



Estimating the Current Mean of a Process Subject to Abrupt Changes

Author(s): Emmanuel Yashchin

Source: *Technometrics*, Vol. 37, No. 3, (Aug., 1995), pp. 311-323

Published by: American Statistical Association and American Society for Quality

Stable URL: <http://www.jstor.org/stable/1269915>

Accessed: 04/04/2008 15:25

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=astata>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We enable the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

# Estimating the Current Mean of a Process Subject to Abrupt Changes

Emmanuel YASHCHIN

Department of Mathematical Sciences  
IBM Corporation  
Thomas J. Watson Research Center  
Yorktown Heights, NY 10598

This article discusses estimation of the current process mean in situations in which this parameter is subject to abrupt changes of unpredictable magnitude at some unknown points in time. It introduces performance criteria for this estimation problem and discusses in detail the relative merits of several estimation procedures. I show that an estimate based on exponentially weighted moving average of past observations has optimality properties within the class of linear estimators, and I propose alternative estimating procedures to overcome its limitations. I consider two primary types of estimation procedures, Markovian estimators, in which the current estimate is obtained as a function of the previous estimate and the most current data point, and adaptive estimators, based on identification of the most recent changepoint. We give several examples that illustrate the use of the proposed techniques.

KEY WORDS: Changepoint; Control charts; Exponentially weighted moving average; Filtering; Process control.

The problem of estimating the current process mean in the presence of abrupt changes was considered by Barnard (1959), Chernoff and Zacks (1964), West (1986), Chen and Tiao (1990), Kenett and Zacks (1992), McCulloch and Tsay (1993), and others. Many of the proposed estimation procedures are Bayesian, and they require a priori assumptions about the location or magnitude of the changes. There exist several non-Bayesian procedures (e.g., see Wetherill 1977), but their properties are often unknown, and it is not clear how they should be assessed. This article introduces a criterion called *inertia* that can be used to compare non-Bayesian estimation procedures for data subject to abrupt changes and applies it to several known and new estimation procedures.

Section 1 describes the concept of inertia. Section 2 shows that exponentially weighted moving average (EWMA) estimators are optimal in the class of linear estimators. It also shows how EWMA estimators can be improved and introduces some more general Markovian procedures. Section 3 discusses adaptive schemes that are based on identifying the last changepoint and using the resulting last stable segment of data to obtain the estimate. Section 4 discusses implementation of the proposed estimation methods. Section 5 considers unknown nuisance scale parameters. Section 6 gives an application.

## 1. MOTIVATION, PERFORMANCE CRITERIA, AND THE BASIC APPROACH

Consider a sequence of independent observations  $\{X_i\}$ , where  $X_i$  has mean  $\mu_i$ , standard deviation  $\sigma_i$ , and cumulative distribution function  $F[(x - \mu_i)/\sigma_i]$ . For example,  $X_i$  may represent the number of defects found in

the  $i$ th produced item, the waiting time of the  $i$ th customer in a queue, the difference of the actual amount of product shipped in the  $i$ th month from the predicted amount, or output from an automatic feedback-forward control device measured at the  $i$ th period of time.

In applications of interest in this article, the means  $\mu_i$  may change abruptly. The changes are infrequent and their times and magnitudes are unpredictable, but it is important to know the current  $\mu_i$ . For example, in many statistical process control applications, the process/product parameters might be adjusted to improve quality if it is known by how much  $\mu_i$  deviates from its target value.

In this article, I assume that the sequence of means  $\{\mu_i\}$  is piecewise constant. Because I am interested in non-Bayesian approaches, I focus on estimates  $\hat{\mu}_i$  that depend only on  $X_1, \dots, X_i$  and not on any prior information about  $\mu_i$ . Further assume that there is a loss function  $L(\hat{\mu}_i, \mu_i)$  and the total loss incurred through period  $i$  is the sum of losses accrued for periods  $1, \dots, i$ . Throughout this article, I shall assume that  $L(\hat{\mu}_i, \mu_i)$  is a function of  $(\hat{\mu}_i - \mu_i)/\sigma_i$ .

The following example may help to clarify the model:

*Example 1.1.* Semiconductor wafers, which are disks about 20 centimeters in diameter, are often processed in lots of about 20 wafers. In one of the hundreds of manufacturing steps, a layer of metal is deposited on each wafer in a lot. Uniform metal thickness is crucial. Usually, the mean metal thickness  $\mu_i$  for a lot is stable across many lots of wafers, but variations in raw materials, maintenance, or other factors can cause abrupt shifts. When the current  $\mu_i$  is known, the shifts are less important because the process can easily be adjusted to bring it on target. The main problem, therefore, is to estimate  $\mu_i$  well.

Every lot contains three test wafers and one measurement of metal thickness is taken from each of them. Denote by  $X_{i1}, X_{i2},$  and  $X_{i3}$  the measurements corresponding to the  $i$ th lot. The average  $X_i = (X_{i1} + X_{i2} + X_{i3})/3$  is reported and used to obtain an updated mean estimate  $\widehat{\mu}_i$ . Suppose that the  $j$ th wafer of the  $i$ th lot is declared nonconforming if its thickness  $X_{ij}$  is more than  $\Delta$  from a target thickness and engineers want to control the average (across lots) fraction of nonconforming wafers. Suppose that  $X_{ij}$  are independent and normal with mean  $\mu_i$  and variance 1. As a result of the decision accepted in this stage, the process mean will be shifted away from the target by  $|\widehat{\mu}_i - \mu_i|$ , and one can define the loss function as the expected fraction of nonconforming wafers,

$$L(\widehat{\mu}_i, \mu_i) = 1 - \Phi(\Delta + \widehat{\mu}_i - \mu_i) + \Phi(-\Delta + \widehat{\mu}_i - \mu_i), \quad (1.1)$$

where  $\Phi$  is the standard normal cdf.

To simplify the presentation, at this point let us assume that  $\sigma_i \equiv \sigma$ , where  $\sigma$  is a known constant (in Sec. 5, I shall discuss more general situations). I shall look for the best sequences of estimators  $\{\widehat{\mu}_i\}$  that possess the following two properties: First, they must be *location equivariant*; that is,  $\widehat{\mu}_i(X_1 + c, \dots, X_i + c) \equiv \widehat{\mu}_i(X_1, \dots, X_i) + c$ , for any constant  $c$ . Second, if  $\mu_i \equiv \mu$  for all  $i > T$ , the distribution of  $\widehat{\mu}_i$  must converge to a limiting (steady-state) distribution that does not depend on  $X_1, X_2, \dots, X_T$ .

To obtain estimators that adapt quickly to abrupt changes, we must first decide on the comparison criteria. In the estimation process we have two sources of loss, (1) loss that occurs even when  $\mu_i$  does not change (*steady-state loss*) and (2) extra loss that occurs when  $\mu_i$  changes. After a change,  $\widehat{\mu}_i$  returns (more or less quickly) to its steady-state distribution. I shall call the extra loss *inertia*. There is a trade-off between steady-state loss and inertia, and my efforts will focus on minimizing inertia for a fixed level of steady-state loss. To put the problem in a more formal framework, suppose that  $\mu_i$  changes once somewhere between times  $T$  and  $T + 1$ . That is,  $\mu_i = \mu - \delta\sigma$  for  $i \leq T$  and  $\mu_i = \mu$  for  $i > T$ . Then  $E_{T,\delta}L(\widehat{\mu}_{T+j}, \mu)$  is the expected loss incurred in period  $T + j, j \geq 1$ . Now let us fix a finite number  $E_0$  and define the class  $LE(E_0)$  of estimators as follows:

*Definition 1.1.* The estimation process  $\{\widehat{\mu}_i\}$  belongs to the class  $LE(E_0)$  if it is location equivariant, and when  $\mu_i = \mu - \delta\sigma$  for  $i = 1, \dots, T < \infty$  and  $\mu_i = \mu$  for  $i > T$ , then  $\lim_{j \rightarrow \infty} E_{T,\delta}L(\widehat{\mu}_{T+j}, \mu) = E_0$ , for any  $T, \delta$ , and  $\mu$ .

If no changes in  $\mu$  occur after time  $T$ , any two sequences of estimators that belong to  $LE(E_0)$  are asymptotically equivalent in the sense that they eventually reach the same expected loss per period. Within the class  $LE(E_0)$ , I prefer the sequences of estimators that adapt most quickly to the

change in mean—that is, have the least excess loss after the change over some horizon of interest,  $H$ . After the change, the expected excess loss in period  $T + j (j \geq 1)$  is  $E_{T,\delta}L(\widehat{\mu}_{T+j}, \mu) - E_0$ .

*Definition 1.2.* Suppose that  $\mu_i = \mu - \delta\sigma$  for  $i \leq T$  and  $\mu_i = \mu$  for  $i > T$ . The inertia of a sequence of estimators  $\{\widehat{\mu}_i\}$  in the class  $LE(E_0)$  is defined by

$$I(\delta) = \lim_{T \rightarrow \infty} \sum_{j=1}^H [E_{T,\delta}L(\widehat{\mu}_{T+j}, \mu) - E_0]. \quad (1.2)$$

Note that  $I(\delta)$  implicitly depends on  $H, E_0, F,$  and  $L$ . For large  $H$  (in most practical cases 50 is large enough), the dependence of  $I(\delta)$  on  $H$  becomes negligible because sensible estimation procedures typically reach the steady state within 50 observations after the change. In this article we shall use the horizon  $H = \infty$ , though all the significant digits reported in numeric values remain unchanged for any  $H \geq 50$ . Similarly, values of  $T$  around 50 are typically high enough to achieve the limiting value  $I(\delta)$  in (1.2).

$I(\delta)$  can be interpreted as the excess loss caused by lack of information about a change: If  $\{\widehat{\mu}_i\}$  can detect the time  $T$  and magnitude  $\delta$  of a change immediately,  $I(\delta) = 0$ . Estimators that depend on more than just the current  $X_i$  do not have zero inertia. In Example 1.1,  $I(\delta)$  is the expected increase in the total number of nonconforming wafers after the change that can be attributed to the failure of the estimation scheme to anticipate the exact time and size of an abrupt change in  $\mu$ .

Now the problem of estimating  $\mu_i$  can be formulated in a simple way: Find a procedure in the class  $LE(E_0)$  that minimizes  $I(\delta)$  in some sense. For example, minimize  $I(\delta_0)$ , where  $\delta_0$  is a fixed “most likely” value of  $\delta$ ; the maximum value of  $I(\delta)$  (minimax procedure); or the mean  $E(I(\delta))$  computed under the assumption that  $\delta$  is random with a known distribution.

Minimizing over all kinds of estimators can be difficult, though. To simplify the problem, first limit the type of estimator and then optimize the curve  $I(\delta)$  that can be achieved for that type of estimator. To further simplify the problem, assume quadratic loss  $L(\widehat{\mu}_i, \mu_i) = A[(\widehat{\mu}_i - \mu_i)/\sigma]^2$ .

## 2. MARKOVIAN ESTIMATION PROCEDURES

Among the most popular procedures are those of linear type,

$$\widehat{\mu}_i = a_0X_i + a_1X_{i-1} + \dots + a_{n-1}X_{i-n+1}, \quad (2.1)$$

for some  $1 \leq n \leq \infty$ . For (2.1) to belong to  $LE(E_0)$ , it is necessary that  $a_0 + a_1 + \dots + a_{n-1} = 1$  and  $E_0 \leq A$ . One of the members of this class is the EWMA filter corresponding to  $a_i = (1 - \gamma)\gamma^i$  and  $n = \infty$ , alternatively

defined by means of the process

$$\begin{aligned} \hat{\mu}_0 &= \text{target} \\ \hat{\mu}_i &= X_i + \gamma(\hat{\mu}_{i-1} - X_i), \quad i = 1, 2, \dots, \end{aligned} \quad (2.2)$$

where  $0 \leq \gamma \leq 1$ . The following theorem shows that this procedure is optimal in its class.

*Theorem 2.1.* In the class of linear procedures belonging to  $LE(E_0)$ , the EWMA with  $\gamma = (A - E_0)/(A + E_0)$  has the smallest value of  $I(\delta)$ , for any  $\delta$ .

*Proof.* See Appendix A. One can see that the inertia function of the EWMA is given by

$$I(\delta) = (A\gamma^2\delta^2)/(1 - \gamma^2), \quad (2.3)$$

and, therefore, (2.2) is optimal in the class of linear schemes with average loss per observation *not exceeding*  $E_0$ . Note that there is only one EWMA that belongs to  $LE(E_0)$ .

In the presence of abrupt changes, Theorem 2.1 has a more limited use as compared to the optimality of the EWMA for processes not involving abrupt changes (e.g., see Bather 1963; Muth 1960). Clearly, the best estimating procedures cannot be linear if they are to adapt to changes of high magnitude. The reason is related to the fact that even for the best linear procedure that belongs to  $LE(E_0)$  with  $E_0 < A$  the inertia  $I(\delta)$  increases without limit as  $\delta \rightarrow \infty$ . Now we shall show how to modify the EWMA so as to obtain a scheme with bounded inertia. One way to achieve this goal is by switching to the family of Markovian procedures defined by means of the process

$$\begin{aligned} \hat{\mu}_0 &= \text{target} \\ \hat{\mu}_i &= X_i + \omega(\hat{\mu}_{i-1} - X_i), \quad i = 1, 2, \dots, \end{aligned} \quad (2.4)$$

where  $\omega(z)$  is some bounded function satisfying the conditions (a)  $\omega(0) = 0$ , (b)  $z\omega(z) \geq 0$ , and (c)  $|\omega(z)/z| \leq 1$ . The last two conditions assure that  $\hat{\mu}_i$  is always located between  $\hat{\mu}_{i-1}$  and  $X_i$ . In addition, I shall adhere to the general *principle of consistency* by which  $\hat{\mu}_i$  should not be allowed to be inconsistent with the last observation,  $X_i$  (in other words,  $X_i$  should never fall too far into the tail of  $F[(x - \hat{\mu}_i)/\sigma]$ ). One simple way to comply with this principle is by imposing an additional condition on  $\omega$ —namely, (d)  $-c^*\sigma \leq \omega(z) \leq c_*\sigma$ . In the Gaussian case, one might reasonably choose both  $c^*$  and  $c_*$  somewhere about 2.

The main attractive points of procedures of type (2.4) are their ease of use and simplicity of design and analysis (see App. B). One can see that  $\omega(z) = \gamma z$  corresponds to the EWMA procedure. Two other suggested functions of this type are as follows:

$$\begin{aligned} \omega_1(z) &= \gamma z, & -c^*\sigma/\gamma \leq z \leq c_*\sigma/\gamma \\ &= c_*\sigma, & z > c_*\sigma/\gamma \\ &= -c^*\sigma, & z < -c^*\sigma/\gamma \\ \omega_2(z) &= \gamma z \cdot \exp[-(z/\beta_*\sigma)^2/2], & z \geq 0 \\ &= \gamma z \cdot \exp[-(-z/\beta^*\sigma)^2/2], & z < 0. \end{aligned} \quad (2.5)$$

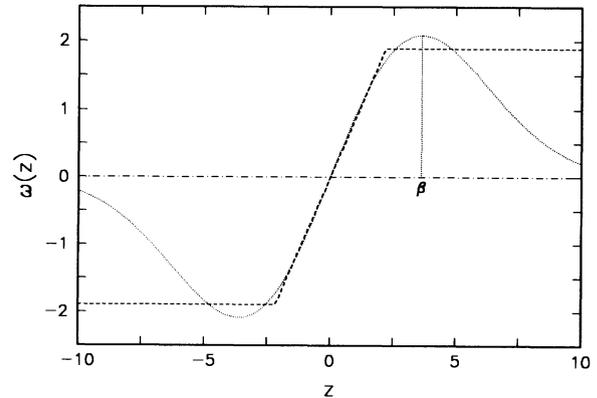


Figure 1. Values of  $\omega(z)$  Corresponding to Symmetric EWMA-C Procedure With Parameters ( $\gamma = .87, c = 1.89$ ) and Markovian Procedure With ( $\gamma = .95, \beta = 3.62$ ). When  $\{X_i\}$  are independent Normal, both procedures belong to the class  $LE(E_0 = 1/9)$  with respect to the quadratic loss function with  $A = 1$ .

Both functions are equivalent to  $\gamma z$  in the vicinity of 0. The parameters  $(c_*, c^*)$  and  $(\beta_*, \beta^*)$ , serve to establish the desired shape of  $I(\delta)$ , whereas  $\gamma$  is used to bring the procedure into the class  $LE(E_0)$ . Some plots of these functions are given in Figure 1. In what follows, I shall refer to Procedures (2.5) with  $c_* = c^* = c$  or  $\beta_* = \beta^* = \beta$  as *symmetric*.

The function  $\omega_1(z)$  can be used when one is interested in putting a bound on  $I(\delta)$  with minimal increase in inertia for moderate or small  $\delta$ . This function has the property that  $-c^*\sigma \leq \hat{\mu}_i - X_i \leq c_*\sigma$ . In this case,  $I(\delta)$  monotonically increases toward the limiting value as  $\delta \rightarrow \infty$ . In what follows, this procedure will be called EWMA-C.

The function  $\omega_2(z)$  enables one to further reduce  $I(\delta)$  as  $|\delta| \rightarrow \infty$  at the expense of increase in  $I(\delta)$  for moderate or small  $\delta$ . Because  $\omega_2(z)$  reaches its maximum (in the domain  $z > 0$ ) when  $z_* = \beta_*\sigma$ , the function  $I(\delta)$  typically reaches its maximum in the region  $\delta < 0$  when  $\delta \simeq -z_*/\sigma = -\beta_*$ . Analogously, the maximum of  $I(\delta)$  in the region  $\delta > 0$  is reached when  $\delta \simeq \beta^*$ .

Markovian procedures are appealing in situations in which most of the changes are of small or moderate magnitude. Their inertia for larger values of  $\delta$  can be strongly reduced (at the expense of higher inertia for smaller changes) by switching to procedures described in Section 3.

### 3. PROCEDURES INVOLVING ESTIMATION OF THE LAST STABLE RANGE

In this section I consider procedures based on estimating, at any point  $i$ , the value of  $r_i$ , which represents the number of observations since the last point of change. To be more precise, the estimate  $\hat{r}_i$  will specify the number of observations that are declared to be relevant when computing the current level of  $\mu$ . Therefore, we shall refer to  $\hat{r}_i$  as the *last stable range*. In the absence of any changes in  $\mu$ , the process  $\{\hat{r}_i\}$  will generally be stationary. The mean of

this process is selected so as to provide (together with the associated procedure for estimating  $\mu$ ) the desired level of  $E_0$  and shape of  $I(\delta)$ . Therefore, though one can interpret the value of  $\hat{r}_i$  as the estimated number of observations since the last point of change, it should be kept in mind that the characteristics of this estimate are determined so as to address a different problem—that is, to provide a suitable estimate of  $\mu_i$ .

### 3.1 Estimating the Last Stable Range

In general, the best way of identifying the last point of change involves identification of *all* points of change by analyzing the whole available set of data. Such a complete segmentation problem is, however, very complex, taking into account that the previous process history can be rather long. Furthermore, success in solving this problem depends, to a large extent, on the adequacy of the model over a long period of time—something that holds very rarely in practical applications. Therefore, procedures proposed will concentrate on the last point of change only. These procedures go from the current point back into history and stop in accordance with some given rule. At this point, all of the preceding observations are ignored and the last stable range  $\hat{r}_i$  is identified.

In connection with the problem of estimating the current process mean, the likelihood ratio procedure given here proves to be of special value. Because this procedure is also useful in more complex situations involving multivariate data, autocorrelated data, or vector parameters, it will be described in a general form. Denote

$$M(n; i) = \max_{\mu_0, 1 \leq r \leq n-1, \mu_1} \log f(X_i, X_{i-1}, \dots, X_{i-n+1} | X_{i-n}, \dots, X_1; \mu_0, r, \mu_1), \quad (3.1)$$

where  $f$  is the conditional density of the last  $n$  observations computed under the assumption that the last  $r$  observations correspond to the parameter  $\mu_0$  and all of the preceding observations correspond to the parameter  $\mu_1$ . The value of  $r$  for which the maximum is achieved is denoted by  $r(n; i)$ . Similarly, denote

$$M_0(n; i) = \max_{\mu} \log f(X_i, X_{i-1}, \dots, X_{i-n+1} | X_{i-n}, \dots, X_1; \mu), \quad (3.2)$$

where  $f$  is the conditional density of the last  $n$  observations computed under the assumption that all of the observations correspond to the parameter  $\mu$ . Now the likelihood ratio procedure for obtaining  $\hat{r}_i$  can be formulated as follows:

*Procedure A.* Select a positive threshold  $h$  and, for  $n = 2, 3, \dots$ , compute

$$d(n; i) = M(n; i) - M_0(n; i), \quad (3.3)$$

until for the first time  $d(n; i) > h$ . At this point, set  $\hat{r}_i = r(n; i)$ . If  $n = i$  is reached and  $d(i, i) \leq h$ , set  $\hat{r}_i = i$ .

For the sake of simplicity, only the case in which  $\{X_i\}$  is independent Gaussian with known  $\sigma$  will be considered in detail. In this case

$$2\sigma^2 M(n; i) = -n\sigma^2 \log(2\pi\sigma^2) - \max_{1 \leq r \leq n-1} \left\{ \sum_{j=1}^r [X_{i-j+1} - \bar{X}_i(r)]^2 + \sum_{j=r+1}^n [X_{i-j+1} - \bar{X}_{i-r}(n-r)]^2 \right\}, \quad (3.4)$$

where  $\bar{X}_i(r) = r^{-1} \sum_{j=1}^r X_{i-j+1}$ . Similarly,

$$2\sigma^2 M_0(n; i) = -n\sigma^2 \log(2\pi\sigma^2) - \sum_{j=1}^n [X_{i-j+1} - \bar{X}_i(n)]^2 \quad (3.5)$$

and the maximal discrepancy  $d(n; i)$  to be used in Procedure A becomes

$$d(n; i) = \max_{1 \leq r \leq n-1} \frac{r(n-r)}{2n\sigma^2} [\bar{X}_i(r) - \bar{X}_{i-r}(n-r)]^2 \quad (3.6)$$

Of course, Procedure A is just one of many that can be suggested for identifying the last stable range. My choice of this procedure is based not only on its simplicity and power of the resulting estimates of  $\mu$ , but also on formal comparisons against some other procedures of this kind. Criteria for such comparisons were described by Yashchin (1992).

### 3.2 Adaptive Estimation of the Current Mean

Once the estimate  $\hat{r}_i$  of the last stable range is available, one can proceed to obtain  $\hat{\mu}_i$ . I suggest the following scheme:

*Adaptive Exponentially Weighted (AEW) Estimation Scheme.* Select the scheme parameters ( $h \geq 0, 0 < \gamma \leq 1$ ) and, for every point  $i$ , compute the estimate  $\hat{\mu}_i$  as follows:

Step 1: Find an estimate  $\hat{r}_i$  of the last stable range by using Procedure A.

Step 2: If  $\hat{r}_i = \hat{r}_{i-1} + 1$ , then set

$$\hat{\mu}_i = \frac{\gamma(1 - \gamma^{\hat{r}_i - 1})}{1 - \gamma^{\hat{r}_i}} \hat{\mu}_{i-1} + \frac{1 - \gamma}{1 - \gamma^{\hat{r}_i}} X_i. \quad (3.7)$$

Otherwise, set

$$\hat{\mu}_i = \frac{(X_i + \gamma X_{i-1} + \dots + \gamma^{\hat{r}_i - 1} X_{i - \hat{r}_i + 1})}{(1 + \gamma + \dots + \gamma^{\hat{r}_i - 1})}. \quad (3.8)$$

In what follows, the preceding procedure will be referred to as the AEW( $h, \gamma$ ) estimation scheme. The process of selecting a suitable pair ( $h, \gamma$ ) essentially represents a one-dimensional search because only the pairs corresponding to the chosen value of  $E_0$  are acceptable. Note that the AEW( $h = \infty, \gamma$ ) reduces to the EWMA scheme with parameter  $\gamma$ . One can see that if  $\hat{r}_i = \hat{r}_{i-1} + 1$  then (3.7) is simply (3.8) presented in a recursive way. In general, when changes in the mean of  $\{X_i\}$  are not very

frequent, one will typically observe that  $\hat{r}_i = \hat{r}_{i-1} + 1$ , which leads to (3.7) being used most of the time.

4. COMPARISONS OF AEW AND MARKOVIAN PROCEDURES

To illustrate the performance of the AEW and Markovian procedures, some values of  $I(\delta)$  computed for the case in which  $\{X_i\}$  forms an independent Normal sequence with  $\sigma = 1$  are presented in Table 1. The loss function is assumed to be quadratic with  $A = 1$ . Every considered procedure belongs to the class LE(1/9). In other words, when  $\mu$  is fixed, all of these procedures are "equivalent" to the EWMA scheme with  $\gamma = .8$  because their steady-state variances are all equal to  $\sigma^2/9$ .

The class of EWMA-C procedures is obtained by increasing  $\gamma$  from .8 and simultaneously decreasing the values of  $c^*$  and  $c_*$  (to simplify the presentation, only the case  $c^* = c_* = c$  is considered). Table 1 shows that, as  $\gamma$  grows, the inertia decreases for large values of  $\delta$  at the expense of higher inertia for small  $\delta$ . This trend continues until  $\gamma = .87$ , which corresponds to a minimax procedure within the EWMA-C family with the largest possible inertia equal to 10.5. For all the EWMA-C procedures with  $\gamma > .87$ , the corresponding values of  $I(\delta)$  are uniformly worse compared to the scheme with  $\gamma = .87$  (in decision theory, such procedures are called *inadmissible*). Another

interesting phenomenon worth mentioning is that, as  $\gamma$  increases to .90, the corresponding value of  $c$  needed to achieve  $E_0 = 1/9$  decreases to 1.86. As  $\gamma$  continues to grow larger and reaches 1, however, the corresponding  $c$  increases to 2.25. This phenomenon can be explained as follows: For a fixed value of  $c$ , the variance of  $\{\hat{\mu}_i\}$  is *not* a monotonically decreasing function in  $\gamma$ . This variance initially decreases in  $\gamma$  because most of it is related to moderate variation in the preceding values of the observations. The impact of observations in the recent past that strongly deviated from the preceding ones is "washed out" by subsequent geometric averaging. As  $\gamma$  approaches 1, however, the impact of such extreme observations in the recent past becomes much more pronounced, which causes an increase in the variance of  $\{\hat{\mu}_i\}$ .

The Markovian procedures based on  $\omega_2(z)$  with  $.95 \leq \gamma \leq 1$  are slightly worse than the EWMA-C procedures with  $.85 \leq \gamma \leq .87$  for changes of magnitudes up to  $5\sigma$  but tend to be progressively better for changes of larger magnitudes. Note that the procedure with  $\gamma = 1$  turns out to be inadmissible: Its inertia is uniformly worse than that of the AEW scheme with  $\gamma = .82$ . Finally, the AEW schemes are definitely best for changes exceeding  $2\sigma$ . For smaller changes, one can find a Markovian scheme that has a lower inertia. The AEW ( $\gamma = .85$ ) is

Table 1. Values of Inertia  $I(\delta)$  Corresponding to EWMA, Symmetric EWMA-C Procedures, Symmetric Markovian Procedures Based on  $\omega_2(z)$  Given by (2.5), and AEW Procedures

Parameters		$\delta$								
		.5	1	1.5	2	3	4	5	6	7
$\gamma$	$c$	EWMA-C								
.80	$\infty$	.44	1.78	4.00	7.11	16.0	28.4	44.4	64.0	87.1
.85	1.95	.59	2.22	4.46	6.73	9.68	10.6	10.8	10.8	10.8
.87	1.89	.66	2.44	4.77	6.97	9.58	10.3	10.5	10.5	10.5
.90	1.86	.81	2.89	5.44	7.64	10.0	10.7	10.8	10.8	10.8
.95	1.91	1.31	4.36	7.58	10.0	12.4	13.0	13.1	13.2	13.2
$\gamma$	$\beta$	Markovian with $\omega_2(z)$								
.80	$\infty$	.44	1.78	4.00	7.11	16.0	28.4	44.4	64.0	87.1
.85	5.62	.48	1.85	3.92	6.42	11.7	16.0	18.6	19.3	18.4
.90	4.20	.55	2.06	4.18	6.47	10.4	12.4	12.6	11.4	9.46
.95	3.62	.66	2.42	4.71	6.96	10.2	11.2	10.5	8.91	7.12
1.00	3.36	.88	3.09	5.73	8.12	11.1	11.6	10.5	8.68	6.88
$\gamma$	$h$	AEW								
.80	$\infty$	.44	1.78	4.00	7.11	16.0	28.4	44.4	64.0	87.1
.81	8.45	.58	2.56	5.14	7.21	8.80	7.73	4.80	2.83	2.23
.82	7.52	.75	2.94	5.40	7.17	8.18	6.60	3.98	2.58	2.18
.83	7.05	.84	3.21	5.56	7.19	7.90	6.15	3.65	2.49	2.18
.85	6.41	1.07	3.70	5.93	7.21	7.40	5.56	3.34	2.39	2.17
.90	5.73	1.68	4.83	6.86	7.68	7.20	5.12	3.10	2.44	2.26
.95	5.33	2.58	6.25	7.82	8.22	7.30	5.07	3.20	2.57	2.42
1.00	5.22	4.27	8.06	8.84	8.83	7.51	5.17	3.30	2.67	2.51

NOTE: All of the estimation procedures belong to the class LE( $E_0 = 1/9$ ) with respect to the quadratic loss function, with  $A = 1$ . The observations  $\{X_i\}$  are independent Normal. The AEW values were obtained by simulation (the standard errors of the entries do not exceed 1%).

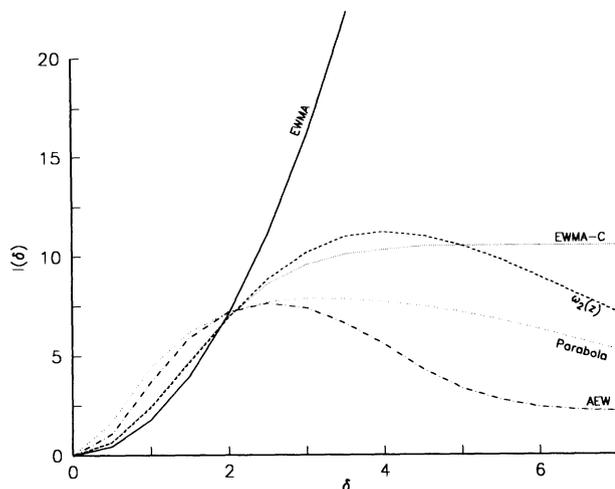


Figure 2. Values of Inertia  $I(\delta)$  Corresponding to EWMA, Symmetric EWMA-C Procedure With Parameters ( $\gamma = .87, c = 1.89$ ), Markovian Procedures Based on  $\omega_2(z)$  Given by (2.5) with ( $\gamma = .95, \beta = 3.62$ ), AEW Procedure with ( $\gamma = .85, h = 6.41$ ), and the Parabola Procedure. All of the estimation procedures belong to the class  $LE(E_0 = 1/9)$  with respect to the quadratic loss function, with  $A = 1$ . The observations  $\{X_i\}$  are independent Normal with a known  $\sigma$ .

a minimax procedure: It assures that for a change of any magnitude the inertia will not exceed 7.89—something no other scheme (among those considered) can assure. Note that the AEW schemes with  $\gamma \geq .95$  should definitely be avoided because their inertia is worse compared to AEW schemes with  $\gamma \leq .9$ .

In Figure 2 the plots of  $I(\delta)$  corresponding to selected procedures of various types are shown. This figure includes the curve corresponding to the parabola method discussed by Barnard (1959), Woodward and Goldsmith (1964), Wetherill (1977), and others. As can be seen from the figure, the method performs uniformly worse compared to the AEW ( $\gamma = .85$ ), and thus it is hardly suitable for practical use. The main problem with the parabola method is related to the fact that it takes too many points from the previous regime into consideration.

#### 4.1 Some Guidelines

Based on considerations similar to those described in the foregoing paragraphs, I have developed some informal guidelines for choosing a suitable estimation scheme:

1. Use EWMA only in cases in which efficient estimation of  $\mu$  following a change of moderate or large magnitude is not important. Such situations typically occur when a large change is likely to lead to external intervention—for example, because of direct damage or in wake of its detection by a monitoring system.

2. To obtain a scheme for which the largest possible inertia is smallest (i.e., a minimax scheme), always use the AEW approach. Also use the AEW approach when it is especially important to obtain an efficient estimate of  $\mu$  following a change of moderate or large magnitude.

3. Use EWMA-C when small changes are of primary concern but it is also desirable to limit the exposure with respect to larger changes.

4. Use a Markovian scheme based on  $\omega_2(z)$  when small changes are of primary concern but it is also important to have a low inertia with respect to very large changes.

5. Once the type of procedure is selected, examine the curves  $I(\delta)$  corresponding to values  $\gamma_0 \leq \gamma < 1$ , where  $\gamma_0 = (a - E_0)/(a + E_0)$ —that is, the smoothing parameter of the EWMA (see Theorem 2.1). As  $\gamma$  increases, the inertia gets lower for intermediate and large changes at the expense of small changes. Stop when a scheme with suitable  $I(\delta)$  is identified.

The preceding policy contains some fuzzy notions like “small” or “large” that need to be clarified with the context of a specific situation. For example, consider the situation involving independent Normal observations and suppose that a scheme that produces (under stable conditions), estimates  $\{\hat{\mu}_i\}$  with standard deviation  $\sigma/3$ , can be tolerated. Then one is facing the problem of finding a suitable scheme in the class  $LE(1/9)$ . As can be seen from Figure 2, in this setting a change of magnitude below  $2\sigma$  can be classified as “small” and changes above  $4\sigma$  as “large.”

#### 4.2 Example: The Role of Transformations

Consider the following situation arising in the production process of integrated circuit chips from semiconductor wafers. As noted in Example 1.1, wafers are usually processed lot by lot, with about 20 wafers per lot. At all times, the surface of each wafer must be kept exceptionally clean. Therefore, the process steps are usually followed by placing the wafer lots into a rinser/dryer, where they are rinsed by deionized filtered water. The water droplets are then spun off in the drying phase. In this process it is important to remove contaminating particles exceeding a certain size from the surface of the wafers. Before a lot is placed in the rinser/dryer, its state is assessed on the basis of a particle count on one of its wafers. These counts are used to monitor the level of contamination in the preceding station. In general, contamination undergoes various regimes depending on the state of equipment and other factors—and these regimes usually change in an abrupt fashion. The particle counts recorded before the rinser/dryer stage are important for providing feedback on the environment—but their immediate use is to determine how long the lot should be processed (the “dirtier” the process is, the longer it takes to wash the lots). For every level  $\mu$  of contamination, the optimal processing time is known, and it can be read off the operator’s nomogram. The main problem, therefore, is to estimate  $\mu$  representing the present regime on the basis of incoming data.

Suppose that under normal conditions the number of contaminating particles can be described by a Poisson random variable with mean about six particles per wafer

and it is desired to derive an estimation procedure with the purpose of setting the rinsing/dryer processing time. Clearly, the process mean is no longer a location parameter of the population of interest, because the shape of the Poisson distribution changes along with its mean. The transformed process,

$$X_i = 2(Y_i + .375)^{1/2}, \quad i = 1, 2, \dots, \quad (4.1)$$

where  $\{Y_i\}$  is the original process of Poisson counts corresponding to successive test wafers, does, however, have a variance very close to 1 for any mean of  $Y_i$  greater than 1. Thus, the problem becomes how to estimate the current process means  $\{\mu_i\}$  of  $\{X_i\}$ , where the transformed distributions form (approximately) a location family (e.g., see Johnson and Kotz 1969).

On the basis of previous experience and simulation studies, the process engineer is content to work with estimation procedures that under stable conditions produce unbiased estimates  $\{\hat{\mu}_i\}$  having a variance of  $1/9$ . The problem, therefore, is to identify a procedure in the class  $LE(1/9)$  that has a suitable inertia curve  $I(\delta)$ . Though most of the changes are anticipated to be of magnitude not exceeding 2, the engineer would like to limit his exposure with respect to changes of very large magnitude. Therefore, the EWMA-C scheme with  $\gamma = .85$  and  $c^* = c_* = 1.95$  appears quite appropriate (the inertia curve corresponding to this scheme is given in Table 1). In Figure 3 this scheme is applied to a sequence of counts of contaminating particles [note that the  $y$ -scale of the plot has been transformed in accordance with (4.1)]. For the sake of comparison, we also apply the EWMA scheme with  $\gamma = .8$  and the AEW scheme with  $(\gamma = .85, h = 6.41)$ . The plot shows two changes in the level of  $\{Y_i\}$ , at points 20 and 41. The EWMA-C procedure adapts to the second change much faster than the usual EWMA. The AEW adapts better to the first change. It detects very quickly the presence of the second change but overreacts to the point 42.

Variance-stabilizing transforms exist for many other families of distributions. After such a transform, the proposed framework of analysis becomes relevant again, provided that our basic assumptions related to the properties of loss functions (as outlined in Sec. 1) are acceptable for the transformed data. A more elaborate discussion on this subject was given by Yashchin (1992).

### 5. ESTIMATION OF THE MEAN WHEN THE VARIANCE IS UNKNOWN

In many practical situations, the scale parameter  $\sigma$  of the process also needs to be estimated and monitored on a periodic basis. In such situations, using a scheme  $\{\hat{\mu}_i\}$  derived under the assumption that  $\sigma$  is at some fixed level (when in fact it is at a higher level) can lead to unnecessary corrective actions and introduce a condition known in the area of process control as "hunting." In this section we consider schemes that are better suited to keeping the variance of  $\{\hat{\mu}_i\}$  under control and thus help to avoid

1	7	11	5	21	9	31	11	41	12	51	6
2	4	12	7	22	8	32	13	42	2	52	4
3	9	13	5	23	8	33	9	43	4	53	6
4	9	14	7	24	8	34	11	44	7	54	6
5	2	15	3	25	13	35	13	45	2	55	8
6	10	16	4	26	10	36	15	46	4	56	5
7	3	17	8	27	6	37	6	47	7	57	6
8	6	18	4	28	10	38	10	48	6	58	9
9	6	19	5	29	11	39	11	49	7	59	3
10	5	20	5	30	3	40	12	50	4	60	6

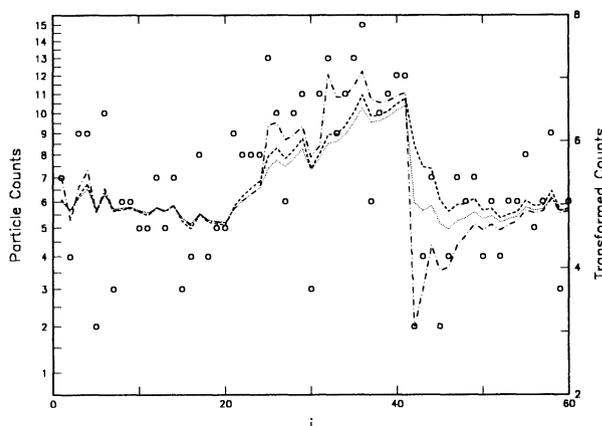


Figure 3. EWMA (with  $\gamma = .8$ , dashed line), EWMA-C (with  $\gamma = .85, c^* = c_* = 1.95$ , dotted line), and AEW (with  $\gamma = .85, h = 6.41$ , dashed-dotted line) Estimating Procedures Applied to Counts of Contaminating Particles Found on Successive Wafers Transformed in Accordance With (4.1). The raw data is given in the table above the plot. Both procedures belong to the class  $LE(E_0 = 1/9)$ .

this condition. We shall assume that  $\{X_i\}$  are independent and distributed in accordance with the cdf  $F[(x - \mu_i)/\sigma_i]$  and that  $L(\hat{\mu}_i, \mu_i)$  is a function of  $(\hat{\mu}_i - \mu_i)/\sigma_i$ . Under such conditions, it is natural to demand that any "good" estimation scheme be location-scale equivariant; that is,  $\hat{\mu}_i(bX_i + c, bX_{i-1} + c, \dots) \equiv b\hat{\mu}_i(X_i, X_{i-1}, \dots) + c$ , for any  $i, b$ , and  $c$  (see Sec. 1). To compare various estimation schemes, we restrict our attention to a class of schemes that produce, under stable conditions, the same expected loss per unit time:

**Definition 5.1.** The estimation process  $\{\hat{\mu}_i\}$  belongs to the class  $LSE(E_0)$  if it is location-scale equivariant, and when  $\sigma_i = \sigma$  for all  $i, \mu_i = \mu - \delta\sigma$  for  $i = 1, \dots, T < \infty$ , and  $\mu_i = \mu$  for  $i > T$ , then  $\lim_{j \rightarrow \infty} E_{T,\delta} L(\hat{\mu}_{T+j}, \mu) = E_0$ , for any  $T, \delta, \mu$ , and  $\sigma$ .

The inertia function  $I(\delta)$  for such procedures is also defined by (1.2).

The  $LSE(E_0)$  class of schemes can be simply constructed on the basis of procedures considered in the previous sections. First of all, because any linear procedure (2.1) belongs to  $LSE(E_0)$ , Theorem 2.1 continues to

hold for this class. Furthermore, the class of Markovian procedures can be generalized by substituting any scale-equivariant estimator  $\hat{\sigma}_{i-1}$  for  $\sigma$  into the approximately chosen function  $\omega(z|\sigma)$  in (2.5). In cases in which changes in  $\sigma$  are of relatively low magnitude, the results of the previous sections suggest the following estimator of the current level of  $\sigma^2$ :

$$\begin{aligned} \hat{\sigma}_0^2 &= \hat{\sigma}_1^2 = \sigma_0^2 \\ \hat{\sigma}_i^2 &= \min \{ c_\sigma \hat{\sigma}_{i-1}^2, \gamma_\sigma \hat{\sigma}_{i-1}^2 + (1 - \gamma_\sigma)(X_i - X_{i-1})^2/2 \}, \end{aligned} \tag{5.1}$$

where  $\sigma_0$  is some initial value of  $\sigma$ . Because  $(X_i - X_{i-1})^2/2$  is an unbiased estimate of  $\sigma^2$ , the sequence  $\{\hat{\sigma}_i^2\}$  is a scale-equivariant estimating scheme that is robust with respect to possible changes in the process mean. The value of  $\gamma_\sigma < 1$  is chosen close to 1, depending on the relevant range of changes in  $\sigma$ . By choosing  $\gamma_\sigma = 1$ , we would effectively make the assumption that  $\sigma = \sigma_0$  and end up with the  $LE(E_0)$  class. The constant  $c_\sigma$  is chosen somewhere above 1, depending on the estimated standard deviation of  $\hat{\sigma}_{i-1}$ . The main rationale for using a finite value of  $c_\sigma$  is to prevent the estimate  $\hat{\sigma}_i^2$  of the current variance from being too strongly affected by a possible abrupt change in  $\mu$ . Because for the Normal case

$$\text{var}(\hat{\sigma}_i^2/\sigma^2 | c_\sigma = \infty) = (1 - \gamma_\sigma)(2 + \gamma_\sigma)/(1 + \gamma_\sigma), \tag{5.2}$$

the choice

$$c_\sigma \simeq 1 + [(1 - \gamma_\sigma)(2 + \gamma_\sigma)/(1 + \gamma_\sigma)]^{1/2} \tag{5.3}$$

ensures that, even in the case of a very strong change in  $\mu$ , its effect on the estimated current variance will never exceed one standard deviation.

Substituting (5.1) into (2.5) and using any of the resulting functions  $\omega_1(z|\hat{\sigma}_{i-1}^2)$  or  $\omega_2(z|\hat{\sigma}_{i-1}^2)$  in (2.5) yields a location-scale equivariant joint estimation scheme  $\{\hat{\mu}_i, \hat{\sigma}_i^2\}$ . Unfortunately, it is no longer easy to compute  $E_0$  and  $I(\delta)$  numerically because  $\{\hat{\mu}_i\}$  is no longer a Markov chain. This computation, however, can be done by using Monte Carlo methods. A suitable Markovian procedure that belongs to  $LSE(E_0)$  can usually be obtained by first selecting an appropriate  $LE(E_0)$  scheme under the assumption that  $\sigma = \sigma_0$ . Then the value  $\gamma_\sigma$  (typically in the range between .9 and 1) is selected on the basis of anticipated fluctuations in  $\sigma$ : The smaller the anticipated changes in  $\sigma$ , the closer  $\gamma_\sigma$  can be chosen to 1. Finally,  $c_\sigma$  is computed in accordance with (5.3) and the parameters of the original  $LE(E_0)$  scheme are adjusted so as to obtain an expected loss per unit time  $E_0$  for the final LSE procedure. Table 2 gives the values of  $I(\delta)$  corresponding to the symmetric Markovian procedure with  $\omega_2(z|\hat{\sigma}_{i-1}) = \gamma z \bullet \exp[-(z|\beta\hat{\sigma}_{i-1})^2/2]$  from  $LSE(1/9)$ . The observations are assumed to be independent normal with standard deviation  $\sigma$ , and the values of  $\hat{\sigma}_{i-1}$  are obtained by using Process (5.1) with parameters  $\gamma_\sigma = .97$  and  $c_\sigma = 1.2$ .

Once again, the LSE procedures obtained on the basis of (2.4) only extend the tracking ability of the EWMA to larger values of  $\delta$  and limit the inertia for very large  $\delta$ . In cases in which further reduction of  $I(\delta)$  with respect to moderate and large  $\delta$  is desired, however, one should

Table 2. Values of Inertia  $I(\delta)$  Corresponding to Location-Scale Equivariant Procedures That Belong to  $LSE(E_0 = 1/9)$

Parameters		$\delta$								
		.5	1	1.5	2	3	4	5	6	7
		<i>Markovian with <math>\omega_2(z)</math></i>								
$\gamma$	$\beta$									
.80	$\infty$	.44	1.78	4.00	7.11	16.0	28.4	44.4	64.0	87.1
.85	5.80	.49	1.87	3.95	6.47	11.9	16.6	19.8	21.1	20.5
.90	4.34	.57	2.11	4.24	6.58	10.7	13.1	13.8	12.8	10.9
.95	3.77	.71	2.52	4.86	7.21	10.7	12.2	11.9	10.5	8.50
1.00	3.52	.94	3.25	6.03	8.59	12.0	13.1	12.4	10.7	8.60
		<i>AEW</i>								
$\gamma$	$h$									
.80	$\infty$	.44	1.78	4.00	7.11	16.0	28.4	44.4	64.0	87.1
.81	9.00	.60	2.51	5.10	7.29	9.15	8.32	5.53	3.17	2.35
.82	8.05	.72	2.81	5.36	7.26	8.48	7.26	4.58	2.82	2.21
.83	7.50	.82	3.11	5.59	7.30	8.13	6.71	4.17	2.72	2.20
.85	6.78	1.04	3.61	5.98	7.38	7.73	6.03	3.64	2.56	2.18
.90	5.93	1.64	4.81	6.94	7.85	7.40	5.52	3.34	2.50	2.26
.95	5.53	2.60	6.40	8.10	8.52	7.60	5.41	3.39	2.62	2.45
1.00	5.40	4.30	8.46	9.28	9.21	7.85	5.52	3.46	2.72	2.55

NOTE: The top part of the table corresponds to symmetric Markovian schemes based on  $\omega_2(z)$ . The bottom part represents the AEW schemes. The observations  $\{X_i\}$  are independent Normal. An entry corresponding to  $\delta$  gives the inertia resulting from an abrupt shift in  $\mu$  by  $\delta\sigma$ . The values of  $\hat{\sigma}_{i-1}$  used in the procedures at point  $i$  are obtained by using the Process (5.1) with parameters  $\gamma_\sigma = .97$  and  $c_\sigma = 1.2$ . The values were obtained by simulation (the standard errors of the entries do not exceed 1%).

use the AEW approach. To obtain an LSE( $E_0$ ) version of the AEW estimation scheme, one can follow the steps outlined previously—that is, by starting with a satisfactory AEW( $h, \gamma$ ) scheme from LE( $E_0$ ), computed under the assumption that  $\sigma = \sigma_0$ . Then a suitable value of  $\gamma_\sigma$  is selected and the corresponding  $c_\sigma$  is computed by using (5.3). Finally, the values  $\hat{\sigma}_{i-1}$  as defined by (5.1) are substituted for  $\sigma$  in this procedure and the value of  $h$  is adjusted so as to assure that the expected loss per observation (under stable conditions) is  $E_0$ .

The values of  $I(\delta)$  corresponding to several AEW schemes useful in practical applications are given in Table 2. The values of  $\hat{\sigma}_{i-1}$  used in (3.6) at point  $i$  are obtained by using the process (5.1) with parameters  $\gamma_\sigma = .97$  and  $c_\sigma = 1.2$ . As can be seen from this table, the fact that  $\sigma$  needs to be estimated means that a higher value of  $h$  is needed to obtain an LSE(1/9) procedure. Accordingly, the values of  $I(\delta)$  are generally higher than those given in Table 1. The difference reflects the price to be paid for giving up the assumption that  $\sigma$  is known.

As noted previously, the process of selection of a suitable scheme starts by assuming that  $\sigma$  is known, which enables one to select a scheme by following steps outlined in Section 4.1. In the subsequent process of selecting a suitable value of  $\gamma_\sigma$ , one can make use of another concept—namely, the inertia curve—with respect to changes in  $\sigma$ . This curve is defined by means of the formula

$$I_\sigma(f) = \lim_{T \rightarrow \infty} \sum_{j=1}^H \{E_{T,f}^{(\sigma)} L(\hat{\mu}_{T+j}, \mu) - E_0\}, \quad (5.4)$$

where  $E_{T,f}^{(\sigma)} L(\hat{\mu}_{T+j}, \mu)$  is the expected loss incurred  $j$  units of time after the change in  $\sigma$  by a factor  $f$  at time  $T$ . This concept enables one to design an LSE( $E_0$ ) scheme that achieves a desired trade-off between excess losses associated with changes in  $\mu$  and  $\sigma$ .

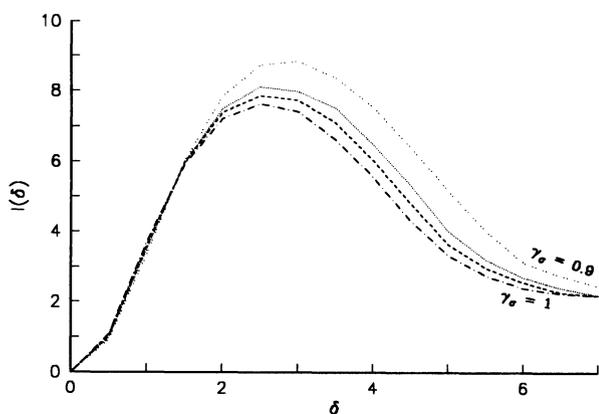


Figure 4. Values of Inertia  $I(\delta)$  Corresponding to AEW Procedures With  $\gamma = .85$  That Belong to LSE( $E_0 = 1/9$ ). The remaining scheme parameters are  $h = 6.41, \gamma_\sigma = 1$  (dotted-dashed line),  $h = 6.78, \gamma_\sigma = .97, c_\sigma = 1.2$  (dashed line),  $h = 7.15, \gamma_\sigma = .95, c_\sigma = 1.3$  (dotted line), and  $h = 8.35, \gamma_\sigma = .9, c_\sigma = 1.4$  (twin-dotted line).

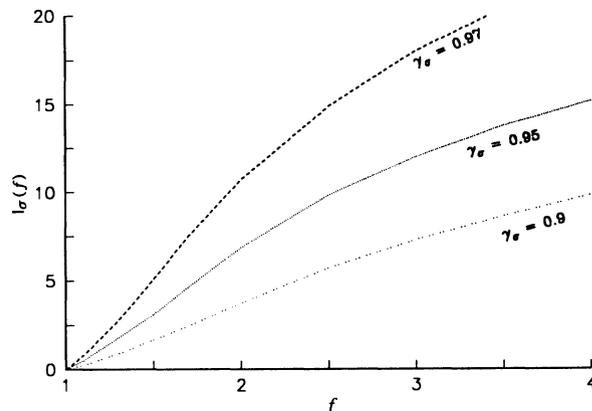


Figure 5. Values of Inertia  $I_\sigma(f)$  With Respect to Changes in  $\sigma$  Corresponding to AEW Procedures With  $\gamma = .85$ . All the procedures belong to LSE( $E_0 = 1/9$ ). The scheme parameters are given in the title of Figure 4.

As an example, consider three AEW schemes with  $\gamma = .85$  that belong to LSE(1/9). The inertia curves  $I(\delta)$  and  $I_\sigma(f)$  corresponding to these schemes are shown in Figures 4 and 5, respectively. The first scheme ( $h = 6.41, \gamma_\sigma = 1$ ) is, essentially, an LE(1/9) scheme derived under the assumption that  $\sigma = \sigma_0$ . This scheme lacks the capability to adapt to a different level of  $\sigma$ . If it turns out that  $\sigma > \sigma_0$ , the process  $\{\hat{\mu}_i\}$  will tend to follow fluctuations in  $\{X_i\}$  too aggressively, leading to increased loss per unit time. The LSE schemes are obtained by decreasing  $\gamma_\sigma$ . The scheme ( $h = 7.15, \gamma_\sigma = .95, c_\sigma = 1.3$ ) is capable of adapting to a new level of  $\sigma$ , but it comes at the expense of higher inertia with respect to changes in  $\mu$ . This trend continues for the scheme ( $h = 8.35, \gamma_\sigma = .9, c_\sigma = 1.4$ ), which assures lower values of  $I_\sigma(f)$  at the expense of higher values of  $I(\delta)$ . Based on the curves shown in Figures 4 and 5, the designer is able to select a suitable scheme. In general, the more stable are the values of  $\sigma_i$ , the closer to 1 can  $\gamma_\sigma$  be chosen. For example, if the scheme with  $\gamma_\sigma = .9$  is adopted, a single change in  $\mu$  of magnitude from  $2\sigma$  to  $6\sigma$  will “cost” an additional one unit of inertia, compared to the scheme with  $\gamma_\sigma = .95$  (see Fig. 4). On the other hand,  $I_\sigma(1.3 | \gamma_\sigma = .95) - I_\sigma(1.3 | \gamma_\sigma = .9) \simeq 1$ . Therefore, if changes in  $\sigma$  typically do not exceed 30% and occur as often as changes in  $\mu$ , one can expect satisfactory performance from the scheme  $\{\hat{\mu}_i, \hat{\sigma}_i\}$  consisting of AEW( $h = 7.15, \gamma = .85$ ) and (5.1) with ( $\gamma_\sigma = .95, c_\sigma = 1.3$ ). If such changes in  $\sigma$  occur less often, one should use a larger value  $\gamma_\sigma$ —for example, .97. Inertia curves corresponding to this scheme can also be found in Figures 4 and 5.

### 6. AN EXAMPLE

Consider once again the situation arising in the production process of integrated circuit chips from semiconductor wafers. Let us focus on one of the process steps in which the layer of silicon dioxide covering the wafer must be etched, by chemical means, until the layer of

*Table 3. The Generalized Markovian Procedure  $\{\tilde{\mu}_i\}$  and the Location-Scale Equivariant AEW Procedure  $\{\hat{\mu}_i\}$  Corresponding to the Example*

$i$	$\bar{x}_i$	$\hat{\sigma}_i$	$\tilde{\mu}_i$	$\hat{\tau}_i$	$\hat{\mu}_i$
1	1.006	.060	1.001	1	1.006
2	1.037	.059	1.005	2	1.023
3	.944	.059	.997	3	.992
4	.957	.059	.993	4	.981
5	1.012	.058	.995	5	.990
6	1.035	.057	.999	6	1.000
7	.917	.058	.987	7	.982
8	1.067	.060	.999	8	1.000
9	1.121	.060	1.022	9	1.023
10	.935	.063	1.009	10	1.007
11	.911	.062	.994	11	.990
12	1.030	.063	.998	12	.997
13	1.018	.062	1.000	13	1.000
14	.941	.062	.993	14	.990
15	1.192	.068	1.056	15	1.023
16	1.142	.067	1.068	16	1.043
17	1.138	.066	1.076	3	1.154
18	1.188	.065	1.095	4	1.165
19	1.080	.066	1.093	5	1.142
20	1.228	.067	1.120	6	1.163
21	1.153	.067	1.123	7	1.161
22	1.141	.066	1.125	8	1.157
23	1.179	.065	1.131	9	1.161
24	1.190	.064	1.138	10	1.166
25	1.184	.063	1.143	11	1.170
26	.880	.069	1.029	1	.880
27	.951	.068	1.018	2	.918
28	.875	.068	.990	3	.902
29	.870	.067	.970	4	.892
30	.811	.066	.934	5	.870
31	.871	.066	.927	6	.870
32	.890	.065	.923	7	.875
33	.866	.064	.916	8	.873
34	.794	.064	.894	9	.857
35	.868	.063	.891	10	.859
36	.854	.062	.887	11	.858
37	.905	.062	.889	12	.867
38	.885	.061	.888	13	.870
39	.885	.060	.888	14	.872
40	.977	.060	.902	15	.890

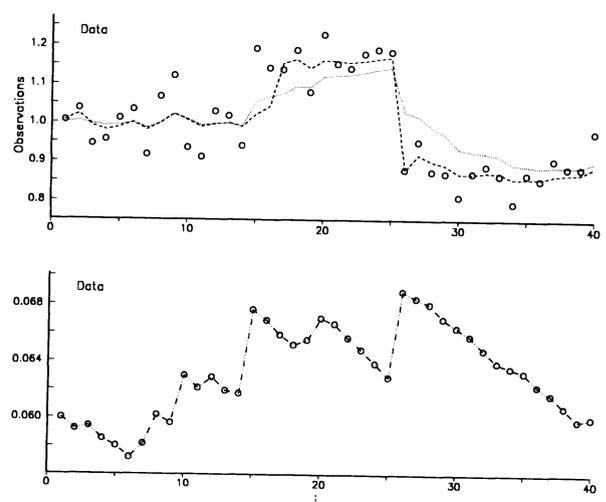
NOTE: The observations  $\{\bar{x}_i\}$  correspond to successive sample averages of four measurements taken from each wafer. Both procedures belong to the class  $LSE(E_0 = 1/9)$ . The third column contains the current estimates  $\{\hat{\tau}_i\}$  used in these procedures and computed by using the Process (5.1) with parameters  $\gamma_\sigma = .97$  and  $c_\sigma = 1.2$ , starting at  $\sigma_0 = .06$  mic. The values  $\{\hat{\tau}_i\}$  correspond to the estimated last stable range (i.e., location of the last point of change) and are used in the AEW procedure.

metal beneath it is reached. Wafers are processed one at a time and the processing time depends on the average of four measurements of oxide film thickness taken at preselected locations on the wafer—in general, thicker layers require longer etching times. Analysis of the process data suggests that the sample means  $\{\bar{X}_i\}$  tend to be independent with moderate changes in variance but possibly strong changes in mean. Therefore, if the process mean corresponding to the current regime were known, the best etching time could be determined by means of a

known formula. The main problem in this situation is how to adapt, as quickly as possible, to the new process mean.

Consider a set of data corresponding to 40 successive wafer averages shown in Table 3. Two procedures are used to estimate the current process level. The first one,  $\{\tilde{\mu}_i\}$ , is the LSF Markovian procedure (2.4) with  $\omega_2(z | \hat{\sigma}_{i-1})$  given by (2.5). The scheme parameters are ( $\gamma = .9, \beta = 4.34$ ) and the value  $\{\hat{\sigma}_{i-1}\}$  is obtained by using Process (5.1) with parameters ( $\gamma_\sigma = .97, c_\sigma = 1.2$ ). The initial (and target) levels of  $\{\bar{X}_i\}$  and its standard deviation are target = 1 micron and  $\sigma_0 = .06$  micron, respectively. The second estimation scheme,  $\{\hat{\mu}_i\}$ , is the LSE version of the AEW procedure with ( $\gamma = .85, h = 6.78$ ) and the values  $\{\hat{\sigma}_{i-1}\}$  used in (3.6) to determine the last point of change are obtained by means of the process (5.1). The values of  $I(\delta)$  corresponding to these two schemes can be found in Table 2.

Both of the processes under consideration belong to the class  $LSE(1/9)$ ; that is, their steady-state variance is  $\sigma^2/9$ . In other words, both  $\{\tilde{\mu}_i\}$  and  $\{\hat{\mu}_i\}$  belong to the same class as the EWMA with  $\gamma = .8$ . Their adaptive capability with respect to larger changes, however, is much higher. The computed values of these sequences, along with the values of  $\{\hat{\sigma}_i\}$ , are given in Table 3. The corresponding plot is shown in Figure 6. As can be seen from this figure, the data undergo two abrupt changes in  $\mu$ , after the points 14 and 25. The LSE Markovian procedure initially tracks the first (smaller) change better, but by the third point after the change, the AEW procedure is able to identify the last point of change and, as a result, it tracks the subsequent data better. In particular, it immediately identifies the second point of change and adapts accordingly, but the LSE Markovian procedure behaves in a more conservative



*Figure 6. The Generalized Markovian Procedure  $\{\tilde{\mu}_i\}$  (dotted line) and the Location-Scale Equivariant AEW Procedure  $\{\hat{\mu}_i\}$  (dashed line) Corresponding to the Example and Table 3. Both procedures belong to the class  $LSE(E_0 = 1/9)$ . The bottom plot contains the current estimates  $\{\hat{\sigma}_i\}$  used in these procedures and computed by using the Process (5.1).*

fashion. These observations illustrate the general property that the LSE Markovian procedure can be expected to perform better than the LSE version of the AEW scheme with respect to smaller changes but worse with respect to larger changes.

7. CONCLUDING REMARKS

To decide which estimation procedure is suitable in any given situation, it is not necessary to impose a model on the process of changes. In fact, such models frequently do not have solid foundations and can lead to schemes with low tracking capability. Similarly, tuning the estimation scheme to past data, as is frequently done when handling engineering process control models (e.g., see Hunter 1986) can hardly be recommended for models involving abrupt changes because the pattern of changes that has been observed in the past is not necessarily of the type that one wants protection against in the future. The two-stage approach discussed in this article enables one to design estimation schemes based on the information that the process owner is likely to possess. At the first stage, we restrict our attention to the schemes that generate, under stable conditions, the same loss per observation,  $E_0$ . In the second stage, we select a scheme with a suitable (low) inertia. The selection process typically reduces to a one-dimensional search because of the constraint related to  $E_0$ .

In spite of the optimality property of the EWMA procedure proven in Theorem 2.1, the procedure cannot be recommended except in the very special situations noted in Section 4.1. This procedure is of value only insofar as it can be modified so as to overcome its drawbacks.

The function  $I(\delta)$  can be used as a basis for a risk index computed by imposing a distribution on the values of  $\delta$ , but such an averaging can also mask some important features of the procedure of interest, such as inadmissibility. Even if such an approach is adopted to simplify the selection process, it is still recommended to examine the function  $I(\delta)$  itself before a procedure is selected for use.

The proposed approach can be used in situations in which the distribution of  $\{X_i\}$  depends on nuisance parameters that also need to be estimated on the basis of the same data. In particular, when the nuisance parameter corresponds to the scale of  $\{X_i\}$ , the methods discussed in Section 5 lead to estimation, the properties of which can be easily assessed within the proposed framework, especially when  $\{\hat{\mu}_i\}$  belongs to LSE( $E_0$ ). In many practical situations in which no LE or LSE estimation is possible, one can find data transformations that enable one to handle the problem in this framework, as illustrated by Example 4.2.

The relative merits of the different schemes discussed in this article are by no means universal: They may depend on the particular form of the loss function, as well as assumptions of Normality and independence. The main purpose of this work is to put forward a simple and coherent framework for selecting an estimation scheme appropriate in a given situation. The basic ideas used for comparison

and derivation of estimation procedures discussed in this work are, however, fully usable in more general situations involving serially correlated or multivariate data, though the details will vary. The proposed approach is also appropriate for estimating the current level of process parameters other than the population mean, as well as for handling other types of change in these parameters.

ACKNOWLEDGMENTS

I thank Betty J. Flehinger and Jonathan R. Hosking (IBM Research) for useful consultations and discussions on this subject and related examples. Thanks are also due to the editor, associate editor, and the referees for substantial help in improving the quality of this article.

APPENDIX A: PROOF OF THEOREM 2.1

For the Procedure (2.1), the expected loss  $j$  units of time after the change at time  $T$  is

$$E_{T,\delta}L(\hat{\mu}_{T+j}, \mu) = A\sigma^{-2} \cdot E \left[ \sum_{k=0}^{j-1} a_k (X_{T+j-k} - \mu) + \sum_{k=j}^{n-1} a_k (X_{T+j-k} - \mu + \delta\sigma) - \delta\sigma \sum_{k=j}^{n-1} a_k \right]^2 = A \cdot \sum_{k=0}^{n-1} a_k^2 + A\delta^2 \left( \sum_{k=j}^{n-1} a_k \right)^2, \quad j = 1, 2, \dots, n-1, \quad (A.1)$$

and, after  $n$  units of time, the adaptation is complete; that is,

$$E_{T,\delta}L(\hat{\mu}_{T+j}, \mu) \equiv E_0 = A \cdot \sum_{k=0}^{n-1} a_k^2, \quad j \geq n. \quad (A.2)$$

The problem, therefore, becomes as follows: minimize

$$I(\delta)/(A\delta^2) = \sum_{j=1}^{n-1} \left( \sum_{k=j}^{n-1} a_k \right)^2$$

with respect to  $\{a_k\}$  and  $n$  subject to constraints

$$\sum_{k=0}^{n-1} a_k^2 - E_0/A = 0 \quad (A.3)$$

and

$$\sum_{k=0}^{n-1} a_k - 1 = 0. \quad (A.4)$$

For a fixed value of  $n$ , this is a fairly typical quadratic optimization problem. When  $n^{-1} < E_0/A < 1$ , its solution can be found by introducing the Lagrange function

$$G = \sum_{j=1}^{n-1} \left( 1 - \sum_{k=0}^{j-1} a_k \right)^2 + \lambda \left( \sum_{k=0}^{n-1} a_k^2 - E_0/A \right) + \lambda_1 \left( \sum_{k=0}^{n-1} a_k - 1 \right) \quad (A.5)$$

and solving the equations

$$\frac{\partial G}{\partial a_i} = -2 \sum_{j=i+1}^{n-1} \left( \sum_{k=j}^{n-1} a_k \right) + 2\lambda a_i + \lambda_1 = 0, \quad i = 0, 1, \dots, n-2$$

$$\frac{\partial G}{\partial a_{n-1}} = 2\lambda a_{n-1} + \lambda_1 = 0, \quad (A.6)$$

together with the constraints (A.3) and (A.4). Now, by leaving the bottom equation intact and subtracting each equation from the one above it, the system (A.6) can be written in the form

$$\begin{bmatrix} \lambda & -(1+\lambda) & -1 & -1 & \dots & -1 \\ 0 & \lambda & -(1+\lambda) & -1 & \dots & -1 \\ 0 & 0 & \lambda & -(1+\lambda) & \dots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & \lambda & -(1+\lambda) \\ 0 & 0 & \dots & \dots & 0 & \lambda \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{n-2} \\ a_{n-1} \end{bmatrix} = -\frac{\lambda_1}{2} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \quad (A.7)$$

Denote  $\gamma = (1 + 2\lambda - \sqrt{1 + 4\lambda}) / (2\lambda)$ . Then, after some algebra, one can show that the solution of the system (A.7, A.3, A.4) can be represented in the form

$$a_j = \frac{(1 - \gamma)\gamma^j(1 + \gamma^{2n-2j-1})}{1 - \gamma^{2n}}, \quad j = 0, 1, \dots, n-1, \quad (A.8)$$

where  $\gamma$  is determined from (A.3); that is,

$$\frac{1 - \gamma}{1 + \gamma} \cdot \frac{1 - \gamma^{4n} + 2n(1 - \gamma^2)\gamma^{2n-1}}{(1 - \gamma^{2n})^2} = E_0/A. \quad (A.9)$$

The value of  $\lambda_1$  is then determined from the last equation of (A.6). Theorem 2.1 is obtained as a limiting case of (A.8) and (A.9) when  $n \rightarrow \infty$ .

It is interesting to note that (A.8) also gives weights of the optimal linear predictor for the integrated first order moving average process (see Wu, Hosking, and Doll 1992).

### APPENDIX B: DESIGN AND ANALYSIS OF MARKOVIAN ESTIMATION SCHEMES

In this appendix, I shall briefly discuss the computation procedure for analysis of Markovian procedures. This procedure served as a basis for software used for analysis of schemes discussed in this article. It shows that the problem of design and analysis can be solved by using conventional analysis of Markov chains.

To design a scheme that belongs to  $LE(E_0)$  in the case in which  $\{X_i\}$  are independent, I shall discretize the space of  $\{\hat{\mu}_i\}$  and approximate this process by a discrete

Markov chain. Let the discretization interval be  $D\sigma$ , and let  $\{mD\sigma, m = 0, \pm 1, \pm 2, \dots\}$  represent the discretized values of  $\hat{\mu}_i$ . Denote by  $P(\mu)$  the transition matrix of this chain corresponding to the process mean  $\mu$ . For example, for the symmetric Markovian scheme based on  $\omega_2(z)$  given by (2.5), the elements of this matrix are

$$P_{mj}(\mu) = P\{(j - .5)D\sigma \leq X + \omega_2(mD\sigma - X) \leq (j + .5)D\sigma\}$$

$$= F[mD + \beta t_\gamma((j - m + .5)D/\beta) - \mu/\sigma] - F[mD + \beta t_\gamma((j - m - .5)D/\beta) - \mu/\sigma], \quad (B.1)$$

where  $F$  is the cdf of the normalized  $X$  and  $t_\gamma(y)$  is the solution of the equation  $t[1 - \gamma \exp(-t^2/2)] = y$ . Now the vector of steady-state probabilities  $\pi_0 = \{\pi(m)\}$  is computed by solving the equations

$$\pi_0 = \pi_0 P(\mu), \quad \pi_0 1 = 1, \quad (B.2)$$

and the steady-state loss per unit time is  $\sum_m \pi_0(m)L(mD\sigma, \mu)$ . For a fixed value of  $\gamma$ , a value  $\beta$  is found for which this sum is equal to  $E_0$ .

To obtain the values of inertia function for a given pair  $(\gamma, \beta)$  corresponding to a scheme from  $LE(E_0)$ , let us first compute the vector  $\pi_j = \{\pi_j(m)\}$  that represents the distribution of the discretized values of the scheme given that a change from  $\mu$  to  $\mu + \delta\sigma$  has occurred  $j$  observations ago. These values are determined from the relation

$$\pi_j = \pi_{j-1} P(\mu + \delta\sigma), \quad j = 1, 2, \dots, H, \quad (B.3)$$

where  $P(\mu + \delta\sigma)$  is the transition matrix corresponding to the mean  $(\mu + \delta\sigma)$ . Now  $I(\delta)$  is obtained by

$$I(\delta) = \sum_{j=1}^H \left[ \sum_m \pi_j(m)L(mD\sigma, \mu + \delta\sigma) - E_0 \right]. \quad (B.4)$$

[Received April 1992. Revised November 1994.]

### REFERENCES

Barnard, G. (1959), "Control Charts and Stochastic Processes," *Journal of the Royal Statistical Society*, Ser. B, 21, 239-257.  
 Bather, J. (1963), "Control Charts and Minimization of Costs," *Journal of the Royal Statistical Society*, Ser. B, 25, 49-80.  
 Chen, C., and Tiao, G. C. (1990), "Random Level-Shift Time Series Models, ARIMA Approximations and Level-Shift Detection," *Journal of Business & Economic Statistics*, 8, 83-97.  
 Chernoff, H., and Zacks, S. (1964), "Estimating the Current Mean of a Normal Population Which Is Subject to Changes in Time," *The Annals of Mathematical Statistics*, 35, 999-1018.  
 Hunter, J. S. (1986), "The Exponentially Weighted Moving Average," *Journal of Quality Technology*, 18, 203-210.  
 Johnson, N. L., and Kotz, S. (1969), *Distributions in Statistics* (Vols. 1-4), New York: John Wiley.  
 Kenett, R., and Zacks, S. (1992), "Tracking Algorithms for Processes With Change Points," unpublished manuscript.

- McCulloch, R. E., and Tsay, R. S. (1993), "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series," *Journal of the American Statistical Association*, 88, 968-978.
- Muth, J. F. (1960), "Optimal Properties of Exponentially Weighted Forecasts," *Journal of the American Statistical Association*, 55, 299-306.
- West, M. (1986), "Bayesian Model Monitoring," *Journal of the Royal Statistical Society, Ser. B*, 48, 70-78.
- Wetherill, G. B. (1977), *Sampling Inspection and Quality Control*, London: Chapman and Hall.
- Woodward, R., and Goldsmith, P. L. (1964), *Cumulative Sum Techniques* (ICI Monograph 3), London: Oliver and Boyd.
- Wu, L. S., Hosking, J. R. M., and Doll, J. M. (1992), "Business Planning Under Uncertainty," *International Journal of Forecasting*, 8, 545-557.
- Yashchin, E. (1992), "On the Problem of Estimating the Current Mean of Processes Subject to Abrupt Changes," Research Report RC#17923, IBM, Yorktown Heights, NY.