Change-point Problems
IMS Lecture Notes – Monograph Series (Volume 23, 1994)

LIKELIHOOD RATIO METHODS FOR MONITORING PARAMETERS OF A NESTED RANDOM EFFECT MODEL

BY EMMANUEL YASHCHIN
IBM Research Division

In many practical situations the variance of a set of measurements can be attributed to several known sources of variability. For example, if several measurements of each item of a lot are taken, one may need to deal not only with the within-item variability, but also with item-to-item-within-lot and lot-to-lot components of variability. In such cases conventional control charts tend to produce an unacceptably high rate of false alarms and, in general, represent a rather weak diagnostic tool. This paper shows how to build a control system, based on Likelihood Ratio Tests, capable of controlling the mean and variance components of a nested random effect model. The strong points and weaknesses of this approach are compared to those of competing methods.

1. Introduction. One of the major aims of Statistical Process Control (SPC) is to achieve the condition where all the parameters related to a given manufacturing, business, ecological or similar process, conform to some prescribed on-target behavior. The means by which this aim is achieved include not only direct process adjustments, but also identification and neutralization of so called *special* causes of unfavorable changes in parameters of interest. Success in this form of control depends, to a large extent, on identification of a suitable model for the observed process. Once this is done and a systematic corrective action is taken to reduce variability due to predictable effects of feed-forward and feed-back variables, process control activity concentrates on monitoring the adequacy of the model and the levels of its parameters.

In many practical situations, the relevant model involves the mean and various measures of variability for a given type of measurement. For example, in situations where production is on a lot-by-lot basis and several measurements are taken at random from each lot, one will usually need to monitor not only the measure of total variance, but also its individual components. This is important because of the following reasons. First, knowing which component

---

AMS 1991 Subject Classification: Primary 62L10; Secondary 62N10

Key words and phrases: Change-point control charts, random effects, run length, variance components.

of variance is out of control is important for diagnosing the problem, since different components are usually affected by different special causes. Second, the ability of screening procedures to improve the outgoing quality depends strongly on the individual variance components: for a fixed total variance, the higher the proportion of variance due to nested factors the more difficult it is to screen out defective product.

In practical applications, the presence of several controllable causes of variability is usually revealed by the phenomenon of an unusually high rate of false alarms produced by standard $\bar{X} - R$ charts based on standard textbook procedures (e.g. see Wells and Smith (1991)). The main reason for this is related to the commonly made assumption that the lot-to-lot or higher components of variance are negligibly small which, in turn, leads to too tight control limits. The standard prescription in such situations involves increasing the Shewhart-type control limits on the chart for the process mean and supplementing it with the moving range chart. Yashchin (1991) discussed monitoring of the process mean and individual components of variance in a nested random effect model by using the Cusum technique. Woodall and Thomas (1991) gave a review of SPC involving several components of variability.

The strategy discussed here is based on the Likelihood Ratio Approach (e.g. see Basseville and Benveniste (1986)), which is known to lead to optimal procedures in several settings, including serially correlated data (eg. see Lorden (1971), Moustakides (1986) and Bansal and Papantoni-Kazakos (1986)). In Section 2 the general random effect model is introduced and the basic monitoring procedures are described. Section 3 deals with design of control schemes for the grand process mean and the lowest component of variance. In Sections 4 and 5 procedures for monitoring higher components of variance are discussed. Finally, Section 6 contains some concluding remarks.

2. The Model and Monitoring Strategy. For the sake of simplicity, the discussion in this work will be related to a specific situation arising in the process of manufacturing integrated circuits (chips) used in computers. Chips are typically processed as part of a wafer, which is a thin disk about 20 cm in diameter. Each wafer contains approximately 100 square shaped chips. Only at the final stages of processing are the wafers diced to produce individual chips. For most of the production process (which can take months), wafers are handled in lots. For example, when a given tool is used to perform one of hundreds of steps required to turn a raw wafer into a set of chips, the whole lot is processed as a single unit. Typically, a lot contains about 20 wafers. Our discussion will focus on a specific process step which deposits a thin layer of silicon oxide onto the surface of a wafer. To accomplish this step, the lot is placed inside a tool and, after being processed for a specified period of time, is

taken out for measurements and further processing by tools down the line. In this step, it is very important to assure the correct thickness of the oxide layer as well as its uniformity, so as to prevent electrical defects or degradation in performance of the final product.

2.1. The Model. In the present work, the oxide thickness corresponding to the population of diced chips will be assumed to follow a nested random effect model,

$$X_{irn} = \mu + L_i + W_{r(i)} + E_{irn} \quad i = 1,2,\ldots \quad r = 1,2,\ldots R, \quad n = 1,2,\ldots N \quad (2.1)$$

where X_{irn} is the thickness corresponding to the n-th chip on the r-th wafer of the i-th lot, μ is the grand process mean, $L_i \sim N(0,\sigma_b)$ is the random effect of the i-th lot, $W_{r(i)} \sim N(0,\sigma_w)$ is the nested effect of the r-th item in the i-th lot and $E_{irn} \sim N(0,\sigma)$ is the random noise representing the efect of the n-th measurement taken from the r-th item of the i-th lot. Without loss of generality, we shall assume that R is the number of wafers randomly selected from each lot for the purpose of monitoring, and that N represents the size of a random sample of oxide film measurements taken from each wafer.

In the model (2.1), σ_b, σ_w and σ represent the lot-to-lot, wafer-to-wafer-within-lot and within-wafer components of variability. Together with μ, these parameters are of primary interest. Efficient monitoring procedures can be based on the control sequence of sufficient statistics $\{\hat{\mu}_i, \hat{\sigma}^2_{i\bullet}, \hat{\sigma}^2_i\}$, $i = 1,2,\ldots$, where

$$\begin{aligned}
\hat{\mu}_i &= \frac{1}{RN}\sum_{r=1}^{R}\sum_{n=1}^{N} X_{irn} \\
\hat{\sigma}^2_{i\bullet} &= \frac{1}{R-1}\sum_{r=1}^{R}\left(\bar{X}_{ir\bullet} - \hat{\mu}_i\right)^2 \qquad (2.2) \\
\hat{\sigma}^2_i &= \frac{1}{R(N-1)}\sum_{r=1}^{R}\sum_{n=1}^{N}\left(X_{irn} - \bar{X}_{ir\bullet}\right)^2
\end{aligned}$$

and, in accordance with the conventional notation, $\bar{X}_{ir\bullet}$ is the average of the N measurements taken from the r-th wafer of the i-th lot.

Before proceeding with design of control schemes for monitoring the parameters of (2.1), one must perform a thorough analysis of available data, so as to establish the relevance of the model as well as the acceptable and rejectable levels of its parameters. One must give due attention to the possibility that some lots may contain outliers and to the fact that historic data sets may reflect the presence of changes associated with raw materials or corrective actions (possibly related to other characteristics of the product) and thus not

have a fixed mean. Therefore, conventional estimation procedures developed for designed nested experiments may turn out not to be appropriate in situations involving process control. Yashchin (1991) discusses the problem of estimation of variance components in this environment.

The results of this article remain relevant in situations when more than two nested components of variance are present. In such situations, monitoring of the higher components can be completed by merely reducing the appropriate lowest nested groups of data to their averages and behaving as if no individual measurements within such groups ever existed. For example, the reader will notice that, in what follows, monitoring of the highest variance component, σ_b of (2.1), is solely based on the within-item averages (which represent sufficient statistics for this purpose) and ignores the individual measurements.

The Monitoring Strategy. Once the process has been analyzed, one should specify the acceptable region (operating window), Ω_0, and the unacceptable region, Ω_1, for every monitored parameter. In this process one will generally take into account specification limits as well as the values that other parameters can take under normal operating conditions. This article discusses monitoring based on the Likelihood Ratio Approach, which can be briefly summarized as follows:

Likelihood Ratio (LR) Strategy: Trigger an out of control signal at time T if for some $l \geq 1$

$$D_l^* \stackrel{\text{def}}{=} L_{l1}^* - L_{l0}^* > h, \tag{2.3}$$

where $h \geq 0$ is a pre-specified signal level, L_{l1}^* is the maximum log-likelihood of the data observed within the last l periods of time under the assumption that the controlled parameter changed from some value in Ω_0 to a value within Ω_1 l units of time ago and L_{l0}^* is the similar maximum log-likelihood achieved under the assumption that the controlled parameter stays within Ω_0.

The above strategy leads to powerful procedures for a wide class of situations involving control of univariate and multivariate processes with or without serial correlation (e.g. see Lorden (1971), Nikiforov (1983), Basseville (1988), Telksnys (1986)). Its main drawback is related to the fact that one needs to go to the very beginning of the data in order to decide whether a signal is to be triggered at time T. This is obviously impractical and, therefore, the scheme is usually run by choosing a value L and triggering a signal only if h is exceeded for values $1 \leq l \leq L$. In effect, it amounts to running a truncated SPRT backwards in time. Alternatively, one could determine the depth L_T based on the previous history in the following way:

Regenerative Likelihood Ratio (LRL) Strategy: Given that at time T the last regeneration point was registered L_T units of time ago, trigger a signal if

$D_l^* > h$ for *some* $1 \leq l \leq L_T$. If $D_l^* \leq 0$ for *every* $1 \leq l \leq L_T$, declare T the new regeneration point.

In the above procedure, the first regeneration point can be set to represent the first moment of time where it is believed that the monitored parameter is in Ω_0, i.e. the point at which the control scheme is re-initiated, e.g. after repair of the faulty tool. Actual implementation of this strategy depends on the concrete situation and on assumptions about the nuisance parameters. For example, in cases where nuisance parameters need to be estimated, one may need to expand the depth of search for l in the above procedure beyond L_T.

When the sequence of statistics used to monitor a given parameter is iid and Ω_0 and Ω_1 reduce to a single point (i.e. control scheme is designed under the assumption that values of the parameter before and after the change are known), the LR and RLR strategies lead to identical schemes. Indeed, these strategies are associated with the V-mask and Page's Cusum procedures, respectively, and thus equivalence follows from the well known result of Kemp (1961). When the observations follow a one-dimensional distribution that belongs to the exponential family and Ω_0 and Ω_1 represent non-overlapping intervals on the real line, the LR and RLR strategies once again turn out to be equivalent (the proof of this statement is non-trivial and thus will not be given here). In more general situations, the LR and RLR strategies lead to different schemes having a roughly comparable statistical performance. In cases where it can be assumed that all the parameters stay at a constant level for relatively long periods of time between changes, the LR strategy tends to be slightly more powerful. In other cases it is advisable to use the RLR approach.

In general, performance of control schemes is characterized by the Run Length (RL), which represents the number of observations (in the case of the model (2.1), the number of lots) taken before an out of control signal is triggered. The problem of design of control schemes thus involves finding control scheme parameters such that the Average Run Length (ARL) is sufficiently large when the controlled parameter is in Ω_0 and sufficiently small when it is in Ω_1. In the next three sections an approach to monitoring the parameters of the basic model (2.1) based on the LR/RLR strategy, is developed.

Control Schemes for Monitoring μ and σ. The strategy outlined in the previous section is relatively easy to implement for two of the four parameters of interest, μ and σ. First of all, note that $\hat{\mu}_i \sim N(\mu, \sigma_{\bullet\bullet})$, where

$$\sigma_{\bullet\bullet}^2 = \sigma_b^2 + \sigma_\bullet^2/R, \quad \text{and } \sigma_\bullet^2 = \sigma_w^2 + \sigma^2/N. \tag{3.1}$$

Clearly, $\sigma_\bullet^2$ represents the variance of the wafer sample mean. The sim-

plest way to set up an upper scheme for detecting a change in μ upwards from the domain $\mu \leq \mu_0$ to $\mu \geq \mu_1$ is by fixing some historically prevalent and practically acceptable value of $\sigma_{\bullet\bullet}^2$ and designing the scheme under the assumption that the variance of $\{\hat{\mu}_i\}$ is equal to this value. In accordance with the usual Cusum convention, define $k_\mu = (\mu_0 + \mu_1)/2$ and let $\bar{\mu}_l = (1/l)\sum_{T-l+1}^{T} \hat{\mu}_i$. Then, it is not difficult to show that

$$D_l^* = \begin{cases} -l(\mu_1 - \bar{\mu}_l)^2/2\sigma_{\bullet\bullet}^2 & \text{if } \bar{\mu}_l \leq \mu_0 \\ -l(\mu_1 - \mu_0)(\bar{\mu}_l - k_\mu)/\sigma_{\bullet\bullet}^2 & \text{if } \mu_0 < \bar{\mu}_l \leq \mu_1 \\ l(\bar{\mu}_l - \mu_0)^2/2\sigma_{\bullet\bullet}^2 & \text{if } \bar{\mu}_l > \mu_1 \end{cases} \tag{3.2}$$

Since $D_l^* \leq 0$ when $\bar{\mu}_l \leq k_\mu$, the upper LR control scheme can be formulated as follows:

Upper LR scheme for μ: Trigger an out of control signal at time T if for some $l \geq 1$

$$\begin{gathered} k_\mu \leq \bar{\mu}_l \leq \mu_1 \quad \text{and} \quad l(\mu_1 - \mu_0)(\bar{\mu}_l - k_\mu)/\sigma_{\bullet\bullet}^2 > h_\mu \quad \text{or} \\ \bar{\mu}_l > \mu_1 \quad \text{and} \quad l(\bar{\mu}_l - \mu_0)^2/2\sigma_{\bullet\bullet}^2 > h_\mu \end{gathered} \tag{3.3}$$

In a similar way, one can define a lower scheme for detecting downward changes in μ by applying an upper scheme to reflected sequence $\{-\hat{\mu}_i\}$. A combination of two one-sided schemes then leads to a two-sided LR control scheme for μ.

Control of the within-item standard deviation, σ, is based on the control sequence $\hat{\sigma}_i^2$ defined by (2.2). Denote $\bar{\sigma}_l^2 = (1/l)\sum_{T-l+1}^{T} \hat{\sigma}_i^2$. Under the assumptions of the model (2.1), the density of this estimate based on the last l lots is given by

$$f_l(t \mid \sigma) = \frac{1}{\Gamma(lv/2)} \left(\frac{lv}{2\sigma^2}\right)^{lv/2} t^{(lv/2)-1} \exp\left(-\frac{lvt}{2\sigma^2}\right), \quad t > 0, \tag{3.4}$$

where $v = R(N-1)$, the number of degrees of freedom associated with a single lot. In most practical situations the acceptable and rejectable regions for σ correspond to $\sigma \leq \sigma_0$ and $\sigma \geq \sigma_1$, respectively, with $\sigma_0 \leq \sigma_1$, i.e. one is primarily interested in detecting changes upwards. In this case the score associated with the last l lots can be represented by

$$D_l^* = \begin{cases} 0.5lv\left[\ln(\bar{\sigma}_l^2/\sigma_1^2) + 1 - \bar{\sigma}_l^2/\sigma_1^2\right] & \text{if } \bar{\sigma}_l^2 \leq \sigma_0^2 \\ 0.5lv\left[\ln(\sigma_0^2/\sigma_1^2) + \bar{\sigma}_l^2(\sigma_0^{-2} - \sigma_1^{-2})\right] & \text{if } \sigma_0^2 \leq \bar{\sigma}_l^2 \leq \sigma_1^2 \\ 0.5lv\left[\ln(\sigma_0^2/\bar{\sigma}_l^2) - 1 + \bar{\sigma}_l^2/\sigma_0^2\right] & \text{if } \bar{\sigma}_l^2 > \sigma_1^2 \end{cases} \tag{3.5}$$

Denoting the reference value k_σ by

$$k_\sigma = \frac{2\ln(\sigma_1/\sigma_0)}{\sigma_0^{-2} - \sigma_1^{-2}} \tag{3.6}$$

and taking into account that the score is positive only if $\bar{\sigma}_l^2 > k_\sigma$, the LR scheme for monitoring σ can be formulated as follows:

Upper LR scheme for σ: Trigger an out of control signal at time T if for some $l \geq 1$

$$\begin{aligned} k_\sigma \leq \bar{\sigma}_l^2 \leq \sigma_1^2 \quad &\text{and} \quad 0.5lv(\sigma_0^{-2} - \sigma_1^{-2})(\bar{\sigma}_l^2 - k_\sigma) > h_\sigma \quad \text{or} \\ \bar{\sigma}_l^2 > \sigma_1^2 \quad &\text{and} \quad 0.5lv\left[\ln(\sigma_0^2/\bar{\sigma}_l^2) - 1 + \bar{\sigma}_l^2/\sigma_0^2\right] > h_\sigma \end{aligned} \tag{3.7}$$

As noted in the previous section, schemes for μ and σ based on the RLR strategy are equivalent to those obtained under the LR strategy, since the distributions of the sequences $\{\hat{\mu}_i\}$ and $\{\hat{\sigma}_i^2\}$ belong to the exponential family. The limited scope of the present article does not enable me to discuss the problem of computing the values h_μ and h_σ that assure a given low rate of false alarms. I will only mention the fact that, analogously to the case of usual Cusum schemes, design of two-sided LR procedures can be decomposed into separate design of the underlying upper and lower schemes. Specifically, the ARL and SDRL (standard deviation of run length) of a two-sided scheme can still be well approximated in terms of its one-sided counterparts by using the formulas (e.g. see Kemp (1961), Yashchin (1985))

$$\begin{aligned} 1/ARL &\simeq 1/ARL^+ + 1/ARL^- \\ (SDRL/ARL)^2 &\simeq (SDRL^+/ARL^+)^2 + (SDRL^-/ARL^-)^2 - 1 \end{aligned} \tag{3.8}$$

4. Control schemes for monitoring σ_w. For other components of variance, direct application of the LR or LRL strategies leads to more complex procedures, mainly because of the effect of the nuisance parameters. When monitoring σ_w, the relevant nuisance parameter is σ. For the sake of simplicity, we shall use instead the nuisance parameter $\eta = \sigma^2/N$ (η represents the part of the variance of the within-item average that is explained by the within-item variability. Monitoring of σ_w is based on the sequence of bivariate statistics, $\{\hat{\sigma}_{i\bullet}^2, \hat{\sigma}_i^2\}$. The components of this sequence are independent and distributed as $(\sigma_w^2 + \eta)V_1[v_1]$ and $\eta V_2[v_2]$, where $v_1 = R - 1$ and $v_2 = R(N-1)$ are degrees of freedom associated with $\hat{\sigma}_{i\bullet}^2$ and $\hat{\sigma}_i^2$, respectively, and $V_j[v]$ is a Chi-square random variable with v degrees of freedom divided by v.

At time T, the logarithm of the likelihood function based on the last l

lots can be represented as

$$L_l(\sigma_w, \eta \mid \hat{\sigma}_{i\bullet}, \hat{\eta}_i, \quad i = T-l+1, \ldots, T) \propto C - lv_1[\ln(\sigma_w^2 + \eta) + M_1/(\sigma_w^2 + \eta)] - lv_2[\ln \eta + M_2/\eta] \tag{4.1}$$

where C does not depend on the parameters, σ_w^2 and η, and

$$M_1 = \frac{1}{l} \sum_{i=T-l+1}^{T} \hat{\sigma}_{i\bullet}^2 \overset{\text{dist}}{=} (\sigma_w^2 + \eta) V_1[rv_1],$$
$$M_2 = \frac{1}{l} \sum_{i=T-l+1}^{T} \hat{\sigma}_i^2/N \overset{\text{dist}}{=} \eta V_2[rv_2]. \tag{4.2}$$

M_1 and M_2 are sufficient statistics related to σ_w and η. To form a score corresponding to the last l lots in a control scheme for σ_w, one must first find the maximal values of L corresponding to acceptable and rejectable areas, i.e. find

$$L_{l0}^* = \max_{\sigma_w \le \sigma_{w0}, \eta \ge 0} L_l(\sigma_w, \eta \mid M_1, M_2) \quad \text{and}$$
$$L_{l1}^* = \max_{\sigma_w \ge \sigma_{w1}, \eta \ge 0} L_l(\sigma_w, \eta \mid M_1, M_2). \tag{4.3}$$

The LR control scheme for σ_w is then derived in a usual way: a signal is triggered at time T if $L_{l1}^* - L_{l0}^* > h_w$ for some $l \ge 1$ and signal level h_w. The RLR procedure in this case is not equivalent to the LR procedure.

It turns out that finding L_{l0}^* and L_{l1}^* is not difficult, provided one is able to find the maximal value of $L_l(\sigma_w, \eta)$ for a *fixed* value of σ_w. This, in turn, is achieved by solving the equation

$$\frac{dL_l(\sigma_w, \eta)}{d\eta} \propto \frac{lv_1}{\sigma_w^2 + \eta}\left(\frac{M_1}{\sigma_w^2 + \eta} - 1\right) + \frac{lv_2}{\eta}\left(\frac{M_2}{\eta} - 1\right) = 0, \tag{4.4}$$

which leads to a cubic equation in η,

$$v_1(\sigma_w^2 + \eta)\eta^2 + v_2(\sigma_w^2 + \eta)^2\eta - M_1 v_1 \eta^2 - M_2 v_2(\sigma_w^2 + \eta)^2 = 0. \tag{4.5}$$

It is not difficult to see that a positive root $\eta^*(\sigma_w)$ of (4.5) in the interval between $\eta_1 = \max(M_1 - \sigma_w^2, 0)$ and $\eta_2 = M_2$ always exists. If such a root is unique in this interval (which can be checked by simply solving (4.5) using standard formulas), it represents the maximizing value of $L_l(\sigma_w, \eta)$. In these rare cases where the other two roots are also real and belong to the interval $[\eta_1, \eta_2]$, they must also be checked and $\eta^*(\sigma_w)$ is then selected to be the root that globally maximizes (4.1).

Now the problem of finding L_{l0}^* and L_{l1}^* is reduced to the problem of maximizing $L_l(\sigma_w, \eta^*(\sigma_w))$ in the areas $\sigma_w \le \sigma_{w0}$ and $\sigma_w \ge \sigma_{w1}$, respectively.

By setting the partial derivatives of $L_l(\sigma_w, \eta)$ by σ_w^2 and η to zero, as in (4.4), one concludes that the maximum likelihood estimators (MLE's) of these parameters are

$$\hat{\eta} = M_2, \quad \hat{\sigma}_w^2 = \max(M_1 - \hat{\eta}, 0) \tag{4.6}$$

Furthermore, one can show that the function $L_l(\sigma_w, \eta^*(\sigma_w))$ increases when $\sigma_w < \hat{\sigma}_w$, reaches its maximum at $\hat{\sigma}_w$ and decreases when $\sigma_w > \hat{\sigma}_w$. Therefore L_{l0}^* and L_{l1}^* are determined as follows:

a) If $\hat{\sigma}_w \leq \sigma_{w0}$ set $L_{l0}^* = L_l(\hat{\sigma}_w, \hat{\eta})$ and $L_{l1}^* = L_l(\sigma_{w1}, \eta^*(\sigma_{w1}))$.

b) If $\sigma_{w0} < \hat{\sigma}_w \leq \sigma_{w1}$ set $L_{l0}^* = L_l(\sigma_{w0}, \eta^*(\sigma_{w0}))$ and $L_{l1}^* = L_l(\sigma_{w1}, \eta^*(\sigma_{w1}))$.

c) If $\hat{\sigma}_w > \sigma_{w1}$ set $L_{l1}^* = L_l(\hat{\sigma}_w, \hat{\eta})$ and $L_{l0}^* = L_l(\sigma_{w0}, \eta^*(\sigma_{w0}))$.

Note that in the upper monitoring scheme only cases b) and c) are relevant, since in case a) the score D_l^* cannot be positive.

Table 4.1: The ARL's and standard deviations of the RL's (in parentheses) of the LR and Markovian Cusum procedures for monitoring σ_w. The acceptable and rejectable regions are $\sigma_w \leq 1$ and $\sigma_w \geq 2$, respectively.

σ^2/N	σ_w^2	$h = 7.6$ LR Procedure	$h = 19.1$ $k_w(1) = 2.05$
1	1	250 (270)	251 (248)
1	2	41.8 (40.8)	38.1 (34.6)
1	3	16.8 (16.1)	17.3 (14.5)
1	4	11.9 (10.8)	11.2 (8.9)
2	1	154 (153)	64.9 (62.5)
2	2	49.4 (46.5)	24.9 (22.7)
2	3	25.1 (24.2)	14.6 (12.7)
2	4	16.4 (14.9)	10.4 (8.7)

Now we shall compare the performance of the above procedure (applied in terms of the LR strategy) to that of an alternative Markovian Cusum scheme defined by computing recursively the process

$$\begin{aligned} S_w(0) = 0, S_w(i) = \max\{S_w(i-1) \\ + (\hat{\sigma}_{i\bullet}^2 - \hat{\sigma}_i^2/N - k_w(\tilde{\eta})), 0\} \quad i = 1, 2, \end{aligned} \tag{4.7}$$

and triggering a signal when $S_w(i) > h_w$. In the above process $\hat{\sigma}_{i\bullet}^2 - \hat{\sigma}_i^2/N$, $i = 1, 2$, are unbiased estimates of σ_w^2 and the reference value is chosen in accordance with the formula

$$k_w(\tilde{\eta}) = -\tilde{\eta} + \frac{\ln[(\sigma_{w1}^2 + \tilde{\eta}) / (\sigma_{w0}^2 + \tilde{\eta})]}{(\sigma_{w0}^2 + \tilde{\eta})^{-1} - (\sigma_{w1}^2 + \tilde{\eta})^{-1}} \tag{4.8}$$

where $\tilde{\eta}$ is some historically prevalent value of η. It is not difficult to see that when σ is known, the above procedure is the LR scheme corresponding to $\Omega_0 = \{\sigma_{w0}\}$ and $\Omega_1 = \{\sigma_{w1}\}$ (e.g. see Yashchin (1991)). To compare the schemes, consider the case with $R = 2, N = 4, \sigma^2_{w0} = 1, \sigma^2_{w1} = 4, \tilde{\eta} = 1$, and select the signal levels of both procedures so as to obtain the same ARL=250 when $\sigma_w = \sigma_{w0} = 1$. This requirement leads to signal levels 7.6 (LR scheme) and 19.1 (Markovian Cusum). By (4.8), the reference value of the latter is $k_w(1) = 2.05$. Table 4.1 contains the ARL's corresponding to cases when $\eta = 1$ (nominal value) and $\eta = 2$.

As can be seen from Table 4.1, the LR procedure is slightly less sensitive if η is at its nominal level. When η increases, however, both schemes experience degradation in performance. The LR procedure is much more robust in terms of the rate of false alarms. On the other hand, the Markovian Cusum is more robust in terms of sensitivity. The choice of which procedure to use should depend on the type of degradation that one is more willing to tolerate in a given application. The above example indicates that a control chart for monitoring σ_w should never be used alone: it should be accompanied by a control chart for detecting changes in η. Should a significant increase in η occur, this chart is likely to pick it up before it causes a false alarm in the chart for monitoring σ_w.

5.Control Schemes for Monitoring σ_b. Control of σ_b is based on the bivariate sequence, $\{\hat{\mu}_i, \hat{\sigma}^2_{i\bullet}\}$. The stochastic behavior of this sequence depends on a single nuisance parameter, $\eta_\bullet = \sigma^2_\bullet/R$ (see (3.1)). The decision whether to signal at time T on the basis of the information corresponding to the last l lots depends on the pair of sufficient statistics,

$$\begin{aligned} \widetilde{M}_1 &= \frac{1}{l-1} \sum_{i=T-l+1}^{T} (\hat{\mu}_i - \bar{\mu}_l)^2 \stackrel{\text{dist}}{=} (\sigma^2_b + \eta_\bullet) V_1[\tilde{v}_1] \\ \widetilde{M}_2 &= \frac{1}{l} \sum_{i=T-l+1}^{T} \hat{\sigma}^2_{i\bullet}/R \stackrel{\text{dist}}{=} \eta_\bullet V_2[\tilde{v}_2] \end{aligned} \tag{5.1}$$

where $\tilde{v}_1 = l - 1$ is the number of degrees of freedom associated with σ_b and $\tilde{v}_2 = l(R-1)$ is the number of degrees of freedom associated with $\eta_\bullet$. In this setting, the problem of controlling σ_b becomes similar to that of controlling σ_w in the sense that at each stage one is dealing with a single "lot" containing the last l groups of R averages and σ^2_b can be treated as an item-to-item variance within this artificial lot. To simplify the presentation, we shall use the letter ζ instead of $\eta_\bullet$ in what follows.

The logarithm of the likelihood function based on the last l lots is

$$L_l(\sigma_b, \zeta \mid \widetilde{M}_1, \widetilde{M}_2) \propto C - \tilde{v}_1[\ln(\sigma_b^2 + \zeta) + \widetilde{M}_1/(\sigma_b^2 + \zeta)] - \tilde{v}_2[\ln \zeta + \widetilde{M}_2/\zeta], \tag{5.2}$$

where C does not depend on the parameters. Next, one needs to find

$$L_{l0}^* = \max_{\sigma_b \le \sigma_{b0}, \zeta \ge 0} L_l(\sigma_b, \zeta \mid \widetilde{M}_1, \widetilde{M}_2) \quad \text{and}$$
$$L_{l1}^* = \max_{\sigma_b \ge \sigma_{b1}, \zeta \ge 0} L_l(\sigma_b, \zeta \mid \widetilde{M}_1, \widetilde{M}_2). \tag{5.3}$$

To find these values, note that to maximize $L_l(\sigma_b, \zeta)$ for a *fixed* value of σ_w, one needs to solve the cubic equation in ζ,

$$\tilde{v}_1(\sigma_b^2 + \zeta)\zeta^2 + \tilde{v}_2(\sigma_b^2 + \zeta)^2\zeta - \widetilde{M}_1\tilde{v}_1\zeta^2 - \widetilde{M}_2\tilde{v}_2(\sigma_b^2 + \zeta)^2 = 0. \tag{5.4}$$

This equation is very similar to (4.5) except that now $\tilde{v}_1$ and $\tilde{v}_2$ depend on l. Once again, a positive root $\zeta^*(\sigma_b)$ that maximizes (5.2) lies in the interval between $\zeta_1 = \max(\widetilde{M}_1 - \sigma_b^2, 0)$ and $\zeta_2 = \widetilde{M}_2$. Taking into account the fact that the MLE's of ζ and σ_b^2 are

$$\hat{\zeta} = \widetilde{M}_2, \hat{\sigma}_b^2 = \max(\widetilde{M}_1 - \hat{\zeta}, 0), \tag{5.5}$$

L_{l0}^* and L_{l1}^* are determined in the same way as in the case of σ_w, i.e.

a) If $\hat{\sigma}_b \le \sigma_{b0}$ set $L_{l0}^* = L_l(\hat{\sigma}_b, \hat{\zeta})$ and $L_{l1}^* = L_l(\sigma_{b1}, \zeta^*(\sigma_{b1}))$.

b) If $\sigma_{b0} < \hat{\sigma}_b \le \sigma_{b1}$ set $L_{l0}^* = L_l(\sigma_{b0}, \zeta^*(\sigma_{b0}))$ and $L_{l1}^* = L_l(\sigma_{b1}, \zeta^*(\sigma_{b1}))$.

c) If $\hat{\sigma}_b > \sigma_{b1}$ set $L_{l1}^* = L_l(\hat{\sigma}_b, \hat{\zeta})$ and $L_{l0}^* = L_l(\sigma_{b0}, \zeta^*(\sigma_{b0}))$.

Therefore, the LR control scheme for monitoring σ_b can be formulated as follows: Trigger a signal at time T if $L_{l1}^* - L_{l0}^* > h_b$ for some $l \ge 2$ and signal level h_b. As in the case with σ_w, only b) and c) are relevant in the upper LR (or RLR) procedures.

One important point should be made about procedures for monitoring σ_b discussed above. These procedures can only be recommended in cases where μ remains stable for prolonged periods of time. If μ tends to undergo abrupt changes without entering into the respective unacceptable region, these changes will be picked up by the scheme for σ_b, leading to incorrect diagnosis. Practical experience suggests that, since μ is frequently subject to abrupt changes and controlled adjustments, robustness of control schemes for variance components with respect to such events is of primary importance. Of course, schemes for σ_w and σ automatically possess this property, because they are based on control sequences that do not depend on μ. On the other hand, the

sequence $\{\hat{\mu}_i\}$ used in control of σ_b *does* depend on μ. To obtain a robust procedure, it is advisable to use

$$\widetilde{M}_1' = \frac{1}{2(l-1)} \sum_{i=T-l+1}^{T} (\hat{\mu}_i - \hat{\mu}_{i-1})^2 \quad \text{and} \quad \tilde{v}_1' = \frac{2(l-1)^2}{3l-4} \tag{5.6}$$

instead of $\widetilde{M}_1$ and $\tilde{v}_1$ in the above procedures. This is justified by the fact that $\widetilde{M}_1'$ is much more robust with respect to changes in μ and its distribution is well approximated by that of $(\sigma_b^2 + \eta_\bullet)V_1[\tilde{v}_1']$.

6. Conclusions. When it is reasonable to assume that the data originates from the model (2.1) or a more general model of this type, one can organize an efficient system for monitoring the underlying parameters by decomposing the data stream into control sequences associated with the grand process mean and individual components of variability and designing appropriate control schemes based on these sequences. There are several approaches to design of a such a system, among them the Cusum approach discussed in Yashchin (1991) and the LR/RLR methodology presented in this article. The latter approach has two major advantages. First of all, it delivers a high statistical power while using very few scheme parameters. Second, its degree of protection against false alarms is relatively robust with respect to possible fluctuations in the level of nuisance parameters. On the other hand, it has several drawbacks: design, analysis and implementation of LR/RLR strategy is fairly complex, the *form* of control schemes depends on the model (unlike the Cusum approach, which uses a standard form of a scheme) and their sensitivity is highly dependent on the levels of nuisance parameters. I feel that there are many situations, especially those involving simultaneous monitoring of many parameters, in which one should give this approach serious consideration. However, its practical success will largely depend on availability of software that automates the implementation and presents its graphical and statistical power to manufacturing engineers "under the hood."

REFERENCES

BANSAL, R. K. and PAPANTONI-KAZAKOS (1986). An Algorithm for Detecting a Change in a Stochastic Process, *IEEE Tran. Inform. Theor., IT-32, No. 2, pp. 227–235.*

BASSEVILLE, M. *(1988). Detecting Changes in Signals and Systems – A Survey, Automatica, Vol. 24(3), pp. 309–326.*

BASSEVILLE, M. and BENVENISTE, A. (ed.) (1986). Detection of Abrupt Changes in Signals and Dynamical Systems, *Lecture Notes in Control and Information Sciences*, Vol. 77, Springer–Verlag, Berlin.

KEMP, K. (1961). The Average Run Length of the Cumulative Sum Chart when a V-mask is used, *J. Roy. Statist. Soc. (B)* **23**, pp. 149–153.

LORDEN, G. (1971). Procedures for Reacting to a Change in Distribution, *Ann. Math. Statist.* **42**, pp. 1897–1908.

MOUSTAKIDES, G. V. (1986). Optimal Stopping Times for Detecting Changes in Distributions, *Ann. Statist.*, **14**, No. 2, pp. 1379–1387.

NIKIFOROV, I. V. (1983). Sequential Detection of Abrupt Changes in Time Series Properties, (in Russian), Nauka, Moscow.

TELKSNYS, L., ed. (1986). Detection of Changes in Random Processes, Optimization Software, Inc., New York.

WELLS, S. W., and SMITH, J. D. (1991). Making Control Charts Work for You, *Semiconductor International,* **14**, No. 10, pp. 86–89.

WOODALL, W. H. and THOMAS, E. V. (1991). Statistical Process Control with Several Components of Variability: A Review, to appear in *IEEE Trans.*

YASHCHIN, E. (1985). On a Unified Approach to the Analysis Of Two-sided Cumulative Sum Schemes With Headstarts, *Adv. Appl. Probab.* **17**, pp. 562–593.

YASHCHIN, E. (1991). Control of Variance Components, IBM Res. Rep. RC #17095, Yorktown Heights, NY.

MATHEMATICAL SCIENCES DEPARTMENT
IBM RESEARCH DIVISION
T. J. WATSON RESEARCH CENTER
YORKTOWN HEIGHTS, NY 10598