MACHINE MONITORING
USING PROBABILITY THRESHOLDS
AND SYSTEM OPERATING CHARACTERISTICS

Stephen M. Pollock
Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109 - 2117

and

Jeffrey M. Alden
G.M. R&D Center
Warren, MI 48090

IOE Tech Report 95-14

1. Introduction

We are concerned with detecting a change in the underlying condition of a process, when the available observations are related only probabilistically to this condition. This situation has long been of concern to statisticians, engineers, economists, epidemiologists, etc. In this paper we address the specific context of monitoring a discrete-part production machine, with the objective of effectively determining when to shut the machine down for maintenance or replacement. Applications to other areas such as quality control, health, military surveillance, or economic analysis should be readily apparent.

Before presenting any details, we list some definitions and initial assumptions, in order to clarify the general setting that governs our analysis:

a) There is an underlying time interval that characterizes the operation of the machine, most often the "part production cycle time". All times and intervals are subsequently measured in units of this time interval.

b) The machine can be in only one of two conditions: "good" or "bad" (denoted G and B, respectively). By the G condition we mean the machine can operate in such a way that it is "in control" or "normal" or otherwise able to produce acceptably; by the B condition we mean it is "out of control", "failed" or only able to produce bad parts (scrap).

c) The machine starts in G but at some random variable operating time T (called the *failure time*) goes to B. This is called a "failure event", or more simply, a "failure".

d) Observations of "signals" which are probabilistically related to the machine's condition are made at fixed, pre-determined times.

e) Immediately following any observation one of two possible actions can be made: "do nothing" or "take an action consistent with believing the machine is in B". The latter investigative action is called a *check*.

f) When a check is made, production is stopped and the condition of the machine becomes known with certainty. A check that finds the machine in G, called a *false alarm*, returns the machine to operation (in G) after an interval of length g. A check that finds the machine in B, called a *true alarm*, re-sets it to "as-new" condition – or, equivalently, replaces it by a new (identical) machine – after an interval of length b. This event is called a *renewal*. Typically $g < b$ since renewal often requires fixing or replacing something, while checking when the machine is in G may only require a brief inspection to ascertain that it is in fact in G.

g) The process (observations, failures, checking, etc.), which we call "monitoring", con-

tinues indefinitely, with successive failure times assumed to be independent and identically distributed (I.I.D.).

All of these assumptions, of course, must be eventually replaced or relaxed to conform to the realities of any actual processes. On the other hand, important underlying relations among performance measures, and their dependence upon machine parameters and checking strategies, will become evident using this set of simplifying assumptions.

Our fundamental problem, then, is to determine a *policy*, i.e. when to check, knowing the complete history of the process, including its age, the elapsed time since the last check and the values of all observations made to date. There are two competing concerns that underly the determination of an effective policy: checking soon enough so that the machine does not operate too long in condition B, while not checking so often that the machine is shut down unnecessarily. In other words, an advantage is gained when a policy raises an alarm that detects the occurrence of B soon after it happens. However, it is also desirable to avoid costly false alarms that result in shutting the machine down to check it when it is still in condition G.

Trading off (or constraining) the costs of delayed failure detection and false alarms is the basis of most quality control and control chart procedures developed over the past sixty years [see Shewhart (1931), Duncan (1956), Roberts (1966), Johnson and Leone (1962), Montgomery (1980) and Lorenzen and Vance (1986), for a historical perspective, and Basseville and Nikoforov (1993) for a contemporary and extensive comparison and evaluation of a variety of methods]. The most common approaches to "optimizing" these procedures [e.g. Moskowitz, Plante and Chun (1989), Saniga (1989)] use economic models that explicitly incorporate costs ascribable to false alarms and delayed checking. Such trade-off analyses and economic approaches depend upon specific policy structures that, although they have intuitive appeal and lead to easy computation or evocative charting methods, can be inefficient or arbitrary in nature.

More important, such methods (indeed, most statistical process control approaches) often essentially ignore what is known about the machine's failure time distribution. In contrast, we explicitly make use of this distribution to present and evaluate a checking policy that is "optimal" according to a broad set of criteria. Our approach is based on earlier work by Shiryaev (1978), Pollak and Siegmund (1975), Pollak (1985) and others. It is also motivated by a natural inclination to develop checking policies that become, in appropriate limits, those that have been shown in the literature to be optimal for well-studied situations where either

no observations or *perfect* observations are made. The former has been long analysed under the rubric of "optimal maintenance policies" [see, for example, the seminal work of Barlow et al. (1963)]; the latter has been the subject of "optimal replacement policies" [see, for an early example, Page (1954)]. These two special cases "bracket" the capabilities of any realizable system that involves information gathering.

2. Performance Measures

The use of *any* checking policy ultimately results in performance measures of interest to decision makers. Before one can find effective checking policies, then, it is important to define these measures and understand the relations among them. We choose to list these measures in three general categories, along with parameters and variables used in our analysis.

Type One (False Alarm) Measures.

False alarms are costly since resources are used to process each alarm, and production time is often lost as well. Measures that account for these include:

$r_f \equiv$ *false alarm rate* $\equiv$ the expected number of false alarms per unit time,

$p_f \equiv$ fraction of total time spent processing false alarms,

$\mu \equiv$ expected number of false alarms until failure,

ARL0 $\equiv$ the expected machine operating time until a check, given the machine *starts and remains* in G

Type Two (Late Detection) Measures.

Being slow to stop a machine that is in B leads to the production of bad parts or to lost production. Relevant measures include:

$D \equiv$ (random variable) time between failure and the next check,

$\delta \equiv$ *expected detection time* $\equiv$ E(D), also called EDD – "expected delay in detection" – see Marcellus (1993),

$r_t \equiv$ *true alarm rate* $\equiv$ the expected number of true alarms per unit time,

$p_t \equiv$ fraction of total time spent processing true alarms,

ARL1 $\equiv$ the expected machine operating time until a check, given the machine *starts and remains* in B

We note here that, in spite of their popularity in the literature (and in practice), we avoid the use of the average run length measures ARL0 and ARL1 for two reasons:

a) the hypothetical situation (required for computing ARL0) where the machine is "forced" to remain in G until the occurence of the first alarm is hard to justify. Its interpretation is particularly unclear if a policy allows the machine to fail before the first alarm;

b) ARL1 is defined only for the situation where the machine *starts and remains* in condition B – a situation that is rarely realized in practice.

These shortcomings have been pointed out before by Woodall (1985,1986), Svoboda (1991) and others.

Composite Measures.

Other performance measures of operational interest can be expressed as simple functions of Type One and Type Two measures. (In the following definitions, the term *total time* means operating time plus checking time):

$p_S \equiv$ fraction of total time the machine is in B and producing scrap,

$p_B \equiv$ fraction of total time the machine is in B (either producing scrap or being replaced),

$p_G \equiv$ fraction of total time the machine is in G (i.e. producing usable parts),

$r \equiv$ *total alarm rate* $\equiv$ the expected number of alarms of any type per unit time,

$p_0 \equiv$ fraction of total time spent processing alarms of any type.

One can combine performance measures to obtain the overall cost per unit time. Computing this cost rate, however, using the simplest linear model, requires the following cost coefficients:

$K_f =$ fixed cost per false alarm,

$V_f =$ cost per unit time spent processing a false alarm, including lost production,

$K_t =$ fixed cost per true alarm,

V_t = cost per unit time spent when processing a true alarm (i.e. when the machine is stopped and found to be in B), including lost production, and

V_d = cost per unit time of producing scrap while in condition B, including lost production.

The resulting total cost per unit time, c_T, can then be expressed as:

$$c_T = K_f r_f + V_f p_f + K_t r_t + V_t b r_t + V_d p_S. \tag{1}$$

Computing c_T requires knowing all these cost coefficients (minimizing c_T requires knowing at least their ratios), which in many cases are difficult (if not impossible) to obtain. For this reason, and because the method we propose to use can be implemented without the availability of these cost coefficients, we do not directly pursue cost-minimization. Our approach concentrates, instead, on computing *non-cost* performance measures and using these to choose among various checking policies. If cost coefficients are available, however, the form of the policy we produce can be readily manipulated numerically to produce a decision rule that minimizes c_T.

3. System Operating Characteristics

We are ultimately concerned with providing decision makers with checking policies that:

a) are easy to understand and implement;

b) readily allow input of, and sensitivity analysis with respect to, important parameters, including the expected value of the failure time, and the discriminatory capabilities of various sensors; and

c) do not require an explicit assessment of the hard-to-estimate cost coefficients in equation (1).

The means by which we present the consequences of using *any* particular monitoring policy is the *System Operating Curve* or *System Operating Characteristic* (SOC) [see Pollock (1965) and Rapoport et al. (1973) for early development]. The SOC is a simple graphical plot involving two axes – one showing a Type One measure, the other a Type Two (or composite) measure. A single point on the SOC represents a pair of performance measures attainable by using a particular checking policy with a machine characterized by specified

parameter values. A family of such points represents the range of output measures attainable by changing one or more policy variables available to the decision maker.

For example, a SOC could be a plot of p_S (the fraction of time the machine is producing scrap) versus p_0 (the fraction of time spent checking alarms of any type). Consider a set of such SOC's, as shown in Figure 1, one for each of three different hypothetical monitoring situations A, B and C on the same machine. Each curve represents the set of operating points (i.e., values of p_S and p_0) achievable by using different values of a particular policy variable.

Figure 1: System Operating Characteristics (SOC) for three hypothetical situations A, B and C.

Assuming that the costs of implementing A, B and C are all the same, situation C is clearly better than A or B, since it has either a lower p_S for any given p_0, or it has a lower p_0 for any given p_S. (Perhaps C allows the monitoring of signals that are not available to either A or B). If forced to chose between A or B, however, A would be preferred to B when the advantage of achieving early detection (i.e. small p_S) is more important than the disadvantage of spending a large fraction of time checking alarms.

The SOC is similiar to the *Receiver Operating Characteristic* (ROC) used in telecommu-

Figure 2: A single cycle with n checks ($n-1$ of which are false alarms). Vertical arrows indicate checking actions and renewals; T = operating time until failure which occurs at time $T+(n-1)g$; D = detection time; t_i = time of i^{th} check (a false alarm for $i < n$); t_n = time condition B is detected; and $t_n + b$ is the time the machine re-enters condition G (which ends the cycle).

nication and signal detection theory, and is also clearly related to the *power curve* or *operating characteristic* of fixed-sample or sequential hypothesis testing. The word *system* is used to emphasize the fact that it is a combination of the machine, the monitoring device *and* a checking policy that is being represented. Note that, given appropriate cost coefficients, equation (1) can be used to produce values of c_T for any point on the SOC curve.

4. Relationships Among Performance Measures

We now present general relations among the performance measures that exist for *any* reasonable checking/monitoring procedure. To obtain these, we define a *cycle* to be the time between renewals. Figure 2 shows a cycle that contains n checks at times $t_1, t_2, \cdots, t_n$. There are $n-1$ checks that find the machine in G (i.e., there are $n-1$ false alarms), each requiring time g. The last check in the cycle finds the machine in B and takes time b. Since the machine is stopped during checks of any type, in any cycle the *operating* time until failure, T, differs from the total time until failure which equals T plus the time spent checking (while in G) prior to failure.

From Figure 2 we see that $\overline{L}$, the expected cycle length when there are n checks (hence

$n-1$ false alarms), is

$$\overline{L} = E(T) + g(n-1) + E(D) + b. \tag{2}$$

Since μ is the expected number of false alarms until failure and $\delta \equiv E(D)$, the overall expected cycle length is

$$\overline{L} = E(T) + g\mu + \delta + b. \tag{3}$$

An elementary use of the fundamental renewal theorem gives

$$r_f = \frac{\mu}{E(T) + g\mu + \delta + b} \tag{4}$$

and

$$r_t = \frac{1}{E(T) + g\mu + \delta + b}. \tag{5}$$

The ratio of these two equations gives

$$\mu = \frac{r_f}{r_t}. \tag{6}$$

Solving for δ in equation (5) gives

$$\delta = \frac{1}{r_t} - E(T) - g\mu - b = \frac{1}{r_t} - E(T) - g(\frac{r_f}{r_t}) - b. \tag{7}$$

The expected time the machine spends in B is $\delta + b$. Since true alarms occur at a rate r_t,

$$p_B = (\delta + b)r_t, \tag{8}$$

The times to process a false alarm and a true alarm are g and b, respectively. Since these occur at rates r_f and r_t

$$p_f = gr_f \tag{9}$$

and

$$p_t = br_t. \tag{10}$$

Finally, by definition,

$$r = r_f + r_t \tag{11}$$

$$p_0 = p_t + p_f. \tag{12}$$

$$p_S = p_B - p_t \tag{13}$$

$$p_G + p_S + p_f + p_t + p_0 = p_G + p_S + 2p_0 = 1 \tag{14}$$

Relations (2) through (14) hold for *any* checking procedure. Thus, having computed any pair of Type One and Type Two measures (such as r_t and r_f) and knowing $E(T)$, b, and g, allows the calculation of δ, p_B, p_G, p_t and p_f, etc.

For expositional simplicity, in the remainder of this paper we set the lengths of time needed to perform checks to unity, i.e. $b = g = 1$. Converting computations for arbitrary b and g values (even when they are zero) is straightforward, as shown in Appendix C.

5. The Basic Monitoring Process

We now present additional assumptions and notation used to define the monitoring process, determine an optimal checking policy, and to compute performance measures that result from its use. Let:

$W(t) \equiv$ the cumulative distribution for the failure time $T = \text{prob.}\{T \leq t\}$.

$i =$ observation number, $i = 1, 2, ...$

$\tau_i =$ time at which the i^{th} observation is made.

Since observations (and therefore checking opportunities) are limited to the times τ_i, it is useful to define the function $f(i)$, the probability that failure occurs between the $(i-1)$st and ith *observation*, so that

$$f(i) \equiv W(\tau_i) - W(\tau_{i-1}),\ i = 1, 2, 3, \ldots, \tag{15}$$

where $\tau_0 \equiv 0$ (even though no observation is actually made at time 0).

Let $C_t \equiv$ condition of the machine at time t. The observation made at time τ_i is the random variable X_i, having a p.d.f. $f_{X_i}(\cdot)$, depending upon the machine condition as follows:

$$f_{X_i}(x) = \begin{cases} p(x); & \text{if } C_{\tau_i} = G,\ i = 1, 2, \ldots, \\ q(x); & \text{if } C_{\tau_i} = B,\ i = 1, 2, \ldots \end{cases} \tag{16}$$

The random vector $\underline{X}_n$ of observations is defined as

$$\underline{X}_n \equiv (X_1, X_2, \ldots, X_n),$$

and its realization $\underline{x}_n$ is

$$\underline{x}_n \equiv (x_1, x_2, \ldots, x_n).$$

Any checking policy can be viewed as being a decision rule to determine whether or not to check the machine at time τ_n given the set of observations $\underline{x}_n$. For example:

a) using the ordinary Shewhart chart (see, e.g. Montgomery [1996]), the decision is based upon only the last observation x_n: if this value falls outside pre-determined control limits then a check (i.e. a "search for an assignable cause") is made. The chart's control limits are policy variables, and varying them will produce an associated SOC.

b) using Shewhart charts with supplementary runs tests, if K out of the last N observations fall within a pre-specified zone, then a check is made. Here, K and N and the control limits are policy variables that, when varied, will produce the associated SOC.

SOCs resulting from the use of these charts are shown and compared in Lele (1995). Other policies, e.g. CUSUM or EWMA charts, make use of various functions of some (or all) of the observations $\underline{x}_n$ [see Basseville and Nikoforov (1993)] . It is important to note, however, that most of these policies are, in some sense, ad-hoc. In contrast, we now introduce a class of policies based upon certain optimality conditions.

6. The Probability Threshold Rule (PTR)

A particularly attractive form of checking policy can be based upon a simple proposition: since the observations at times τ_1 through τ_n provide information about machine condition, this information can be used to "update" the probability that $C_{\tau_n} = B$. Specifically, the probability

$$P_n(\underline{x}_n) \equiv \text{prob.}\{C_n = B | C_0 = G, X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\},$$

can be used as the fundamental element in the definition of the Probability Threshold Rule (PTR): Check when $P_n(\underline{x}_n)$ first equals or exceeds some *threshold probability* p^*.

If $W(t)$ is geometric, it is well known (e.g., see Girshick and Rubin [1952], Shiryaev [1963], or Pollock [1965] for early references, or Pollak [1987] for a more recent one) that the PTR is "optimal" since a threshold-type policy variable p^* exists that minimizes either:

a) the expected detection time δ for a given value of false alarm rate r_f; or

b) the false alarm rate r_f for a given value of expected detection time δ; or

c) the cost per unit time, c_T, given by equation (1)

From an *operational* point of view this means that, given a particular value of p^*, a decision maker can achieve the associated performance level shown as a point on a SOC. Or, equivalently, from an *analysis* or *strategic* point of view, given the ability to efficiently compute $P_n(\underline{x}_n)$ the decision maker can vary values of p^* to explore trade-offs among various performance measures. For example, if curve C of Figure 1 represents a typical SOC associated with using a PTR policy: large p_S and small p_0 are produced when $p^* \to 1$, and small p_S and large p_0 are produced when $p^* \to 0$.

Note that the PTR has only one "free" parameter, p^*.

7. Computing $P_n(\underline{x}_n)$

The computation of $P_n(\underline{x}_n)$ follows from the definition of conditional probability:

$$\begin{aligned} P_n(\underline{x}_n) &\equiv \text{prob.}\{T \le \tau_n | \ \underline{X}_n = \underline{x}_n\} \\ &= \frac{\text{prob.}\{T \le \tau_n \cap \underline{X}_n = \underline{x}_n\}}{\text{prob.}\{\underline{X}_n = \underline{x}_n\}} \end{aligned} \tag{17}$$

where

$$\text{prob.}\{T \le \tau_n \cap \underline{X}_n = \underline{x}_n\} = \sum_{j=1}^{n} f(j) \Pi_{i=1}^{j-1} p(x_i) \Pi_{k=j}^{n} q(x_k) \tag{18}$$

and

$$\text{prob.}\{\underline{X}_n = \underline{x}_n\} = \ \text{prob}\{T \le \tau_n \cap \underline{X}_n = \underline{x}_n\} + \sum_{j=n+1}^{\infty} f(j) \Pi_{i=1}^{n} p(x_i). \tag{19}$$

Substituting (18) and (19) into (17), and dividing the numerator and denominator by $\Pi_{i=1}^{n} p(x_i)$, gives

$$P_n(\underline{x}_n) = \frac{\sum_{j=1}^{n} f(j) \Pi_{k=j}^{n} L(x_k)}{\sum_{j=1}^{n} f(j) \Pi_{k=j}^{n} L(x_k) + \overline{F}(n)} \tag{20}$$

where

$$L(x_i) \equiv q(x_i)/p(x_i) \tag{21}$$

is the likelihood ratio for condition B given $X_i = x_i$, and

$$\overline{F}(n) \equiv \sum_{i=n+1}^{\infty} f(i) = \text{prob}\{T > \tau_n\}.$$

Although computing $P_n(\underline{x}_n)$ directly from equation (20) is straightforward, it is computationally advantageous, instead, to use the "odds in favor of condition B" (also called the odds ratio)

$$R_n(\underline{x}_n) \equiv P_n(\underline{x}_n)/(1 - P_n(\underline{x}_n)). \tag{22}$$

(For notational convenience, the argument $\underline{x}_n$ is suppressed for the remainder of this paper, e.g., $R_n(\underline{x}_n)$ is written as R_n.) This odds ratio can be obtained directly from equation (20) as

$$R_n = [\overline{F}(n)]^{-1} \sum_{j=1}^{n} f(j) \Pi_{k=j}^{n} L(x_k). \tag{23}$$

This allows a recursive representation for R_n:

$$R_n = \frac{L(x_n)}{\overline{F}(n)} [\overline{F}(n-1) R_{n-1} + f(n)], \tag{24}$$

which can be confirmed by substitution into equation (23). Equation (24) is an excellent way to compute R_n, and thus $P_n = R_n/(1 + R_n)$, since R_n is calculated from the previously obtained R_{n-1} and each new observation x_n by means of a simple addition and multiplication.

Equation (24) clarifies the challenge of computing performance measures associated with the PTR. In particular, consider computing the point in time when P_n first equals or exceeds p^* (at which time the PTR produces a check). This is the equivalent of finding the smallest value of n such that R_n equals or exceeds the "odds threshold"

$$\rho^* \equiv p^*/(1 - p^*). \tag{25}$$

When x_{n+1} is replaced by the random variable X_{n+1}, we see that equation (24) can be viewed as the generator of a Markov Process R_n. This process has as a state space the non-negative real line $\mathbb{R}^+$, with transitions at observation times τ_n governed by the stochastic behavior of X_n, which in turn are governed by the p.d.f.s of equation (16). When an observation of $X_{n+1} = x_{n+1}$ is made, either

(a) $R_{n+1} < \rho^*$, and the process continues; or

(b) $R_{n+1} \geq \rho^*$, and a check is made.

8. Performance Measures for Two Limiting Cases:

In this section we "bracket" performance of the PTR for any realizable monitoring system

by computing performance measures for the limiting worst-case and best-case extremes of monitoring.

No Observations: As a "worst case" bound on the SOC, we can consider the limiting situation where observations provide *no* information, which is equivalent to having $p(x_i) = q(x_i)$, so that $L(x_i) = 1$, for $i = 1, 2 \ldots$. In this case, equation (24) reduces to

$$R_n = [\overline{F}(n)]^{-1}[\overline{F}(n-1)R_{n-1} + f(n)], \tag{26}$$

with solution (given boundary condition $R_0 = 0$):

$$R_n = \frac{F(n)}{\overline{F}(n)}. \tag{27}$$

From the definition of R_n in equation (22), this gives

$$P_n = F(n), \tag{28}$$

a result that holds for any $F(n)$. Clearly, observations of $\underline{x}_n$ have no effect on P_n, which is simply the cumulative distribution for the failure time T evaluated at $T = \tau_n$.

Two performance measures that can be readily calculated are μ and δ. By definition of the PTR, the (deterministic) time to first check, t_1, is the smallest monitoring time such that the threshold probability is exceeded, i.e.

$$t_1 = \min_{t \in \{\tau_i\}} \{t : W(t) \geq p^*\}. \tag{29}$$

Similarly, the time of the j^{th} check, t_j, can be shown to be

$$t_j = \min_{t \in \{\tau_i\}} \{t : \frac{W(t) - W(t_{j-1})}{1 - W(t_{j-1})} \geq p^*; t > \tau_{j-1}\} \tag{30}$$

where $t_0 = 0$.

The computation of μ and δ becomes straightforward if we assume that the inequalities of expressions (29) and (30) are satisfied as equalities. This situation holds if the checking times t_j are not constrained to be in the set $\{\tau_j\}$ – a reasonable assumption if the monitoring provides no information. In this case, the t_j are easily shown to satisfy

$$W(t_j) = 1 - (1 - p^*)^j. \tag{31}$$

The probability that the failure time T is between the $(j-1)^{st}$ and the j^{th} check, i.e. that $\{t_{j-1} < T \leq t_j\}$, is $W(t_j) - W(t_{j-1})$. Since all previous checks will have produced false alarms, we have

$$\mu = \sum_{j=1}^{\infty}(j-1)[W(t_j) - W(t_{j-1})] = 1/p^* - 1 \tag{32}$$

The time late, given $t_{j-1} < T \leq t_j$ is $t_j - T$, so by equation (31) and simple decomposition of expectations, the expected time late $E(D) \equiv \delta$ is

$$\delta = \sum_{j=1}^{\infty}\int_{t=t_{j-1}}^{t_j}(t_j - t)dW(t) \ = p^*\sum_{j=1}^{\infty}(1-p^*)^{j-1}W^{-1}[1-(1-p^*)^j] - E(T). \tag{33}$$

This result is equivalent to similar ones contained in the literature, and has served as the basis for cost-minimizing checking policies over the past thirty years, starting with Barlow and Proschan [1965].

Perfect Monitoring: The limiting case of "perfect" information represents another special situation. In this case the supports of $p(x_i)$ and $q(x_i)$ are disjoint: $L(x_i) = 0$ for all i such that $\tau_i < T$; and $L(x_i) = \infty$ for all i such that $\tau_i \geq T$. From equation (20), we see that *any* non-zero threshold less than one is exceeded at time τ_n if and only if n is such that $\tau_n \geq T$. In this case

$$\delta = E_T[\min_{t\in\{\tau_i\}}\{t - T : t \geq T\}], \tag{34}$$

and the machine is checked only when it fails. From this, it is trivial to show that the true alarm rate is $r_t = 1/(E(T) + b + \delta)$, the false alarm rate is $r_f = 0$, the fraction of time spent processing true alarms is $p_t = b/(E(T) + b + \delta)$ and the fraction of time processing false alarms is $p_f = 0$.

9. Behavior of R_n for Geometric Failure Time Using the PTR

Unfortunately, computation of performance measures for the PTR is extremely difficult for general failure time distributions $W(t)$. (For a recent example dealing with a uniform distribution, see Wang [1995]). However, suppose the failure time T is geometrically distributed with expected value $E(T) = 1/\eta$. Then the cumulative distribution is

$$W(t) = 1 - (1-\eta)^t, \ t = 1, 2, 3 \ldots. \tag{35}$$

Furthermore, assume that observations are made at times $\tau_1 = \tau, \tau_2 = 2\tau, \tau_3 = 3\tau, \ldots$. (This equal-interval assumption is consistent with monitoring practice. The determination

of the optimal value of such checking intervals has been well studied when no monitoring is possible (see, for example, Kaio and Osaki (1989) for a comparison of methods). However, when monitoring is possible the selection of optimal checking times is still an open research question.) The associated probability that a failure will occur between observation i and $(i-1)$ is readily shown to be

$$f(i) = a(1-a)^{i-1},\ i = 1, 2, 3, \ldots, \tag{36}$$

with cumulative distribution

$$F(i) = 1 - (1-a)^{i},\ i = 1, 2, 3 \ldots, \tag{37}$$

where

$$a \equiv 1 - (1-\eta)^{\tau}. \tag{38}$$

Using this distribution, equation (24) reduces to:

$$R_n = \ell(X_n)[R_{n-1} + a], \tag{39}$$

where

$$\ell(X) \equiv L(X)/(1-a). \tag{40}$$

In this case the conditions for checking and continuing, respectively, using equation (39), are:

a) if $\ell(x_{n+1}) < \rho^*/(R_n + a)$, continue; and

b) if $\ell(x_{n+1}) \geq \rho^*/(R_n + a)$, check.

The absorption behavior of the process represented by equation (39), and in particular the distribution of the (random variable) time N until R_N first equals or exceeds the odds threshold ρ^*, has a long and important history of study[(see Shiryaev (1978)], resulting in computational (as contrasted to structural) methods for limiting cases or approximations [e.g., Pollak (1985) and Pollak and Siegmund (1975)]. In the sections that follow, we present a Markov Chain approximation that extends these results, and provides an efficient method for computing performance measures of interest.

10. Markov Process Representation for Geometric Failure Time

All performance measures of Section 4 can be obtained by computing the steady state

probabilities of a mixed continuous-discrete state Markov Process we shall call MPR, created by combining R_n with the machine condition C_n. MPR is defined such that: transitions occur immediately after observations and any associated checking; the state at the end of the n^{th} transition is denoted as $S_n \in \mathcal{S},\ n = 1, 2, \ldots$; and the state space $\mathcal{S}$ is the union of five disjoint sub-spaces – three "singleton" sets (i.e. each containing a single state) and two containing elements in the crossproduct of the open interval $(0, \rho^*)$ and $\{G, B\}$.

These five sub-spaces of $\mathcal{S}$ are (see the schematic representation in Figure A1):

$\mathcal{S}_G \equiv \{(R, G) : R \in (0, \rho^*)\}$, set of states for which $0 < R_n < \rho^*$ and $C_n = G$;

$\mathcal{S}_B \equiv \{(R, B) : R \in (0, \rho^*)\}$, set of states for which $0 < R_n < \rho^*$ and $C_n = B$;

$\mathcal{S}_0 \equiv 0$, renewal state: the state entered after the machine is renewed, i.e., when $R_n = 0$ (or, equivalently, $P_n = 0$) and $C_n = G$;

$\mathcal{S}_G^* \equiv \{(\rho^*, G)\}$, false alarm state: the state entered after checking while the machine is in G, i.e., when $R_n \geq \rho^*$ and $C_n = G$;

$\mathcal{S}_B^* \equiv \{(\rho^*, B)\}$, true alarm state: the state entered after checking while the machine is in B, i.e., when $R_n \geq \rho^*$ and $C_n = B$.

With the distributions of the random variable X_n given by equation (16) and the geometric failure time distribution of equation (36), the transition probabilities among the states in these sets are governed by the evolution of R_n described by equation (39). The key to establishing the Markovian properties of MPR is that the probability the machine goes from G to B at each transition is the constant a (except from $\mathcal{S}_G^*$ where this probability is zero).

We note some of the properties of MPR and define associated probabilities:

a) MPR ergodic, since there is a single closed communicating class of states;

b) the probability of transition from $\mathcal{S}_G^*$ or $\mathcal{S}_B^*$ to $\mathcal{S}_0$ is 1, reflecting the one transition (since $b = g = 1$) needed to check after a false alarm or a true alarm;

c) due to the geometric failure time distribution of equation (36), the single-step transition probability is a for transitions:

   i) from the set $\mathcal{S}_G$ to the set $\mathcal{S}_B \cup \mathcal{S}_B^*$;

ii) from the state $\mathcal{S}_0$ to the set $\mathcal{S}_B \cup \mathcal{S}_B^*$.

d) the state occupancy probabilities are defined to be:

i) $\pi_{0,n} = \text{prob.}\{S_n = \mathcal{S}_0\}$

ii) $\pi_{G,n}^* = \text{prob.}\{S_n = \mathcal{S}_G^*\}$

iii) $\pi_{B,n}^* = \text{prob.}\{S_n = \mathcal{S}_B^*\}$

e) the steady-state probabilities for the singleton states are

$$\begin{aligned} \pi_0 &\equiv \lim_{n\to\infty} \pi_{0,n} \\ \pi_G^* &\equiv \lim_{n\to\infty} \pi_{G,n}^* \\ \pi_B^* &\equiv \lim_{n\to\infty} \pi_{B,n}^*. \end{aligned}$$

f) the cumulative distribution functions for the elements of the sets $\mathcal{S}_G$ and $\mathcal{S}_B$ are

$$\begin{aligned} F_{G,n}(r) &= \text{prob.}\{S_n \in [(s,G) : 0 < s \le r < \rho^*]\}, \\ F_{B,n}(r) &= \text{prob.}\{S_n \in [(s,B) : 0 < s \le r < \rho^*]\}. \end{aligned}$$

g) the steady state cumulative distribution functions for the elements of the sets $\mathcal{S}_G$ and $\mathcal{S}_B$ are

$$\begin{aligned} F_G(r) &= \lim_{n\to\infty} F_{G,n}(r), \\ F_B(r) &= \lim_{n\to\infty} F_{B,n}(r). \end{aligned}$$

Performance measures are easily obtained from these steady state probabilities and distributions. In particular, by appealing to the ergodic theorem for Markov Processes, we know that

$p_f = \pi_G^* =$ expected fraction of time the process is in the false alarm state;

$p_t = \pi_B^* =$ expected fraction of time the process is in the true alarm state;

$p_0 = \pi_0 =$ fraction of time the process is in the renewal state $= p_t + p_f$ by equation (12).

Given these values for p_f and p_t, the analysis in Section 4 and Appendix C shows how to compute other measures of interest.

The equations needed to calculate π_B^* and π_G^*, given in Appendix A, are special cases of the Fredholm equation of the second kind, which has a long history of theoretical study (e.g., Groetsch (1984) or Brunner (1982)) and numerical means of solution (Schippers (1983)). Indeed these equations have an analogue to those developed by Pollak [1987] to compute an ARL0-type measure. However, as Pollak notes, a general solution method is lacking for even the simplest forms of monitoring distributions $p(x)$ and $q(x)$.

11. Markov Chain Approximation

We now present an approximation useful for finding performance measures by showing that the process MPR, which has *continuous* elements in its state space, can be approximated by a Markov *chain* ("MCR") with a *discrete* state space.

To construct MCR, we identify values of the odds ratio R_n that lie in the interval $(0, \rho^*)$, but are restricted to the finite set

$$\mathcal{S}^1 \equiv \{r_1, r_2, \ldots r_{m-1}\}.$$

The key to this restriction is finding a set of r_i values that "covers" the interval $(0, \rho^*)$ in such a way that the sums over probabilities over $r_i \in \mathcal{S}^1$ well approximate the integrals over $\mathcal{S}$ implicit in equations (A4) and (A5). (Obtaining such a set of r-values is shown, for example, in Section 12 for Bernoulli monitoring.)

Assuming that the elements of $\mathcal{S}^1$ are available, we define a finite discrete state space $\mathcal{R}$ for MCR

$$\mathcal{R} = \mathcal{S}_0 \cup \mathcal{S}_G^* \cup \mathcal{S}_B^* \cup \mathcal{R}_G \cup \mathcal{R}_B.$$

The first three sub-spaces are the singletons previously defined in Section 10 (corresponding to renewal, false alarm and true alarm states, respectively), and the last two are

$$\begin{aligned} \mathcal{R}_G &\equiv \{r_i : r_i \in \mathcal{S}^1,\ C_n = G,\ i = 1, 2, \ldots m-1,\ n = 1, 2, \ldots\} \\ \mathcal{R}_B &\equiv \{r_i : r_i \in \mathcal{S}^1,\ C_n = B,\ i = 1, 2, \ldots m-1,\ n = 1, 2, \ldots\}. \end{aligned}$$

Thus $\mathcal{R}_G$ represents a discrete subset of R_n values while the machine is in G, and $\mathcal{R}_B$ represents a discrete subset of R_n values when the machine is in B.

A simple arbitrary numbering of states allows us to represent MCR as a $(2m+1)$-state ergodic Markov Chain, which we will refer to as "MC," with

- $\sigma_n \equiv$ the state of MC after the n^{th} transition;
- state space $I_{2m+1} \equiv \{0, 1, 2, \ldots, 2m\}$; and
- transition matrix P with elements $[P]_{ij} = p_{ij} \equiv \text{prob.}\{\sigma_n = j | \sigma_{n-1} = i\}$ for $i, j \in I_{2m+1}, n = 1, 2, \ldots$.

Details of the relation between MCR and MC, and finding the values of p_{ij}, are contained in Appendix B.

Given P, and defining

$$\pi_i \equiv \lim_{n\to\infty} \text{prob.}\{\sigma_n = i\},\ i = 0, 1, \ldots 2m,$$

the steady-state probability vector $\pi \equiv \{\pi_0, \pi_1, \pi_2, \ldots, \pi_{2m}\}$ can be obtained by solving the set of linear equations:

$$\pi = \pi P \tag{41}$$

$$\pi \underline{1} = 1 \tag{42}$$

where $\underline{1} \equiv \{1\ 1\ 1\ \ldots 1\}^t$ is the transpose of the unit $(2m+1)$-vector.

Performance measures can be immediately obtained from these equations since π_0 is immediately given, and $\pi^*_G = \pi_m$ and $\pi^*_B = \pi_{2m}$. The next sections show specific results for two examples: monitoring Bernoulli and Normal observations.

12. Markov Chain MCR for Bernoulli Observations and Geometric Failure Time

This section deals the "discretization" of MPR by MC for the special case of Bernoulli observations, i.e. the observations are $X_n \in \{0, 1\}, n = 1, 2, \ldots$, and

$$p(x) = \begin{cases} 1-\alpha & \text{if } x = 0, \\ \alpha & \text{if } x = 1, \end{cases}$$

$$q(x) = \begin{cases} \beta & \text{if } x = 0, \\ 1-\beta & \text{if } x = 1. \end{cases}$$

This situation (which can be viewed as "attributes testing") is a form of classical hypothesis testing: $x = 0$ is "evidence" of condition G (e.g., no defect in an observed manufactured product) and $x = 1$ is evidence of condition B (e.g., a defect is observed). Thus α is analogous to an "error of the first kind," and β to an "error of the second kind."

The likelihood ratio is

$$L(x) = \begin{cases} \beta/(1-\alpha) & \text{if } x = 0, \\ (1-\beta)/\alpha & \text{if } x = 1. \end{cases}$$

By defining

$$\begin{aligned} w_0 &\equiv \frac{\beta}{(1-\alpha)(1-a)}, \\ w_1 &\equiv \frac{1-\beta}{\alpha(1-a)}, \end{aligned}$$

equation (39) can be written

$$R_n = \begin{cases} w_0(R_{n-1} + a) & \text{if } X_n = 0, \\ w_1(R_{n-1} + a) & \text{if } X_n = 1. \end{cases} \tag{43}$$

To identify a useful set $\mathcal{S}^1$ we consider the evolution of the R_n process of equation (43) when the process starts with $R_0 = 0$ (i.e., $P_0 = 0$). The possible values of R_n that can be generated after the first three observations, assuming none exceed ρ^*, are given in Table 1:

Observation Number n	Possible R_n Values
1	aw_0
	aw_1
2	$a(1+w_0)w_0$
	$a(1+w_0)w_1$
	$a(1+w_1)w_0$
	$a(1+w_1)w_1$
3	$a(1+(1+w_0)w_0)w_0$
	$a(1+(1+w_0)w_0)w_1$
	$a(1+(1+w_0)w_1)w_0$
	$a(1+(1+w_0)w_1)w_1$
	$a(1+(1+w_1)w_0)w_0$
	$a(1+(1+w_1)w_0)w_1$
	$a(1+(1+w_1)w_1)w_0$
	$a(1+(1+w_1)w_1)w_1$

Table 1: Possible values of R_n after $n = 1, 2, 3$ observations.

After h observations the maximum number of possible distinct values of R_n that could be generated is clearly $2^{h+1} - 2$. However, an important effect reduces this number, which allows a practical means of discretizing MCR to MC: as h becomes large, some values of R_h satisfy $R_h \geq \rho^*$, in which case either state $\mathcal{S}_G^*$ or $\mathcal{S}_B^*$ occurs and R_{h+1} becomes 0.

Thus we can generate $\mathcal{S}^1 \equiv \{r_1, r_2, \ldots, r_{m-1}\}$ by simply computing all possible R_n-values achievable over a "horizon" of h observations of $X_n, n = 1, 2, \ldots, h$. The advantages of this approach are:

a) only feasible values of r_i are generated,

b) it actually represents the R_n process for $n = 1, 2, \ldots, h$ over the horizon h, and

c) the accuracy of the discrete approximation can be improved by increasing h.

An algorithm for generating m and $\mathcal{S}^1$ is:

1. Set $S_0 = \{0\}$, a horizon $h \geq 1$ and $n = 1$.
2. Generate the set $s_0 = \{w_0(r_i + a)$ for all $r_i \in S_{n-1}$ such that $w_0(r_i + a) < \rho^*\}$.
3. Generate the set $s_1 = \{w_1(r_i + a)$ for all $r_i \in S_{n-1}$ such that $w_1(r_i + a) < \rho^*\}$.
4. Set $S_n = \{r : \ r \in s_0 \cup s_1$ and $r \notin \cup_{i=0}^{n-1} S_i\}$.
5. Increment n by 1.
6. If $n < h$ then go to step 2.
7. Set $\mathcal{S}^1 = \cup_{i=1}^{n} S_i$, $m = |\mathcal{S}^1| + 1$ and stop.

At termination $\mathcal{S}^1$ will be a set of $m-1$ distinct r-values which, when sorted by increasing value, can be labeled $r_1, r_2, \ldots, r_{m-1}$. Computational experience (see Section 13) has shown that even though m increases in h, the accuracy of the discrete approximation, for a wide range of parameters, becomes excellent for $h \leq 10$.

The key to converting MCR into MC is identifying, for any possible realized value of R_n not an element of $\mathcal{S}^1$, the "closest" element to it. Thus the entire evolution of the process, and in particular values of R_n for $n \geq h$, can approximately contained within the states $\{0,\ \mathcal{S}_G^*,\ \mathcal{S}_B^*,\ \mathcal{S}^1 \times G$ and $\mathcal{S}^1 \times B\}$.

A formal procedure for doing this is to define the finite set $W_0^* = \{0 \cup \mathcal{S}^1 \cup \rho^*\}$, with 0 as its 0^{th} element (i.e., $r_0 = 0$) and ρ^* as its m^{th} element (i.e., $r_m = \rho^*$), and define the indices

$$\begin{aligned} J_0(i) &= \text{the index of the closest element in } W_0^* \text{ to } R_{n+1} \text{ given } R_n = r_i \text{ and } x_{n+1} = 0 \\ &= \arg \min_{k \in 0,1,\ldots,m} \{|w_0(r_i + a) - r_k|\} \end{aligned}$$

and

$$\begin{aligned} J_1(i) &= \text{the index of the closest element in } W_0^* \text{ to } R_{n+1} \text{ given } R_n = r_i \text{ and } x_{n+1} = 1 \\ &= \arg \min_{k \in 0,1,\ldots m} \{|w_1(r_i + a) - r_k|\}. \end{aligned}$$

The associated probability transition matrix P can be created (see Appendix B) from two conditional probability transition submatrices P^G and P^B which correspond, respectively, to state transitions conditioned on the machine being in G and B. Matrix P is shown schematically in Figure 3.

The composition of these submatrices is:

$$\left[P^G\right]_{ij} = \begin{cases} 1-\alpha & \text{if } j = J_0(i),\ i \in \{0,1,2,\ldots,m-1\}, \\ \alpha & \text{if } j = J_1(i),\ i \in \{0,1,2,\ldots,m-1\}, \\ 0 & \text{other } i \in \{0,1,\ldots,m-1\},\ j \in \{0,1,\ldots,m\}, \end{cases} \tag{44}$$

$$\left[P^B\right]_{ij} = \begin{cases} \beta & \text{if } j = J_0(i),\ i \in \{0,1,\ldots,m-1\}, \\ 1-\beta & \text{if } j = J_1(i),\ i \in \{0,1,\ldots,m-1\}, \\ 0 & \text{other } i \in \{0,1,\ldots,m-1\},\ j \in \{1,2,\ldots,m\}. \end{cases} \tag{45}$$

13. Numerical Results for Bernoulli Monitoring

To obtain numerical results for Bernoulli monitoring we first generated the set $\mathcal{S}^1$ using the algorithm of the preceding section, and found that the number of states remains less than a few hundred for $h \leq 12$ when $\rho^* \leq .4$, $\alpha \leq .4$, and $\beta \leq .4$, conditions typical for realistic systems.

We next computed, using equations (41) and (42), the steady-state probability vector π. For example, Figure 4 shows $\pi_i, i = 0, 1, ..., m$, plotted against the values of r_i generated by the algorithm of Section 12 for $\alpha = \beta = 0.2$, $a = 0.01$, $h = 8$ and $\rho^* = 0.2$. Note the "jumpy" nature of the steady-state probabilities and the gaps between the values of r_i. This behavior suggests that using an evenly spaced grid over the r axis to represent the possible r_i values

Figure 3: Schematic representation of the transition matrix P from state i to state j showing the use of submatrices P^G and P^B. $[P^B]_{i\neq 0}$ denotes the submatrix created by removing the row associated with $i = 0$ from P^B. Shaded regions represent zero transition probabilities.

would be inefficient due to the resulting inclusion of highly unlikely (or even impossible) values of r_i.

Increasing the horizon h from 8 to 12 (which increases the number of generated states, and thus the computational effort) was shown to have little effect on the probabilities associated with the r_i values. Moreover computed values of the performance measures π_0, π_m and π_{2m} for a wide range of parameter settings, show a maximum absolute error of less than .01 for $h \geq 7$. For this reason, we used $h = 7$ for the remaining computational results.

The resulting performance measures are shown in Figures 5, 6 and 7; each shows a ROC that plots p_S = prob.{producing scrap} versus π_0 = prob.{down for checking}, for selected parameter values.

Figure 5 shows that for a fairly non-informative sensor ($\alpha = \beta = 0.3$), varying ρ^* from .01 to .5 produces a wide range of operating points (i.e., possible values of p_S and π_0). Figure 6 shows the effect of increasing sensor sensitivity to $\alpha = \beta = 0.2$. Figure 7 shows that with an even more informative sensor ($\alpha = \beta = 0.1$), only a few operating points are possible for ρ^* between 0.01 and 0.5. Indeed, for $a = 0.1$ there is only one feasible operating point

Figure 4: Probabilities π_i associated with values of r_i generated by the solution procedure when $\alpha = \beta = 0.2$, $a = 0.01$, $\rho^* = 0.2$, and a horizon of $h = 8$ is used to generate the state space.

Figure 5: System Operating Characteristics with Bernoulli monitoring generated by varying ρ^* from 0.01 to 0.5. with $\alpha = \beta = 0.3$. Points associated with the same value of ρ^* are indicated by arrows .

Figure 6: System Operating Characteristics with $\alpha = \beta = .2$. Increasing the failure probability a tends to collapse ranges of smaller ρ^* into one operating point, for example $\rho^* \in (0.01, 0.3)$ produces a wide range of operating points when $a = 0.01$, but produces only *one* operating point when $a = 0.1$.

Figure 7: System Operating Characteristics of Figure 5 with $\alpha = \beta = .1$. Note the extreme collapse of operating points to just one for all $\rho^* \in (0.01, 0.5)$ associated with a high failure probability ($a = 0.1$).

Figure 8: Decreasing β from 0.3 (as in Figure 5) to 0.2 (above) improves both performance measures.

($p_S = 0.00747$, $\pi_0 = 0.146$) in the range $.01 \leq \rho^* \leq 1$. This insensitivity to the value of ρ^* is one indicator of the robustness of the PTR policy for machines that have high probability of failure within an inter-observation interval.

Figure 8 shows a SOC for an "asymmetric" sensor with $\alpha = 0.3$, and β decreased from 0.3 (as shown in Figure 5) to 0.2. By comparing to Figure 5, we see that decreasing β reduces the detection time as well as the probability that the system is down: a lower β gives greater confidence that observing $x = 0$ implies condition G. This reduces the upward drift of the R_n process for any given set of "zero" observations, which delays the expected time until the next false alarm.

Figure 9 compares different sensors when $a = 0.1$, showing the advantage of having a more informative sensor.

Figure 10 shows an interesting alternative form of an SOC. The two attributes are p_S and prob.{Producing good product} $= p_G$. The latter measure is important since checking time is taken from production capacity even though less scrap is produced. This SOC shows operating points (above the dotted line) that will *never* be optimal with respect to these measures. For example when $a = 0.01$, there exists for each value of ρ^* above 0.26 another ρ^* below 0.26 that produces the *same* throughput of good product with a *lower* scrap rate.

Figure 9: Decreasing α and β from 0.3 to 0.1 improves the System Operating Characteristic .

14. Elements of P for Normal monitoring.

For Normal monitoring the observations X_n are independent normally distributed random variables depending only on machine condition. In particular,we assume the distributions of X_n in equation(16) are given by

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ q(x) &= \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}. \end{aligned}$$

The likelihood ratio is thus

$$L(x) = q(x)/p(x) = e^{\mu x - \mu^2/2}$$

and equation (39) becomes

$$R_n = \gamma e^{\mu X_n}[R_{n-1} + a], \tag{46}$$

where

$$\gamma \equiv \frac{e^{-\mu^2/2}}{1-a}. \tag{47}$$

Since X_n is a *continuous* random variable, the results of Appendix A can be used directly. Moreover, since both $p(x)$ and $q(x)$ are continuous functions of x, it can be shown that the

Figure 10 Operating Characteristic Curve for scrap production versus good production. "Better" is towards the point (1,0): no scrapping and always producing good product (no down time). Note the existence of "dominated" operating points: when $a = 0.01$, for example, for each ρ^* above 0.26 there exists a ρ^* below 0.26 with the same throughput of good product *and* a lower scrap rate. Similar behavior occurs when $a = 0.05$ and $a = 0.1$ at ρ^* near 0.32 and 0.36, respectively.

distribution functions $F_G(r)$ and $F_B(r)$, given by equations (A4) and (A5) respectively, are also continuous on the open interval $(0, \rho^*)$.

Defining the associated probability density functions to be:

$$f_G(r) \equiv \frac{d}{dr} F_G(r)$$

$$f_B(r) \equiv \frac{d}{dr} F_B(r)$$

and noting that the continuation region is

$$\mathcal{C}(r, y) \equiv \{x : x < \frac{1}{\mu} \ln \frac{r}{\gamma(y + a)}\}$$

equations (A4) and (A5) – after differentiating both sides – become:

$$f_G(r) = (1-a) \int_0^{\rho^*} \frac{f_G(y)}{\mu r} p\left(\frac{1}{\mu} \ln \frac{r}{\gamma(a+y)}\right) dy + \frac{(1-a)\pi_0}{\mu r} p\left(\frac{1}{\mu} \ln \frac{r}{a\gamma}\right) \quad 0 < r < \rho^* \quad (48)$$

and

$$f_B(r) = \int_0^{\rho^*} \frac{[f_B(y) + a f_G(y)]}{\mu r} q\left(\frac{1}{\mu} \ln \frac{r}{\gamma(a+y)}\right) dy + \frac{a\pi_0}{\mu r} q\left(\frac{1}{\mu} \ln \frac{r}{a\gamma}\right) \quad 0 < r < \rho^*, \quad (49)$$

the alarm probabilities can be shown to satisfy:

$$\pi_G^* = (1-a) \int_0^{\rho^*} f_G(y) \int_{\rho^*}^{\infty} \frac{1}{\mu r} p\left(\frac{1}{\mu} \ln \frac{r}{\gamma(a+y)}\right) dr\, dy + (1-a)\pi_0 \int_{\rho^*}^{\infty} \frac{1}{\mu r} p\left(\frac{1}{\mu} \ln \frac{r}{a\gamma}\right) dr \quad (50)$$

$$\pi_B^* = \int_0^{\rho^*} [f_B(y) + a f_G(y)] \int_{\rho^*}^{\infty} \frac{1}{\mu r} q\left(\frac{1}{\mu} \ln \frac{r}{\gamma(a+y)}\right) dr\, dy + a\pi_0 \int_{\rho^*}^{\infty} \frac{1}{\mu r} q\left(\frac{1}{\mu} \ln \frac{r}{a\gamma}\right) dr, \quad (51)$$

and the normalizing equation equivalent to equation (42) becomes

$$1 = \pi_0 + \int_0^{\rho^*} f_G(r) dr + \int_0^{\rho^*} f_B(r) dr + \pi_G^* + \pi_B^*. \quad (52)$$

To solve equations (48) through (52), we define

$$g(r) \equiv f_G(r)/\pi_0 \quad (53)$$

$$b(r) \equiv f_B(r)/\pi_0 \quad (54)$$

so that equations (48) and (49) become:

$$g(r) = (1-a)\int_0^{\rho^*} \frac{g(y)}{\mu r} p\left(\frac{1}{\mu}\ \ln\ \frac{r}{\gamma(a+y)}\right) dy + \frac{(1-a)}{\mu r} p\left(\frac{1}{\mu}\ln\frac{r}{a\gamma}\right) \quad 0 < r < \rho^* \qquad (55)$$

and

$$b(r) = \int_0^{\rho^*} \frac{[b(y)+ag(y)]}{\mu r} q\left(\frac{1}{\mu}\ \ln\ \frac{r}{\gamma(a+y)}\right) dy + \frac{a}{\mu r} q\left(\frac{1}{\mu}\ln\frac{r}{a\gamma}\right) \quad 0 < r < \rho^*. \qquad (56)$$

Equations (55) and (56) are a system of Fredholm equations of the second kind. The key to their solution is the nature of their kernels, that is the behavior of

$$\begin{aligned} k_G(y,r) &= \frac{1}{\mu r} p\left(\frac{1}{\mu}\ln\frac{r}{\gamma(a+y)}\right) \\ &= \frac{1}{\mu r\sqrt{2\pi}} e^{-\frac{1}{2\mu^2}\ln^2\frac{r}{\gamma(a+y)}} \end{aligned}$$

and

$$k_B(y,r) = \frac{1}{\mu r\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{1}{\mu}\ln\frac{r}{\gamma(a+y)} - \mu\right)^2}$$

A detailed study of the behavior of these equations is given in Jorna and Pollock (1998). The numerical results shown below were produced by approximating $g(\cdot)$ and $b(\cdot)$ by n^{th} order polynomials formed by truncating a Chebychev series. In particular,we find the polynomial coefficients $c_i^G, i = 1, 2, \ldots, n$ such that

$$g(r) \approx \sum_{i=1}^{n} c_i^G \cos\left((i-1)\cos^{-1}\left(\frac{2r}{\rho^*} - 1\right)\right)$$

with the property

$$\int_{r=0}^{\rho^*} g(r)\, dr \approx \sum_{i=1;\, i \text{ odd}}^{n} \frac{c_i^G \rho^*}{1-(i-1)^2}. \qquad (57)$$

The function $g(x_i)$ can then be evaluated over the set of Chebyshev points x_i, where

$$x_i = \frac{\rho^*}{2}\left(1 + \cos\left(\frac{\pi(i-1)}{n-1}\right)\right) \qquad i = 1, 2, \ldots, n.$$

Figure 11: The conditional density function $g(r) = f_G(r)/\pi_0$ for Normal monitoring, analogous to the plot of π_i for Bernoulli monitoring shown in Figure 4.

Equation (49) is solved, in a similar way, by approximating $b(\cdot)$ with a different Chebychev series and using the $g(\cdot)$ obtained above. That is, the coefficients $c_i^B, i = 1, 2, \ldots, n$, are found such that

$$b(r) \approx \sum_{i=1}^{n} c_i^B \cos\left((i-1)\cos^{-1}\left(\frac{2r}{\rho^*} - 1\right)\right),$$

with the property

$$\int_{r=0}^{\rho^*} b(r)\, dr \approx \sum_{i=1;\, i \text{ odd}}^{n} \frac{c_i^B \rho^*}{1-(i-1)^2}. \tag{58}$$

The performance measures π_G^* and π_B^* are obtained by trapezodial approximation of the integrals in equations (50) and (51) using the values of $g(x_i)$ and $b(x_i)$ at the Chebyshev points x_i. Finally, π_0 is obtained by using equations (57) and (58) in equation (52).

15. Numerical Results for Normal Monitoring

Figure 11 shows $g(r)$ for Normal monitoring with $a = .05, \mu = .5$ and $\rho^* = .2$. Compared to the π_i obtained for Bernoulli monitoring (Figure 4) this distribution is smooth and well behaved except near zero (where it can be shown that $\lim_{r\to 0} F_G(r) = 0$.) However, the behavior of $g(r)$ and $b(r)$ for other parameter values, particularly for large a and small μ, can be extremely erratic (see Jorna and Pollock (1998)).

Figure 12 shows the System Operating Characteristic curves for p_S = prob.{Producing scrap} versus π_0 = prob.{System Down} for $a = 0.05$ and 0.1 while fixing $\mu = 1.5$. As in Figure 5, there is an improvement in the SOC with smaller a.

Figure 12: Operating Characteristic for Normal monitoring, $\mu = 1.5$, for different failure probabilities

Figure 13 shows SOC sensitivity to changing μ, the shift in the expected observation value when the system fails. As μ increases the power of the sensor to discriminate between conditions G and B increases and this improves the SOC curve. This has obvious implications in evaluating sensors with different μ values.

Figure 14 shows an alternative set of SOC curves: a plot of Prob.{Producing scrap} versus Prob.{Producing good product} for $a = 0.05$ and $\mu \in \{0.5, 1.0, 1.5\}$. Note that when the production utilization (i.e. the probability of producing good parts) is about .8, trying to increase this (by decreasing the threshold ρ^*) produces an increase in scrap without appreciable improvement in the production rate.

16. Conclusion

The System Operating Characteristic is an important and evocative tool for the comparison of monitoring policies and for the comparison of alternative observation technologies. The structure set forth in this paper provides a Markov chain based method for the computation of critical performance measures needed to express SOCs when sample observations are Bernoulli random variables. For more general sampling functions $p(x)$ and $q(x)$, Appendix A presents equations that can be solved to find key performance measures.

Figure 13: System Operating Characteristic for Normal monitoring, $a = .05$, for different mean shifts in monitored signal μ.

Figure 14: Alternative Operating Characteristic for Normal monitoring, $a = .05$, for different mean shifts in monitored signal μ.

Acknowledgments

This research was supported jointly by General Motors Research Laboratories and the National Science Foundation under contract # SES9108956. The support of Walter Albers, Jr., who was instrumental in arranging for this unique funding collaboration, is greatly appreciated. The presentation was also appreciably strengthened by many substantive discussions with Richard Marcellus. Much of the computation was performed using code developed by Wei-Ching Wang and Pieter Jorna.

APPENDIX A: Steady State Properties of MPR

In this Appendix, we show how the Chapman-Kolmogorov (C-K) equations can be used to write expressions for steady-state probabilities and distributions for MPR defined in section 10. Although the notation and details may appear formidable, the method is a straightforward extension of the use of C-K equations for finding steady-state solutions to a finite state Markov chain. In the development that follows, it might be helpful to refer to the schematic flow diagram of Figure A1. In this diagram, the "states" $\{\mathcal{R}_G\}$ and $\{\mathcal{R}_B\}$ refer to R-values from equation (39) in the open interval $(0, \rho^*)$ while the machine is in condition G and B, respectively. S_G^*, S_B^*, and 0 are the singleton "false alarm," "true alarm," and "renewal" states as discussed in Section 10. The labels on the transition arrows from singleton states represent governing probabilities; the arrows from the sets of states $\{\mathcal{R}_G\}$ and $\{\mathcal{R}_B\}$, represent the complementary distribution functions $\overline{P}$ and $\overline{Q}$ associated with exceeding the threshold ρ^*, given machine condition G and B, respectively.

Figure A1: Schematic representation of transitions among the states in MPR. Note that 0, S_B^* and S_G^* are singleton states, while $\{\mathcal{R}_G\}$ and $\{\mathcal{R}_B\}$ represent a continuum of states in the open interval $(0, \rho^*)$.

The distribution function $F_{G,n}(r)$ is computed by conditioning on the value of $R_{n-1} = y$,

and noting that $C_n = G$ is only possible if $C_{n-1} = G$. This gives

$$\begin{aligned} F_{G,n}(r) &= \int_{y=0^+}^{\rho^*} \text{prob.}\{0 < R_n \le r \cap C_n = G | R_{n-1} = y \cap C_{n-1} = G\} dF_{G,n-1}(y) \\ &\quad + \pi_{0,n-1} \text{prob.}\{0 < R_n \le r \cap C_n = G | R_{n-1} = 0 \cap C_{n-1} = G\} \end{aligned} \tag{A1}$$

The first term in the integrand of (A1) can be written:

$$\text{prob.}\{0 < R_n \le r | C_n = G \cap R_{n-1} = y \cap C_{n-1} = G\} \text{prob.}\{C_n = G | R_{n-1} = y \cap C_{n-1} = G\}.$$

Using equation (39) (which governs the behavior of R_n when $R_{n-1} = y$), and the fact that $\text{prob.}\{C_n = G | C_{n-1} = G\} = 1 - a$ (which is independent of the value of R_{n-1}), gives

$$\text{prob.}\{0 < R_n \le r \cap C_n = G | R_{n-1} = y \cap C_{n-1} = G\} = (1-a)\text{prob.}\{0 < \ell(X_n)(y+a) \le r | C_n = G\}.$$

Thus

$$\begin{aligned} F_{G,n}(r) &= \int_{y=0^+}^{\rho^*} (1-a)\ \text{prob}\{0 < \ell(X_n)(y+a) \le r | C_n = G\} dF_{G,n-1}(y) \\ &\quad + (1-a)\pi_{0,n-1}\ \text{prob.}\{0 < \ell(X_n)a \le r | C_n = G\}. \end{aligned} \tag{A2}$$

By defining the region $\mathcal{C}(r,y) \equiv \{x :\ \ell(x) < r/(a+y)\}$, $\mathcal{C}(\rho^*, y)$ becomes the set of "continuation" values of the observation x_n. Since $p(x)$, the p.d.f. for X_n given $C_n = G$, is independent of n,

$$F_{G,n}(r) = (1-a) \int_{y=0}^{\rho^*} \int_{x \in \mathcal{C}(r,y)} p(x)\, dx\, dF_{G,n-1}(y) + (1-a)\pi_{0,n-1} \int_{x \in \mathcal{C}(r,0)} p(x)\, dx. \tag{A3}$$

Taking the limit of both sides of equation (A3) as $n \to \infty$ gives the steady state distribution

$$F_G(r) = (1-a) \int_{y=0}^{\rho^*} \int_{x \in \mathcal{C}(r,y)} p(x)\, dx\, dF_G(y) + (1-a)\pi_0 \int_{x \in \mathcal{C}(r,0)} p(x)\, dx. \tag{A4}$$

Similarly, it can be shown that

$$\begin{aligned} F_B(r) &= a \int_{y=0}^{\rho^*} \int_{x \in \mathcal{C}(r,y)} q(x)\, dx\, dF_G(y) + a\pi_0 \int_{x \in \mathcal{C}(r,0)} q(x)\, dx \\ &\quad + \int_{y=0}^{\rho^*} \int_{x \in \mathcal{C}(r,y)} q(x) dx dF_B(y) \end{aligned} \tag{A5}$$

$$\pi_G^* = \int_{y=0}^{\rho^*} \int_{x \notin \mathcal{C}(\rho^*, y)} (1-a)p(x)\, dx\, dF_G(y) + \pi_0 \int_{x \notin \mathcal{C}(\rho^*, 0)} (1-a)p(x)\, dx; \tag{A6}$$

$$\pi_B^* = \int_{y=0}^{\rho^*} \int_{x \notin \mathcal{C}(\rho^*, y)} aq(x)\, dx\, dF_G(y) + \pi_0 \int_{x \notin \mathcal{C}(\rho^*, 0)} aq(x)\, dx + \int_{y=0}^{\rho^*} \int_{x \notin \mathcal{C}(\rho^*, y)} q(x)\, dx\, dF_B(y) \tag{A7}$$

and, finally,

$$\begin{aligned} \pi_0 &= \pi_G^* + \pi_B^* \\ &= \int_{y=0}^{\rho^*} \int_{x \notin \mathcal{C}(\rho^*, y)} [aq(x) + (1-a)p(x)]\, dx\, dF_G(y) \\ &\quad + \int_{y=0}^{\rho^*} \int_{x \notin \mathcal{C}(\rho^*, y)} q(x)\, dx\, dF_B(y) + \pi_0 \int_{x \notin \mathcal{C}(\rho^*, 0)} [aq(x) + (1-a)p(x)]\, dx. \end{aligned} \tag{A8}$$

Figure B1: Schematic representation of the MC states, the MCR sets, and the R-values generated from equation (39).

APPENDIX B: Correspondence between Formulations MC and MCR

The correspondence between elements of $I_{2m+1} \in \{0, 1, 2, \ldots, 2m\}$ of the chain MC and of the elements MCR which lie in the set $\mathcal{S}^1 \times \{G, B\}$ are shown in Figure B1.

State 0 is the "starting" state of MC; since it represents the situation where $P_n = 0$ (and thus $R_n = 0$) it is also the "renewal" state. State m is the false alarm state, since $R_n = \rho^*$ with machine condition G; state $2m$ is the true alarm state, since $R_n = \rho^*$ with machine condition B. The set $\{\mathcal{R}_G\}$ represent states where R_n lies between 0 and ρ^* with machine condition G; $\{\mathcal{R}_B\}$ represents states where R_n lies between 0 and ρ^* with machine condition B.

Transition probabilities for MC, i.e. among the states in I_{2m+1}, are obtained by noting

that:

a) when $R_{n+1} \geq \rho^*$, the state entered after transition $n+1$ is either m or $2m$, depending whether machine condition is either G or B;

b) once in state m or $2m$, since $b = g = 1$, the next transition is into state 0 with probability 1;

c) for any value of $R_n < \rho^*$, if the machine condition is G then with probability a condition B will hold on the next transition;

d) all transitions from condition B to condition G must be made via the true alarm state $2m$.

Thus, transitions from alarm states to the renewal state are

$$\begin{aligned} p_{m,0} &= 1, \\ p_{2m,0} &= 1, \\ p_{m,j} &= 0 \qquad \text{if } j \neq 0, \\ p_{2m,j} &= 0 \qquad \text{if } j \neq 0, \end{aligned}$$

and $p_{ij} = 0$ for $i = m+1, m+2, \ldots, 2m-1;\ j = 0, 1, 2, \ldots, m$.

For the rest of the elements of P, we define the $m \times m$ sub-matrices P^G and P^B, such that

$$\begin{aligned} \left[P^G\right]_{ij} &= \text{prob.}\{\sigma_n = j | \sigma_{n-1} = i \cap C_n = C_{n-1} = G\} \qquad \text{for } i = 0, 1, \ldots, m-1;\ j = 0, 1, \ldots, m, \\ \left[P^B\right]_{ij} &= \text{prob.}\{\sigma_n = j+m | \sigma_{n-1} = i \cap C_n = C_{n-1} = B\} \quad \text{for } i = 0, 1, \ldots, m-1;\ j = 1, 2, \ldots, m, \end{aligned}$$

where σ_n is the state at the end of the n^{th} transition.

The elements of these submatrices are given by the nature of the distributions $p(\cdot)$ and $q(\cdot)$ of equation (16). In terms of these submatrices, the remaining elements of P are given by:

$$\begin{aligned} p_{ij} &= (1-a)[P^G]_{ij} \qquad \text{for } i = 0, 1, \ldots, m-1;\ j = 0, 1, \ldots, m \\ p_{ij} &= a[P^B]_{i,j-m} \qquad \text{for } i = 0, 1, \ldots, m-1;\ j = m+1, \ldots, 2m, \\ p_{ij} &= [P^B]_{i-m,j-m} \qquad \text{for } i = m+1, m+2, \ldots, 2m-1;\ j = m+1, m+2, \ldots, 2m. \end{aligned}$$

The first equation reflects transitions from $C_{n-1} = G$ to $C_n = G$; the second represents transitions from $C_{n-1} = G$ to $C_n = B$; the third equation reflects transitions while the machine condition is B. Matrix P is shown schematically in Figure 3.

APPENDIX C: Performance Measures for Arbitrary b and g

In our development, we assumed checking times of $b = g = 1$. Calculating performance measures for arbitrary b and g for condition B or G, respectively, is described here.

Recalling that $\overline{L}$ is the expected cycle time, we can re-express equation (3) as:

$$\overline{L}(b, g) = E(T) + \mu g + \delta + b$$

which explicitly incorporates as arguments of $\overline{L}$ the checking times b and g, as well as μ, the expected number of false alarms in a cycle.

By definition, let $p_f(b, g)$ be the fraction of time the machine is in the false alarm state. For arbitrary b and g, the cycle time has three components:

a) $\overline{L}(1, 1)$;

b) $b - 1$, the amount of checking time while in B not accounted for in $\overline{L}(1, 1)$;

c) $p_f(1, 1)\overline{L}(1, 1)(g - 1)$, the time consumed in checking false alarms not accounted for in $\overline{L}(1, 1)$, making use of the fact that, when $b = g = 1$, the expected number of false alarms per cycle is $\mu = p_f(1, 1)\overline{L}(1, 1)$.

Hence, the expected cycle time for arbitrary b and g can be written

$$\overline{L}(b, g) = \overline{L}(1, 1) + (b - 1) + p_f(1, 1)\overline{L}(1, 1)(g - 1).$$

Using this result, the fraction of time spent in the false alarm state is

$$p_f(b, g) = \frac{p_f(1, 1)\overline{L}(1, 1)g}{\overline{L}(b, g)},$$

and the fraction of time spent in the true alarm state is

$$p_t(b, g) = \frac{b}{\overline{L}(b, g)}.$$

References

Barlow, R. E., Hunter, L. C. and Proschan, F.,"Optimum Checking Procedures", *J. Soc. Indust. Appl. Math.*, 11, 1078-1095 (1963)

Basseville, M. and Nikiforov, I.V., *Detection of Abrupt Changes: Theory and Application*, Prentice Hall, (1993).

Barlow, R. E., and Proschan, F., *Mathematical Theory of Reliability*, Wiley, New York (1965)

Brunner, H. "A Survey of Recent Advances in the Numerical Treatment of the Volterra Integral and Integro-differential Equations", *Journal Comput. Appl. Math*, 8, 213-229 (1982).

Duncan, A. J., "The Economic Design of $\overline{X}$ Charts Used to Maintain Current Control of a Process," *Journal of the American Statistical Association*, 51, 228–242, (1956)

Girschick, M.A. and Rubin, H. "A Bayes Approach to a Quality Control Model," *Ann. Math. Stat.*, Vol. 23, 114-125 (1952)

Groetsch, C. W., "The theory of Tikhonov Regularization for Fredholm Equations," 104p, Boston Pitman Publication (1984).

Johnson, N.L. and Leone, F.C., "Cumulative Sum Control Charts: Mathematical Principles Applied to Their Construction and Use," Parts I, II, III. *Industrial Quality Control*, Vol. 18, 15-21; Vol. 19, 29-36; Vol 20, 22-28, (1962).

Jorna, P., and Pollock S.M., "Solving a Pair of Linked Fredholm equations Originating in a Machine Monitoring Problem", Technical Report 98-12, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI (1998)

Kaio, N. and Osaki, S., "Comparison of Inspection Policies", *Journal of the Operational Research Society*, V. 40, No. 5, 499-503 (1989)

Lele, S., "Steady-State Analysis of Three Process Monitoring Procedures in Quality Control", Ph. D. Dissertation, University of Michigan, Ann Arbor, MI, (1996)

Lorenzen, T. J. and Vance, L.C. , "The Economic Design of Control Charts: A Unified Approach", *Technometrics*, 28, 3-10, (1986)

Marcellus, R. L. and Jasmani, Z., "A Comparative Study of Cusum Control Charts and Bayesian Process Control," in *Productivity and Quality Management Frontiers - III*, Institute of Industrial Engineers, Norcross, GA. (1991)

Montgomery, D.C. "The Economic Design of Control Charts: A Review and Literature Survey," *Journal of Quality Technology*, Vol. 12, No. 2 75-87 (1980).

Montgomery, D C. *Introduction to Statistical Quality Control*, third edition, John Wiley and Sons, New York. (1996),

Moskowitz, H., Plante, R. and Chun, Y.H. "Economic Design of Continuous Shift Model $\overline{X}$ Process Control Charts," Krannert Graduate School of Management, Purdue University (1989).

Page, E. S., "Continuous Inspection Schemes," *Biometrika* 41, 100–115, (1954)

Pollak, M. "Average Run Lengths of an Optimal Method of Detecting a Change in Distribution," *Ann. Stat.*, Vol. 15, No. 2, 749-779 (1987).

Pollak, M. and Siegmund, D., "Approximations to the Expected Sample Size of Certain Sequential Tests," *Ann. Stat.*, Vol. 6, 1267-1282, (1975).

Pollak, M. "Optimal Detection of A Change in Distribution," *Ann. Stat.*, Vol. 13, No. 1, 206-227 (1985).

Pollock, S.M., "Minimum Cost Checking Using Imperfect Information," *Management Science*, Vol. 13, No 7, pp 206-227 (1965).

Rapoport, A. and Burkheimer,G.J., "Parameters of Discrete Time Models of Detection of Change," *Management Science*, Vol. 19, No. 9, 973-984, (1973)

Roberts, S.W., "A Comparison of Some Control Chart Procedures," *Technometrics*, Vol. 8, 411-430 (1966).

Saniga, I. M., Economic Statistical Control Chart Designs with Aplications to $\overline{X}$ and R Charts", *Technometrics* 31: 313-320, August 1989

Schippers, H., "Multiple Grid Methods for Equation of the Second Kind," Amsterdam, Mathematisch Centrum, 133p (1983)

Shewhart, W.A., "*The Economic Control of the Quality of Manufactured Product*," Macmillan, New York, (1931).

Shiryaev, A.N., "*Optimal Stopping Rules*," Springer-Verlag, New York, (1978).

Shiryaev, A.N. "On Optimum Methods in Quickest Detection Problems," *Prob. Appl.*, Vol. 8, 22-46 (1963).

Svoboda, L., "Economic Design of Control Charts: A Review and Literature Survey," in *Statistical Process Control in Manufacturing*, Marcel Dekker, Inc., New York. (1991)

Wang, W.-C., "Detection of Process Change with Non-Geometric Failure Time Distribution", Ph. D. Dissertation, University of Michigan, Ann Arbor, MI (1995)

Woodall, W. H.,"Weaknesses of the Economic Design of Control Charts," *Technometrics* 28(4), 408–409, (1986)

Woodall, W. H., "The Statistical Design of Quality Control Charts," *The Statistician* 34, 155–160,(1985)