4)$$

A very convenient estimate η_n for the process mean at time n is the EWMA (Z_n) in equation (3). This is particularly convenient in the case where the process mean is not really constant, but is time varying, a situation rather typical in the continuous process industries and discussed further in the next section. Control limits for this EWMV are derived in Appendix B, and some values are presented in Table B1 for the case of independent observations, and in Table B2 for autocorrelated observations.

Figure 2 is a simulation of the same independent observation case of Figure 1, but where a shift in mean of 1.0 ($= 1\sigma$) unit occurs at $k = 50$, followed by a change in variance from 1.0 to 2.0 at $k = 110$. In Figure 2 the plots (a), (b), and (c) show the Shewhart, EWMA, and EWRMS charts as in Figure 1 for this new situation. The Shewhart (a) and EWMA plot (b) both detect the mean shift shortly after $k = 50$. The EWRMS plot (c) detects a shift in mean squared deviation resulting from the mean shift at $k = 50$ and then rises again due to the variance

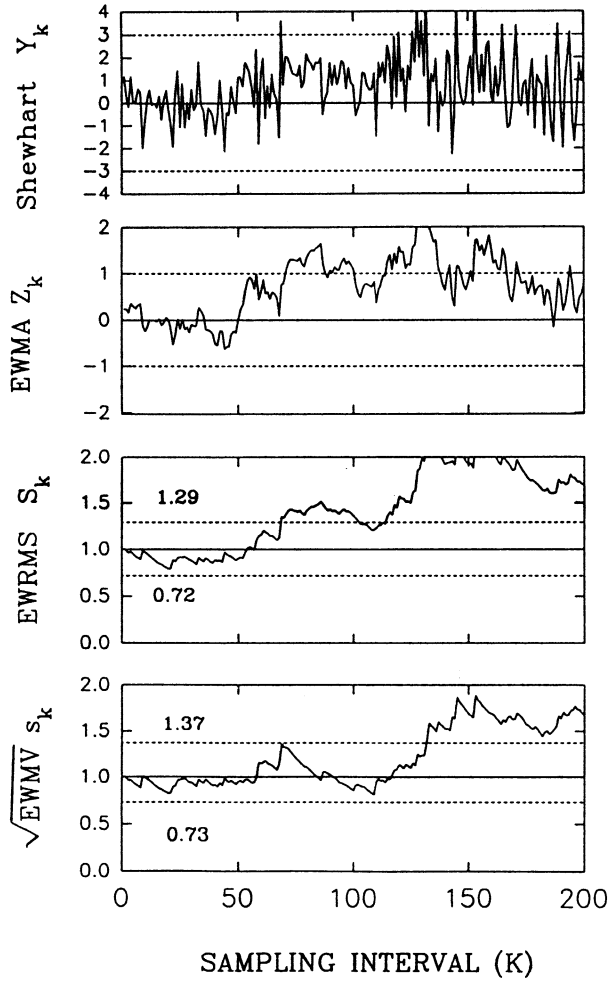

FIGURE 2. Independent Observations with Mean Change from 0.0 to 1.0 at $k = 50$, and a Variance Change from $\sigma^2 = 1.0$ to 2.0 at $k = 110$ with (a) Shewhart Chart on Individuals; (b) EWMA with $\lambda = 0.2$; (c) EWRMS with $r = 0.05$ and $S_0 = 1.0$; (d) $\sqrt{\text{EWMV}}$ with $r = 0.05$ and $s_0 = 1.0$.

increase at $k = 110$. However, the $\sqrt{\text{EWMV}}$ (s_k) in plot (d) of Figure 2 shows that the variance does not change until after $k = 110$. The control limits (0.73, 1.37) on the $\sqrt{\text{EWMV}}$ are those given in Table B1 corresponding to a probability level of $\alpha = 0.01$.

Autocorrelated Observations

The classical assumption that $Y_k = \eta + e_k$ with a constant mean and independent errors is often quite unrealistic for the continuous process industries (see MacGregor and Harris (1990), MacGregor (1990), and Harris and Ross (1991)). A more realistic assumption is that

$$Y_k = \eta_k + e_k \tag{5}$$

where the independent and identically distributed e_k are sampling/measurement errors (i.e., e_k is a white noise process) with mean zero and variance σ_e^2, and the true process mean η_k is time varying and not assumed to be a constant as in the classic Shewhart model. Persistent common cause variations such as might arise from raw material variations, impurities, or upstream process variations can give rise to such time varying process means. Assume that η_k follows a first order autoregressive process (i.e., η_k is AR(1)), so that

$$\eta_k = \phi\eta_{k-1} + \alpha_k \tag{6}$$

where α_k is a white noise sequence (independent of e_k) with mean zero and variance σ_α^2. The parameter ϕ $(-1 \leq \phi \leq 1)$ usually lies between 0 and 1.0 in continuous processes. It can easily be shown (see Anderson (1976, p. 146)) that the combined AR(1) plus white noise process in equation (5) can be equivalently expressed as the ARMA(1, 1) process

$$Y_k = \phi Y_{k-1} + a_k - \theta a_{k-1} \tag{7}$$

where a_k is a white noise process with mean zero and variance σ_a^2, and the autoregressive parameter ϕ is the same as that in equation (6) for the true process mean. The ratio of the random noise variance to the total variance in equation (5) is needed later and can be expressed as (see Anderson (1976, p. 147))

$$\sigma_e^2/\sigma_Y^2 = 1 - \rho_1/\phi \tag{8}$$

$$= 1 - \frac{(1-\phi\theta)(\phi-\theta)}{\phi(1+\theta^2-2\phi\theta)} \tag{9}$$

where ρ_1 is the lag one autocorrelation of the Y_k process. Note that equation (9) is obtained from equation (8) by substituting for the lag one autocorrelation of an ARMA(1, 1) process given in Box and Jenkins (1976, p. 77). A method for estimating the parameters ϕ and θ in this ARMA(1, 1) model is given in Appendix C.

Note that as the ϕ parameter approaches unity (a common situation in the process industries), the time varying mean in equation (6) becomes nonstationary and behaves as a random walk, and the observed Y_k (equations (5) and (7)) behaves as an integrated moving average (IMA) process. For this integrated moving average process the minimum mean squared

error predictor for Y_{n+j} $(j \geq 1)$ given observations $Y_n, Y_{n-1}, \ldots$ is the EWMA (Z_n) in equation (3) with $\lambda = 1 - \theta$. This is also the optimal estimate of the present and future process means η_{n+j} $(j \geq 0)$. The one step ahead prediction error $(Y_{n+1} - Z_n)$ has variance σ_a^2.

Although the EWMA is an optimal predictor for only the IMA(1, 1) model, it is not overly sensitive to the choice of the parameter λ. Therefore, it provides a reasonable prediction for other positively autocorrelated processes. Furthermore, unlike the case with independent data, for autocorrelated data an estimate of λ can be obtained from an in-control reference set of data as that value which minimizes the prediction error sum of squares.

More generally, any time series model identified from the data could be used to estimate the current process level or mean η_k (see Box and Jenkins (1976)). A simple test to check whether the model given in equation (5) or the classic Shewhart model is appropriate for a particular situation is to compute the autocorrelation function and test for the presence of significant autocorrelation at lags $l = 1, 2, \ldots$. Other SPC approaches for treating such autocorrelated data are outlined by Alwan and Roberts (1988). These approaches are analysed in Harris and Ross (1991). However, in this paper the basic control charts augmented with an EWMV chart will be used.

Figure 3 shows a simulation with autoregressive parameter $\phi = 0.90$, and where the independent error variance $\sigma_e^2 = 0.50$ accounts for 50% of the total variance of Y_k. These parameters would imply that the in-control process follows on ARMA(1, 1) model with $\theta = 0.37$ (from equation (8)). At $k = 90$ the variance of both the process mean component (η) and the random component (e) are doubled $(\sigma_\eta^2 = \sigma_e^2 = 1.0)$. Plots (a) and (b) again show the raw data in a Shewhart chart and the EWMA chart using an exponential weighting parameter $\lambda = 0.2$ (the control limits shown are those for independent observations). A good value of λ can be obtained by minimizing the sum of squares of the prediction errors $(Y_{k+1} - Z_k)^2$ on a reference data set. The EWRMS in plot (c) uses an exponential weighting parameter $r = 0.05$ and clearly picks up the change in the total process variance after $k = 90$. The distribution of the EWRMS and the appropriate control limits for it are derived in the Appendix for the case of autocorrelated observations. The control limits $(0.55\sigma_0, 1.49\sigma_0)$ in plot (c) are taken from Table A2 using $\alpha = 0.01$ and $\sigma_0 = 1.0$. The $\sqrt{\text{EWMV}}$ in plot (d) again uses

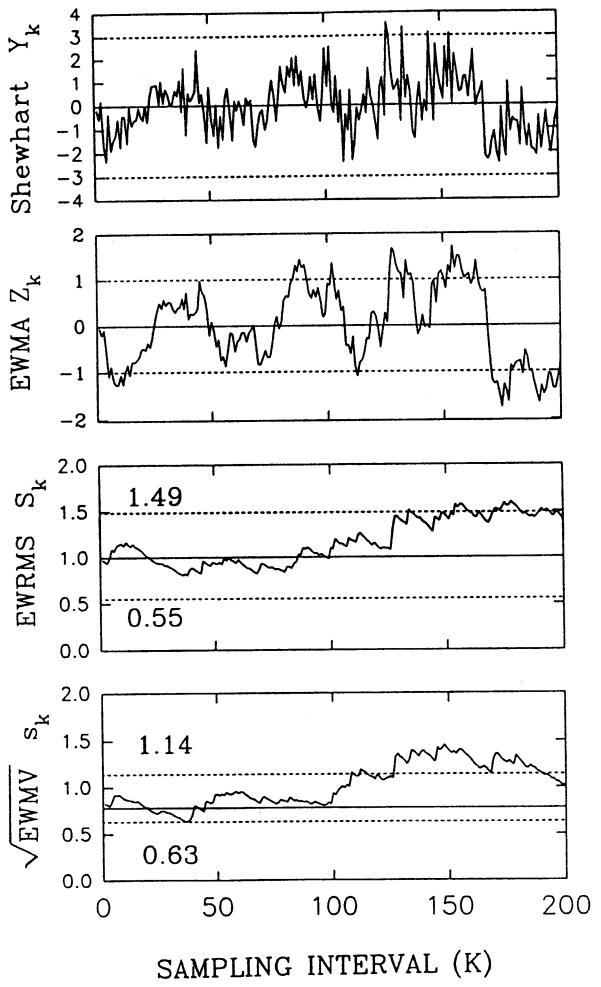

FIGURE 3. Autocorrelated Observations with a Variance Change From $\sigma^2 = 1.0$ to 2.0 at $k = 90$ with (a) Shewhart Chart on Individuals; (b) EWMA with $\lambda = 0.2$; (c) EWRMS with $r = 0.05$ and $S_0 = 1.0$; (d) $\sqrt{\text{EWMV}}$ with $r = 0.05$ and $s_0 = 0.7$.

$r = 0.05$ and is computed using the EWMA (Z_k) as an estimate of the true process level (η_k) at every sampling interval. In this case the EWMA is very close to the optimal one-step ahead prediction of Y_k, and hence the $\sqrt{\text{EWMV}}$ in Figure 3(d) is an estimate of the standard deviation σ_a of the one-step ahead prediction error $a_k = Y_k - Z_{k-1}$. This estimate could be used to provide time varying confidence intervals about the EWMA predictions in order to test whether they are significantly different from target at any time (see, e.g., Montgomery and Mastrangelo (1991)). The change in prediction error variance after $k = 90$ is clearly picked up by this plot. The control limits $(0.63\sigma_0, 1.14\sigma_0)$ used in plot (d) are taken from Table B2.

There are many different ways in which the EWMV can be used. For instance, in the case of equation (5) where the variance inflation (or deflation) could be due either to inflation of the random (measurement/sampling) component (e_k) or to an increase in the variance of the true process (η_k), more than one EWMS/EWMV chart could be used. In the typical case of a positively autocorrelated process ($\phi > 0$) an EWMS about the target using a small value of r (say 0.02) would reach far back into the data ($\nu = 99$). It would measure the longer term total process variation, while an EWMV taken about a time series estimate of the process mean (such as the EWMA), and using a very short memory (i.e., $r = 0.33$ and $\nu = 5$) would measure mainly the local random component of the variance. Note that it is only this latter random component of variance that would be estimated by the standard $\overline{X}$ and R chart. The range R of a sample of m observations collected at one time would give no indication of the variation in the real process mean η_k.

In practice the choice of the parameter r in the EWMV is completely at the discretion of the user. It should be selected to provide an estimate of the type of variability that one wishes to monitor. As shown in Table 1 a large value of r (i.e., $r = 0.33$) has very few degrees of freedom (i.e., $\nu = 5$) and measures short term variability analogous to the R chart when subgroups are available. A small value of r allows the EWMV to reach far back into the data and estimate longer term process variability. Tracking both types of variability would generally be worthwhile.

An Industrial Example

In the manufacture of paper an important property is the dry basis weight (weight/unit area of the paper on a dry basis). A calculation of this property is obtained on-line in most industrial paper machines using data from a scanning beta gauge (measuring total mass/unit area), and from a scanning capacitance gauge (measuring moisture content). In this way a dry basis weight measurement is available once per scan (usually every 30 to 60 seconds). Since the basis weight is strongly affected by the changing properties of the stock fed to the head box (e.g., its consistency, freeness, fiber length distribution, etc.) its behavior is generally non-stationary in the mean and is often well modeled by an IMA(1, 1) time series model. Therefore, an active feedback process control scheme is always used in order to maintain the dry basis weight near its target.

A record of 221 basis weight values from a paper machine under computer control is shown in Figure 4(a). Since the feedback control scheme contained integral action (see, e.g., Smith and Corripio (1985)) the basis weight is stationary with its mean equal to the target ($\eta = \tau$). An EWRMS plot for the data using an exponential weighting parameter $r = 0.05$ is shown in Figure 4(b). A very sudden change in the variance is detected by the EWRMS plot around the 155^{th} scan. Although the reason for this sudden increase in the variance for this data set is not known, it could be due to any number of causes, such as a change in the nature of the disturbances resulting from feedstock changes or upstream process changes, from difficulties in the moisture control loop, and so forth. If an EWRMS chart had been active at the time this data was being collected, the operators or engineer would have been immediately alerted to this out-of-control situation. Then they might have been able to assign a cause, and thereby improve the control system or remove the source of disturbance. The importance of using such SPC charts to monitor active computer control systems has been stressed by MacGregor and Harris (1990).

The upper and lower control limits on the EWRMS plot in Figure 4(b) were obtained as follows. The first 140 observations where the process was very stable and operating well were used to provide the reference distribution. (Normally a larger set of historical

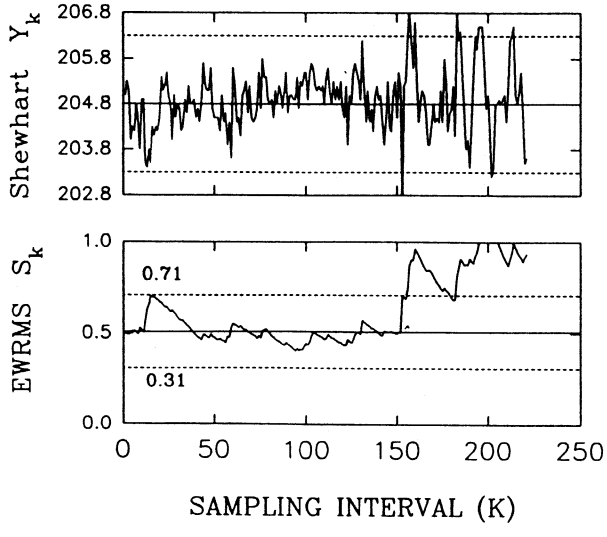

FIGURE 4. Basis Weight Data from an Industrial Paper Machine (a) Basis Weight Scan Averages; (b) EWRMS with $r = 0.05$ and $S_0 = 0.51$.

data would be preferable). The standard deviation of this reference data set was $s_0 = 0.51$ and the estimated lag one autocorrelation was $\rho_1 = 0.40$. An ARMA(1, 1) model (given in equation (7)) was fitted to the data (see Appendix C) yielding $\phi = 0.81$ and $\theta = 0.51$. Using these estimates in equation (9) gave an estimate of $\sigma_e^2/\sigma_Y^2 = 0.50$. Interpolating in Table A1 using these estimates gives the degrees of freedom for the $\chi^2(\nu)/\nu$ distribution as $\nu = 21$. Then using equations (2) with $\alpha = 0.01$ the upper and lower control limits for the EWRMS are found to be $(0.62\sigma_0, 1.40\sigma_0) = (0.32, 0.71)$.

Conclusion

The exponentially weighted moving variance (EWMV) and mean squared deviation (EWMS) charts have been proposed as alternative means of monitoring process variation. They are particularly useful when

- only individual observations are collected and analysed and hence no estimate of variability is available from within subgroups, and
- the observations are positively autocorrelated due to real process mean variations, and hence even the within subgroup variation would only estimate the short term random component of the variability.

The EWMV and EWMS are extremely simple to calculate in a recursive manner. Also, the estimate of the process mean about which the variance is being calculated and the exponential weighting used provide great flexibility for monitoring different types of process variability.

Acknowledgment

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, and the helpful comments of the Editor and a reviewer.

Appendix A

Distribution of the Exponentially Weighted Mean Square

Consider the exponentially weighted mean squared deviation in equation (1) where the n random variables $Y_k - \eta$ have mean zero and are distributed as a multivariate normal distribution with $n \times n$ covariance matrix $\sigma_Y^2\mathbf{P}$ ($\mathbf{P}$ is the correlation matrix of the Y_k's). For the moment let us ignore the second term $((1-r)^n S_0^2)$ on the right side of equation (1), and we shall derive the distribution of S_n^2 for large n. Using an extension of Cochran's theorem (see Box (1954)) it can be shown for large n that S_n^2/σ_Y^2 is distributed as the quantity

$$Q = \sum_{j=1}^{n} \zeta_j \chi_j^2(1) \tag{A1}$$

where the $\chi_j^2(1)$ variates are independent, the ζ_j's are the n real non-zero latent roots of the matrix

$$\mathbf{U} = \mathbf{PR} \tag{A2}$$

and $\mathbf{R}$ is a diagonal matrix with elements $[r, r(1-r), r(1-r)^2, ..., r(1-r)^{n-1}]$. Furthermore, Box (1954) showed that a good approximation to the distribution of Q is given by

$$Q \sim g\,\chi^2(\nu) \tag{A3}$$

where g and ν are given by

$$g = \sum_{j=1}^{n} \zeta_j^2 \Big/ \sum_{j=1}^{n} \zeta_j \tag{A4}$$

and

$$\nu = \left(\sum_{j=1}^{n} \zeta_j\right)^2 \Big/ \sum_{j=1}^{n} \zeta_j^2. \tag{A5}$$

To evaluate g and ν it is not necessary to calculate all the latent roots ζ_j since one is only interested in their sum and sum of squares. These are given by

$$\sum_{j=1}^{n} \zeta_j = \text{trace}(\mathbf{U}) \tag{A6}$$

and

$$\sum_{j=1}^{n} \zeta_j^2 = \text{trace}(\mathbf{UU}). \tag{A7}$$

If one denotes the ij^{th} element of $\mathbf{U}$ by u_{ij}, then

$$\text{trace}(\mathbf{UU}) = \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij}u_{ji}.$$

Since $\mathbf{R}$ is diagonal, the i^{th} column of $\mathbf{U}$ is simply the i^{th} column of the correlation matrix multiplied by the i^{th} diagonal element of $\mathbf{R}$. Thus,

$$u_{ij} = r(1-r)^{j-1}\rho_{|i-j|}$$

where $\rho_0 = 1$. It is easy to verify that

$$\lim_{n\to\infty} \sum_{j=1}^{n} \zeta_j = 1. \tag{A8}$$

Using an expansion of trace(**UU**), one can verify that

$$\lim_{n\to\infty} \sum_{j=1}^{n} \zeta_j^2 = \frac{r\left[1 + 2\sum_{j=1}^{n} \rho_j^2 (1-r)^j\right]}{2-r}. \tag{A9}$$

As a result of equation (A8) it is easily seen from equations (A4) and (A5) that $g = 1/\nu$, and hence

$$Q \sim \chi^2(\nu)/\nu. \tag{A10}$$

Independent Observations

When the observations are independent (i.e., $\rho_{|i-j|} = 0$ for $i \neq j$), then from equations (A5), (A8), and (A9) the degrees of freedom ν in the χ^2 distribution (A10) is given by $\nu = (2-r)/r$. This result has also been stated without proof by Wortham (1972).

Autocorrelated Observations

For any correlation matrix **P** equations (A6) and (A7) can be evaluated, and the appropriate degrees of freedom ν of the χ^2 distribution for Q can be determined. We consider here only the autocorrelated process treated in the paper, namely that given by equations (5) and (6). For this process

$$\begin{aligned} \rho_1 &= (1 - \sigma_e^2/\sigma_Y^2)\phi \\ \rho_j &= \phi^{j-1}\rho_1 \text{ for } j > 1 \end{aligned} \tag{A11}$$

and therefore

$$\lim_{n\to\infty} \sum_{j=1}^{n} \zeta_j^2 = \left(\frac{r}{2-r}\right)\left(1 + \frac{2(1-\sigma_e^2/\sigma_Y^2)^2(1-r)\phi^2}{1-(1-r)\phi^2}\right).$$

TABLE A1. Degrees of Freedom ν for the EWMS with $r = 0.05$ for the Autocorrelated Observation Model of Equations (5) and (6)

σ_Y^2/σ_0^2	ϕ 0.1	0.25	0.5	0.75	0.9
1.00	39.0	39.0	39.0	39.0	39.0
0.90	39.0	39.0	38.8	38.1	36.6
0.50	38.8	37.8	33.7	24.8	14.6
0.10	38.4	35.3	25.9	13.6	6.10

TABLE A2. Control Limit Constants C_5 and C_6 for the EWRMS with $r = 0.05$, Autocorrelated Observations ($\phi = 0.90$), and $\sigma_e^2/\sigma_0^2 = 0.50$

		Box $\chi^2(\nu)/\nu$ Approximation	Johnson Distribution
$\alpha = 0.05$	C_5	0.64	0.73
	C_6	1.36	1.39
$\alpha = 0.01$	C_5	0.55	0.69
	C_6	1.49	1.59

The degrees of freedom ν for the EWMS with an exponential weighting parameter $r = 0.05$ are given in Table A1 for various values of the parameters ϕ and σ_e^2/σ_Y^2. Note that as the process becomes more highly autocorrelated (ϕ large and σ_e^2/σ_Y^2 small) the degrees of freedom in the EWMS fall off rapidly. The upper and lower control limits for the EWRMS based on Box's χ^2 approximation (A10) are given by $(\sigma_0 - C_5\sigma_0, \sigma_0 + C_6\sigma_0)$ where C_5 and C_6 are the square roots of the critical values of the appropriate $\chi^2(\nu)/\nu$ distribution. The control limit constants C_5 and C_6 are provided in Table A2 for an EWRMS with $r = 0.05$, $\alpha = 0.01$ and 0.05, $\sigma_e^2/\sigma_Y^2 = 0.5$, and $\phi = 0.90$.

To employ these tables in a practical situation one must estimate the ratio of the random noise variance to the total observation variance σ_e^2/σ_Y^2 and the value of the AR(1) parameter ϕ. If σ_e^2/σ_Y^2 is known from a prior ANOVA then ϕ can be estimated from equations (A11) as

$$\phi = \frac{\rho_1}{(1 - \sigma_e^2/\sigma_Y^2)}$$

where ρ_1 is the estimated lag 1 autocorrelation coefficient. If estimates of neither σ_e^2/σ_Y^2 nor ϕ are available then an ARMA(1, 1) model can be fitted to a sample of in-control data, and equations (8) or (9) used to obtain estimates, as illustrated in the industrial example in the paper.

Higher-Order Approximations

A theoretical justification for the chi squared approximation (A3) was given by Patnaik (1949). He noted that the approximation was better in the upper tails than the lower tails. To investigate the adequacy of (A3) we also evaluated the first four moments of (A1) and used Johnson curves (see Johnson (1949)) to obtain more accurate control limits. (Note that

the approximation (A10) obtained from Box (1954) uses only a two-moment approximation of the distribution.)

From the quadratic form (A1) the moments of the distribution are (see Box (1954))

$$\mu_1' = \sigma_Y^2 \text{ trace}(\mathbf{U})$$
$$\mu_2 = 2\sigma_Y^4 \text{ trace}(\mathbf{UU})$$
$$\mu_3 = 8\sigma_Y^6 \text{ trace}(\mathbf{UUU})$$
$$\mu_4 = 12\left[4\sigma_Y^8 \text{ trace}(\mathbf{UUUU}) + \mu_2^2\right]. \quad \text{(A12)}$$

Once these moments have been calculated the form and parameters of the Johnson distribution can be determined from tables in Pearson and Hartley (1972) or using the computer program in Hill et al. (1976). In either case it is necessary to evaluate the expressions in (A12). When the observations are independent it is straightforward to show that

$$\mu_1' = \sigma_Y^2$$
$$\mu_2 = \frac{2\sigma_Y^4}{2-r}$$
$$\mu_3 = \frac{8\sigma_Y^6 r^3}{1-(1-r)^3}$$
$$\mu_4 = 12\sigma_Y^8\left\{\frac{4r^4}{[1-(1-r)^4]+4/(2-r)^2}\right\}.$$

When the observations are not independent, the moments can be calculated by selecting a large value for n (> 100) and evaluating equations (A12) numerically.

The lower half of Table 1 shows the control limits based on the Johnson curves for independent observations and Type I error levels $\alpha = 0.01$ and 0.05. For large degrees of freedom these match closely those based on the two-moment chi squared approximation of Box. However, for small degrees of freedom the differences are greater. In particular, the Johnson curves modify the lower control limits most, moving them in closer to the target, while moving the upper limits slightly further away from the target. Similar behavior is observed in the case of autocorrelated observations (Table A2).

Time Varying Control Limits

The control limits derived above are all asymptotic results. When an adjustment is made to the process, for example, to eliminate an assignable cause of variation, the EWMRS chart is restarted with a new initial condition S_0. The asymptotic control limits will then not be appropriate for the early values of the EWRMS. In general the appropriate control limits will be time varying, but will gradually approach the previous asymptotic limits. Using the time varying limits will allow for fast initial response behavior similar to that with EWMA charts (see MacGregor and Harris (1990)).

The development of these time varying limits parallels the development in the previous section. We shall illustrate this using the approximation of Box. Let the initial estimate of the variance S_0^2/σ_Y^2 be approximated as $\chi^2(\nu_0)/\nu_0$, where ν_0 is the effective degrees of freedom associated with S_0^2. Then Q is approximately distributed as $g'\,\chi^2(\nu')$ where

$$g' = \frac{\sum_{j=1}^{n} \zeta_j^2 + \frac{(1-r)^{2n}}{\nu_0}}{\sum_{j=1}^{n} \zeta_j + (1-r)^n} \quad \text{(A13)}$$

and

$$\nu' = \frac{\left(\sum_{j=1}^{n} \zeta_j + (1-r)^n\right)^2}{\sum_{j=1}^{n} \zeta_j^2 + \frac{(1-r)^{2n}}{\nu_0}}. \quad \text{(A14)}$$

The summations in equations (A13) and (A14) are calculated using equations (A6) and (A7). Three cases are noted.

(i) When $\nu_0 < \nu$, then the limits for small values of n are larger than the asymptotic values.

(ii) When $\nu_0 > \nu$, then the limits for small values of n are less than the asymptotic values.

(iii) When $\nu_0 = \nu$, the limits are not a function of n and are the same as the asymptotic limits.

Appendix B

Distribution of the Exponentially Weighted Moving Variance

Using an EWMA to estimate the local mean (η_n) of the process, the EWMV is defined by the coupled equations (3) and (4). It is straightforward to show that s_n^2 can be written as the quadratic form

$$s_n^2 = \mathbf{Y}'(\mathbf{I}-\mathbf{M})'\mathbf{R}(\mathbf{I}-\mathbf{M})\mathbf{Y} + (1-r)^n s_0^2 \quad \text{(B1)}$$

where $\mathbf{Y}$ denotes the vector $(Y_n, Y_{n-1}, \ldots, Y_1)$, the

TABLE B1. Moments of s_n^2/σ_Y^2 and Control Limit Constants C_7 and C_8 for the $\sqrt{\text{EWMV}}$ with $\lambda = 0.20$ for Independent Observations

	r	0.01	0.02	0.05	0.10	0.20	0.33
$E(s_n^2/\sigma_Y^2)$		1.11	1.11	1.11	1.11	1.11	1.11
$\sqrt{\text{Var}(s_n^2/\sigma_Y^2)}$		0.12	0.17	0.27	0.38	0.56	0.72
Box $\alpha = 0.05$	C_7	0.91	0.90	0.80	0.70	0.56	0.42
	C_8	1.12	1.22	1.29	1.40	1.55	1.70
Box $\alpha = 0.01$	C_7	0.87	0.85	0.73	0.61	0.44	0.29
	C_8	1.16	1.27	1.37	1.52	1.73	1.96

matrix $\mathbf{R}$ is as defined in equation (A2), and the matrix $\mathbf{M}$ is

$$\begin{bmatrix} 0 & \lambda & \lambda(1-\lambda) & \lambda(1-\lambda)^2 & \cdots & \lambda(1-\lambda)^{n-1} \\ & 0 & \lambda & \lambda(1-\lambda) & \cdots & \lambda(1-\lambda)^{n-2} \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & 0 & \lambda & \lambda(1-\lambda) \\ & & & & 0 & \lambda \\ & & & & & 0 \end{bmatrix}.$$

Either the two-moment Box chi squared or the four-moment Johnson curve approximation developed in Appendix A can now be used with

$$\mathbf{U} = \mathbf{P}(\mathbf{I} - \mathbf{M})'\mathbf{R}(\mathbf{I} - \mathbf{M}).$$

In the case of independent observations, Sweet (1986) used a normal approximation for the asymptotic distribution of (B1) using the first two moments of (B1). The expected value of s_n^2/σ_Y^2, for independent observations and large n, is given by

$$E(s_n^2/\sigma_Y^2) = \frac{2}{2-\lambda}. \qquad \text{(B2)}$$

This value always exceeds 1 and hence s_n^2 is a biased estimator. The higher moments of (B1) are dependent on both λ and r.

The upper and lower control limits for the EWMV are given by $(\sigma_0 - C_7\sigma_0, \sigma_0 + C_8\sigma_0)$ where C_7 and C_8 are the square roots of the critical values calculated from either the Johnson curve approximation or the $g\,\chi^2(\nu)$ approximation. Values of C_7 and C_8 are shown in Table B1 for $\lambda = 0.20$ and the case of independent data using the approximation of Box. Calculation of the higher-order moments via equations (A12) was impractical due to the slow convergence of the quantities. The first two moments were calculated from closed-form expressions given in Sweet (1986). We note that the limits are wider than the corresponding limits for the EWRMS due to the additional uncertainty in the process mean. For autocorrelated observations following the model in equations (5) and (6) the upper and lower control limits for $\lambda = 0.20$, $r = 0.05$, $\phi = 0.9$, and $\sigma_e^2/\sigma_Y^2 = 0.50$ are shown in Table B2.

TABLE B2. Moments of s_n^2/σ_Y^2 and Control Limit Constants C_7 and C_8 for the $\sqrt{\text{EWMV}}$ with $\lambda = 0.2$ and $r = 0.05$ for the Autocorrelated Observation Model of Equations (5) and (6) with $\phi = 0.90$ and $\sigma_e^2/\sigma_Y^2 = 0.50$

$E(s_n^2/\sigma_Y^2)$		0.78
$\sqrt{\text{Var}(s_n^2/\sigma_Y^2)}$		0.18
Box $\alpha = 0.05$	C_7	0.68
	C_8	1.08
Box $\alpha = 0.01$	C_7	0.63
	C_8	1.14

Appendix C

Fitting an ARMA(1, 1) Model to Time Series Data

For readers not familiar with time series analysis we provide here a very brief discussion on how to estimate the parameters of the ARMA(1, 1) model in equation (7). Given a time series consisting of N mean corrected observations, the parameters ϕ and θ of the ARMA(1, 1) model can be estimated by minimizing the objective function

$$J = \sum_{i=1}^{N} \left(Y_i - \widehat{Y}_i\right)^2$$

where $\widehat{Y}_i$ is the one-step ahead forecast of Y_i, which can be calculated recursively for any trial values of ϕ and θ as

$$\widehat{Y}_i = \phi Y_{i-1} - \theta(Y_{i-1} - \widehat{Y}_{i-1}) \ .$$

The recursion is typically started with $\widehat{Y}_i$ equal to the first observation Y_1. The parameter estimates that minimize the objective function J require an iterative search. This can be accomplished using available time series packages or a general purpose nonlinear estimation routine. For this simple model the parameters can also be estimated by a simple grid search over the region $(-1 \leq \phi \leq 1, -1 \leq \theta \leq 1)$.

The above estimation method is usually justified on the basis that it minimizes the variance of the one-step ahead prediction errors. Under the assumption of normally distributed errors, it is also a conditional maximum likelihood method. A good introduction to time series analysis methods is the text by Cryer (1986).

References

ANDERSON, O. D. (1976). *Time Series and Forcasting.* Buttersworth, London.

ALWAN, L. C. and ROBERTS, H. V. (1988). "Time Series Modeling for Statistical Process Control". *Journal of Economics and Business Statistics* 6, pp. 87-95.

BAUER, P. and HACKL, P. (1978). "The Use of MOSUMS for Quality Control". *Technometrics* 20, pp. 431-436.

BAUER, P. and HACKL, P. (1980). "An Extension of the MOSUM Technique for Quality Control". *Technometrics* 22, pp. 1-8.

BOX, G. E. P. and JENKINS, G. M. (1976). *Time Series Analysis*, 2nd ed. Holden-Day, San Francisco, CA.

BOX, G. E. P. (1954). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: Effect of Inequality of Variance in One-Way Classification". *Annals of Mathematical Statistics* 25, pp. 290-302.

CROWDER, S. V. and HAMILTON, M. (1992). "An EWMA for Monitoring Standard Deviation". *Journal of Quality Technology* 24, pp. 12-21.

CRYER, J. D. (1986). *Time Series Analysis.* Duxbury Press, Boston, MA.

DOMANGUE, R. and PATCH, S. C. (1991). "Some Omnibus Exponentially Weighted Moving Average Statistical Process Monitoring Schemes". *Technometrics* 33, pp. 299-313.

HARRIS, T. J. and ROSS, W. H. (1991). "Statistical Process Control Procedures for Correlated Observations". *Canadian Journal of Chemical Engineering* 69, pp. 48-57.

HILL, I. D.; HILL, R.; and HOLDER, R. H. (1976). "Algorithm A599: Fitting Johnson Curves by Moments". *Applied Statistics* 25, pp. 180-189.

HSU, D. A. (1977). "Test for Variance Shift at an Unknown Time Point". *Applied Statistics* 26, pp. 279-284.

HUNTER, J. S. (1986). "The Exponentially Weighted Moving Average". *Journal of Quality Technology* 18, pp. 203-210.

JOHN, P. W. (1990). *Statistical Methods in Engineering Quality Assurance.* John Wiley & Sons, New York, NY.

JOHNSON, N. L., (1949). "Systems of Frequency Curves Generated by Methods of Translation". *Biometrika* 36, pp. 149-176.

LUCAS, J. M. (1976). "The Design and Use of V-Mask Control Schemes". *Journal of Quality Technology* 8, pp. 1-12.

MACGREGOR, J. F. and HARRIS, T. J. (1990). Discussion of "EWMA Control Schemes" by Lucas and Saccucci. *Technometrics* 32, pp. 23-26.

MACGREGOR, J. F. (1990). "A Different View of the Funnel Experiment". *Journal of Quality Technology* 4, pp. 255-259.

MONTGOMERY, D. C. (1991). *Statistical Quality Control.* John Wiley & Sons, New York, NY.

MONTGOMERY, D. C. and MASTRANGELO, C. M. (1991). "Some Statistical Process Control Methods for Autocorrelated Data". *Journal of Quality Technology* 23, pp. 179-193.

NELSON, L. S. (1980). "The Mean Square Successive Difference Test". *Journal of Quality Technology* 12, pp. 174-175.

NG, C. H. and CASE, K. E. (1989). "Development and Evaluation of Control Charts Using Exponentially Weighted Moving Averages". *Journal of Quality Technology* 21, pp. 242-250.

PATNAIK, P. B. (1949). "The Non-Central χ^2 and F Distributions and Their Applications". *Biometrika* 36, pp. 202-232.

PEARSON, E. S. and HARTLEY, H. O. (1972). *Biometrika Tables for Statisticians* Vol. 2. Cambridge University Press.

SMITH, C. A. and CORRIPIO, A. B. (1985). *Principles and Practice of Automatic Process Control.* John Wiley & Sons, New York, NY.

SMITH, D. E. (1976). "Adaptive Response for Exponential Smoothing: Comparative System Analysis". *Operational Research Quaterly* 25, pp. 421-435.

SWEET, A. L. (1986). "Control Charts Using Coupled Exponentially Weighted Moving Averages". *IEEE Transactions* 18, pp. 26-33.

TRIGG, D. W. and LEACH, A. G. (1967). "Exponential Smoothing with an Adaptive Response Rate". *Operational Research Quarterly* 18, pp. 53-59.

TSAY, R. S. (1988). "Outliers, Level Shifts and Variance Changes in Time Series". *Journal of Forecasting* 7, pp. 1-20.

WORTHAM, A. W. and RINGER, L. J. (1971). "Control via Exponential Smoothing". *Transportation and Logistic Review* 7, pp. 33-39.

WICHERN, D. W.; MILLER, R. B.; and HSU, D. A. (1976). "Changes of Variance in First-Order Autoregressive Time Series Models—with an Application". *Applied Statistics* 25, pp. 248-256.

WORTHAM, A. W. (1972). "The Use of Exponentially Smoothed Data in Continuous Process Control". *International Journal of Production Research* 10, pp. 393-400.

Key Words: *Autocorrelated Observations, Quality Control, Process Variation.*