OPERATING CHARACTERISTICS FOR DETECTION OF PROCESS CHANGE PART I: THEORETICAL DEVELOPMENT

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

and

Jeffrey M. Alden
General Motors Research Laboratories
Warren, MI 48090

August 1992

Technical Report 92-34

1. Introduction

Detecting the occurrence of an underlying process change, when observations are related only probabilistically to the state of the process, has long been a concern of statisticians, engineers, economists, epidemiologists, etc. In this paper we present an analysis of this "detection of change" problem in the context of machine monitoring. Applications to areas such as quality control, health, military surveillance, or economic analysis should be readily apparent.

Fundamentally, the problem is to determine when a system goes "out of control" or "fails" and to do this "as soon as possible." We use the term "policy" to describe a procedure that prescribes when to take an "action" consistent with a process change, after observing available information. Usually, an economic or other advantage is gained when a policy raises an "alarm" that detects the change soon after it happens. However, it is also desirable to avoid costly "false alarms" that may occur while the system is still "in control," i.e., when no process change has occured.

The balancing off of these two criteria – quick detection and few false alarms – is the basis of most quality control and control chart procedures (e.g., CUSUM) developed over the past sixty years (see Shewhart (1981), Roberts (1966), Johnson and Leone (1962), Montgomery (1980), for a historical perspective and related formulations). The most common approach to "optimize" these procedures (e.g. Moskowitz, Plante and Chun (1989)) uses an economic model that explicitly assumes costs are ascribable to these two criteria, or to equivalent operational measures of the procedure. These economic approaches depend upon specific policy structures which, although intuitively appealing, and lead to easy computation or evocative charting methods, are not necessarily optimal in any sense, or even, in some cases, consistent.

Thus, a full understanding of how to effectively and appropriately operate a monitoring (or "surveillance" or "inspection") system, and how to compute the associated performance

measures, is still lacking. As a step towards such an understanding, we address four issues, in the context of a rather simple but generalizable framework, that are relatively neglected in both the literature and practice.

1. A traditional performance measure is the *average run length* (ARL): the expected time until a false alarm occurs given the system remains "in control." This attempts to capture the relative cost associated with false alarms — a policy with a longer ARL typically engenders lower false alarm costs. We propose instead to use the reasonable, computable, and economically supportable measures of "false alarm rate" and "fraction of time processing false alarms" during the entire monitoring process. These measures contribute to the typical costs associated with false alarms: a fixed cost per false alarm (e.g., to allocate resources) which is captured by the false alarm rate, and a variable cost for time spent processing false alarms (e.g., labor hours and lost operating time) which is captured by the fraction of time spent processing false alarms.

2. A traditional "quickness of detection" measure is the expected time to detect that the system has gone "out of control" or "failed." This detection time is used to estimate the cost of operating the system while in a failed condition, or "down" (e.g., while producing scrap). However, there is more. Typically failure costs also depend on the failure occurance (e.g., replacement part costs) and on the time spent after detection to renew the system (e.g., lost production time). To understand total costs over time, the frequency of these cost-incurring events is needed, particularly when one wishes to minimize average cost per unit time. These costs are captured by the failure frequency or "true alarm rate" (assuming all failures are eventually detected), the fraction of time the system is down (i.e., operating while in a failed condition or halted for renewal after a failure detection), and the expected time to detect a failure. Hence, we focus on calculating the true alarm rate, fraction of time down, and expected detection time

while routinely using the monitoring policy. This is in constrast to similar measures traditionally computed given the monitoring process starts when the system first fails.

3. Most analyses of monitoring system performance produce only asymptotic results, involve computational difficulties, are relatively complex and difficult to explain to potential users, or require unrealistic conditions on sampling processes. Our approach, instead, provides a method for the computation of operating performance measures which (although approximate) is computationally straightforward, easy to understand and implement, and holds for a wide range of parameters.

4. Our approach allows sensitivity analysis with respect to important parameters such as the expected operating time until system failure and the relative discriminatory power of various sampling devices. We do not require an explicit assessment of hard-to-estimate costs associated with false alarms and late detections.

2. Formulation of the Basic Monitoring Problem

Consider a system that can be in one of two conditions, G or B (for "Good" and "Bad"). The system starts in G and enters B after some operating time T, where T is a discrete random variable (r.v.) with probability mass function (p.m.f.) $f(t)$; $t = 1, 2, \ldots$. Once the system enters B, it remains there until some exogenous action is taken. When this action, called *checking*, is taken the system is halted (i.e., it stops operating) and its condition is assessed with certainty. Checking may find the system in G or B and, in either case, we assume it is *renewed* (i.e., made as "good-as-new") before operations continue. The time it takes to renew after checking depends on system condition. In particular, this time is c or b if the system is found in G or B, respectively. Typically $c < b$ since renewal from B often requires fixing something while renewal from G may only require a brief inspection to ascertain it is in G. Since the system is halted during renewals, the operating time T until failure differs from the system time until failure. The *system* time until failure is T plus the total time spent checking and renewing (while in G) prior to failure.

For convenience, we use language appropriate to a manufacturing process that starts producing good parts (hence "G") and at some operating time T later it "fails" and then produces bad (hence "B") parts or no parts at all. We do <u>not</u> consider the explicit economic or physical consequences of the process while in either condition. As we will see, these can be subsumed into appropriate performance measures of a procedure for determining if the system has entered B.

We use the following notation:

$\{G, B\}$ = set of possible system conditions,

$C_t \equiv$ condition of the system at time t, $t = 0, 1, 2, \ldots$, and

$P_t \equiv \text{prob}\{C_t = B | C_0 = G\}$, $t = 1, 2, 3, \ldots$.

We now consider the use of an (imperfect) information gathering procedure with a sequence of events as shown in Figure 1. At every observation time $\tau_i = 1, 2, 3, \ldots$ a random variable X_i is observed with probability density function (p.d.f.) $f_{X_i}(\cdot)$. This "monitoring" of X_i provides information about system condition since its p.d.f. depends upon the condition. This information is used to immediately "update" $f(t)$ and determine if checking should be done.

In particular, let

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

define the random vector $\underline{X}_n$ as

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

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

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

and define

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

It is well known (e.g., see Girshik and Rubin [1952], Shiryaev [1963], or Pollock [1965] for early references, or Pollak [1987] for a more recent one) that for this situation, the following form of decision rule is optimal (in that it minimizes average cost per unit time) for a variety of reasonable economic and/or statistical criteria:

> Probability Threshold Rule (PTR): Decide (or take appropriate action consistent with such a decision) that the system is in B when $P_n(\underline{x}_n)$ equals or exceeds some threshold p^*.

Event Flow Chart

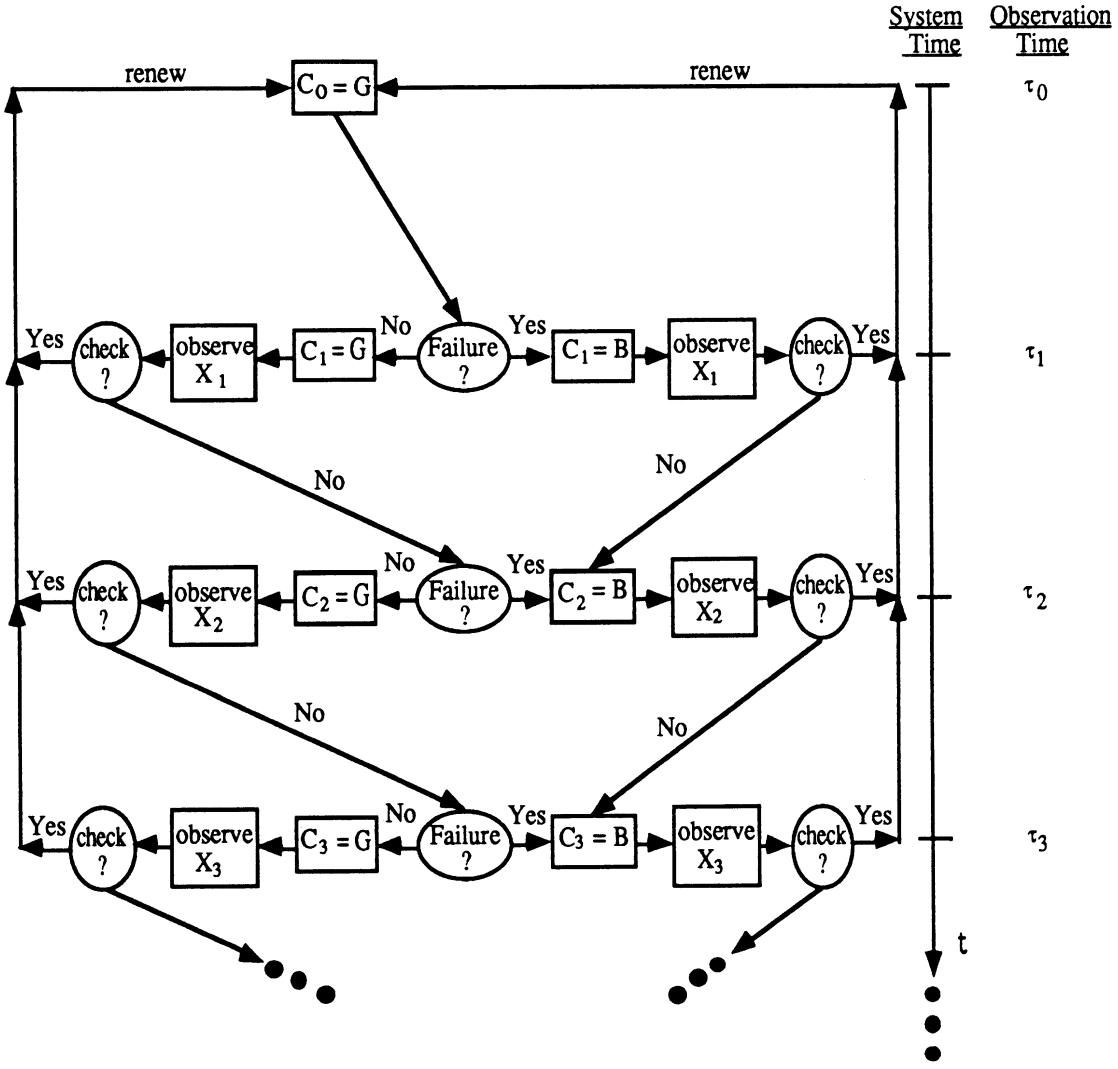

Figure 1: Sequence of events involving possible system "failure," monitoring by observing r.v. X_i, and deciding whether or not to check.

The PTR policy effectively sets the structure of the monitoring procedure: continuously compute $P_n(\underline{x}_n)$ and react when it equals or exceeds p^*. By varying the threshold p^*, the decision maker, given the ability to compute $P_n(\underline{x}_n)$, can explore trade-offs among various criteria and performance measures. Note that there is only this one parameter, p^*, to vary.

The event when $P_n(\underline{x}_n)$ equals or exceeds p^* is called an *alarm*; subsequent exploration to determine the system's actual condition is called *checking*; and making the system "as good as new" is called *renewing*. An alarm is called a *false alarm* if, upon checking, the system is found to be in G, otherwise it is a *true alarm*.

This paper first develops a way to conveniently compute $P_n(\underline{x}_n)$ for any set of observed values $\{x_1, x_2, \ldots x_n\}$. Given this "probability the system has failed given $\underline{x}_n$," we compute five performance measures arising from two competing criteria that characterize the "goodness" of the procedure: the cost associated with false alarms and the cost associated with failures. The first is measured by the false alarm rate and the fraction of time spent processing false alarms. The latter is measured by the true alarm rate, the fraction of time down (i.e., in B), and the expected detection time. These performance measures are formally defined as:

r_f = false alarm rate: the expected number of false alarms per unit time,

p_f = fraction of time spent processing false alarms,

r_t = true alarm rate: the expected number of true alarms per unit of time,

p_B = fraction of time the system is down (in B), and

δ = expected detection time: the expected time after failure until $P_n(\underline{x}_n)$ equals or exceeds the threshold p^*.

These measures capture the essence of the operational performance of PTR: $P_n(\underline{x}_n)$ allows the rule to be followed (given any assigned value of p^*); r_t, p_B, and δ represent the waste

due to failures (failure frequency, fraction of time down, and expected detection time); r_f and p_f represent the waste due to false alarms (false alarm frequency and fraction of time spent processing false alarms).

These measures also support a simple yet reasonable cost model. Given the following costs:

K_f = fixed cost per false alarm,

V_f = variable cost per unit of time spent processing a false alarm,

K_t = fixed cost per true alarm,

V_t = variable cost per unit of time spent while the system is down and a failure has been detected, and

V_d = variable cost per unit of time while the system is down and no failure has been detected,

then the average cost rate for the procedure can often be expressed as

$$K_f r_f + V_f p_f + K_t r_t + V_t(p_B - r_t\delta) + V_d r_t \delta.$$

Even if such a linear cost model model or cost parameters can not be specified, it is important to note that these measures serve to summarize the behavior of the policy.

An excellent way to summarize the performance of *any* procedure that purports to be appropriate for monitoring a system change process is to develop an Operating Characteristic (OC) for the procedure. Similiar to the Receiver Operating Characteristic, used in telecommunication and signal detection theory, and related to the power curve of simple hypothesis testing, an OC is simply a plot of achievable levels of two competing performance measures.

For example, an OC could plot the expected fraction of time down, p_B, versus false alarm rate, r_f. Consider such OC's for three different hypothetical monitoring procedures A, B

and C, as shown in Figure 2. Each curve represents the operating points (i.e., values of p_B and r_f) achievable by varying free parameters within the procedures. In the PTR policy, this is done by varying the threshold p^* to produce large (near 1) p_B and small r_f when $p^* \rightarrow 1$, and small p_B and large r_f when $p^* \rightarrow 0$.

Procedure C is clearly better than A or B, since it has a lower p_B for any given r_f or lower r_f for any given p_B. (Perhaps procedure C is based upon observing variables x_i that are not available to A or B). Procedure A would be preferred to B only in those situations when early detection (i.e. small p_B) is more important than having high false alarm rates.

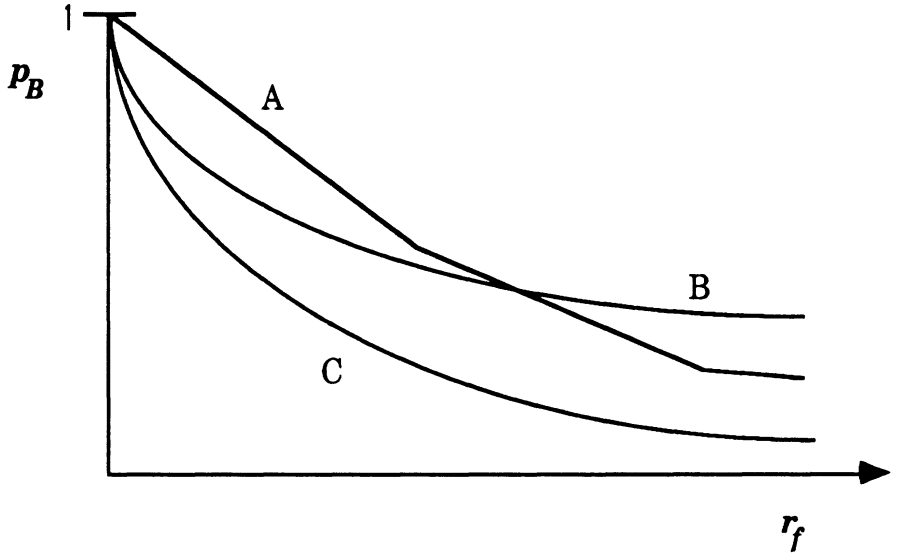

Figure 2: Operating Characteristics (OC) for three hypothetical procedures A, B and C.

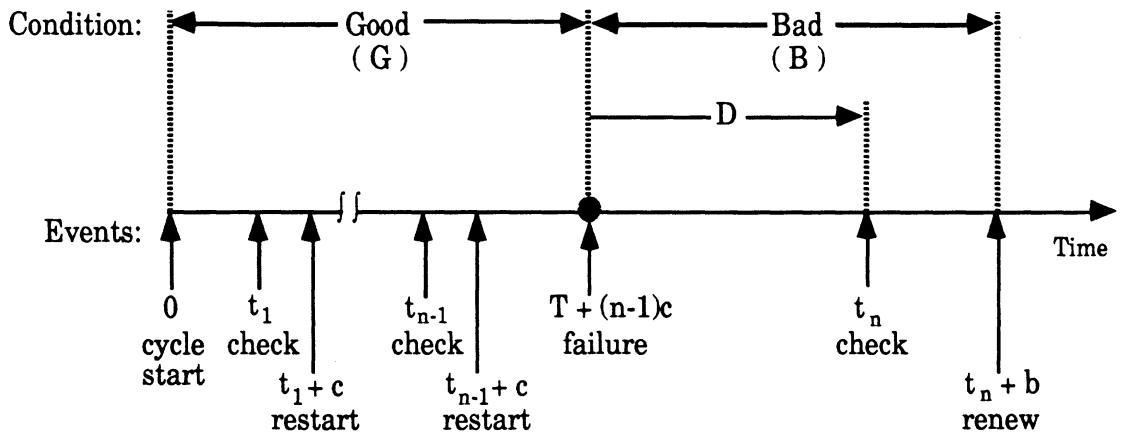

Figure 3: A single cycle of the checking/monitoring procedure. Arrows indicate checking actions and renewals; T = operating time until failure which occurs at system time $T + (n-1)c$; D = detection or response time; t_i = time of i^{th} check — a false alarm for $i < n$; t_n = time condition B is detected; and $t_n + b$ is the time the system re-enters condition G which ends the cycle.

3. Some Relations Among Performance Measures

For a process governed by failure time p.m.f. $f(t)$, where t is the operating time until failure, and *any* reasonable checking/monitoring procedure, there exist general relations among the performance measures. Figure 3 shows a single *cycle* of a procedure that starts with a renewal and ends with a renewal following the first check that finds the system in B. If this cycle contains n checks at times $t_1, t_2, \cdots, t_n$, then $n-1$ checks find the system in G, i.e., there are $n-1$ false alarms, each requiring a time c to process.

As shown in Figure 3, D is the (random variable) response time. The performance measure δ is $E(D)$ by definition, and the expected cycle length is

$$C(b,c) = E(T) + c(n-1) + \delta + b$$

where $E(T) \equiv \sum tf(t)$ is the expected operating time until system failure. Note, the system remains in B during the renewal time interval of length b, and it cannot fail during the

checking time interval(s) of length c while in G.

Using the fundamental renewal theorem, if μ is the expected number of false alarms until failure, then

$$r_f = \frac{\mu}{E(T) + c\mu + \delta + b} \tag{2}$$

and

$$r_t = \frac{1}{E(T) + c\mu + \delta + b}. \tag{3}$$

The ratio of these two equations gives

$$\mu = \frac{r_f}{r_t}. \tag{4}$$

Solving for δ in equation (3) gives

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

Since $\delta + b$ is the time in B which occurs at rate r_t, and c is the lost operating time for each false alarm which occurs at rate r_f,

$$p_B = (\delta + b)r_t, \tag{6}$$

$$p_f = cr_f. \tag{7}$$

These relationships hold for any checking or monitoring procedure. If r_t and r_f, can be calculated, then given the parameters $E(T)$, b, and c, we can calculate μ, δ, p_B, and p_f.

A traditionally used Average Run Length measure, ϕ_G, is the expected time to first alarm given the system remains in condition G. In spite of its popularity, using ϕ_G as a performance measure of a monitoring procedure is questionable for several reasons:

a) the hypothetical situation where the system is "forced" to remain in G until the first alarm (an assumption required for computing ϕ_G) is hard to justify as a realistic one, and in any event never represents reality;

b) its interpretation is unclear if the system fails before the first false alarm: ϕ_G is really a *conditional* expectation, yet it is not computed as such in the literature;

c) in any event, in most situations *any* checking action (not only a false alarm) is costly; thus ϕ_G is less relevant than, for example, the total checking (thus cost-incurring) rate $r = r_f + r_t$.

As a point of completness, we note that if the inter-checking intervals $t_1, t_2 - t, t_3 - t_2, \ldots$ etc. are I.I.D. random variables, then the unconditional expected time to first alarm (false or true), or ϕ, is

$$\phi = E_G(t_1) = E_G(t_2 - t_1) = E_G(t_3 - t_2), \text{ etc.},$$

where E_G is the expectation given the system was found in condition G when last checked. The total checking rate r is then

$$r = r_f + r_t \;\; = \;\; 1/\phi.$$

We stress that for nontrivial decision policies, the inter-checking intervals will *not* be I.I.D. For example, while using the TPR policy with any informative observations, the expected time between checks becomes smaller once the system fails.

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

The computation of $P_n(\underline{x}_n)$ follows from a straightforward application of Bayes' Theorem:

$$\begin{aligned} P_n(\underline{X}_n) &\equiv \text{prob}\{T \leq n | \underline{X}_n = \underline{x}_n\} \\ &= \frac{\text{prob}\{T \leq n \cap \underline{X}_n = \underline{x}_n\}}{\text{prob}\{\underline{X}_n = \underline{x}_n\}} \end{aligned} \tag{8}$$

where

$$\text{prob}\{T \leq n \cap \underline{X}_n = \underline{x}_n\} = \sum_{j=1}^{n} f(j)\Pi_{i=1}^{j-1} p(x_i)\Pi_{k=j}^{n} q(x_k) \tag{9}$$

and

$$\text{prob}\{\underline{X}_n = \underline{x}_n\} = \text{prob}\{T \leq n \cap \underline{X}_n = \underline{x}_n\} + \sum_{j=n+1}^{\infty} f(j)\Pi_{i=1}^{n} p(x_i) \tag{10}$$

Substituting (9) and (10) into (8), and dividing numerator and denominator by $\Pi_{i=1}^{n} p(x_i)$, gives

$$P_n(\underline{x}_n) = \frac{\sum_{j=1}^{n} f(j)\Pi_{k=j}^{n} L(x_k)}{\sum_{j=1}^{n} f(j)\Pi_{k=j}^{n} L(x_k) + \bar{F}(n)} \tag{11}$$

where

$$L(x_i) \equiv q(x_i)/p(x_i) \tag{12}$$

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

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

For notational convenience in the remainder of this paper, the argument $\underline{x}_n$ will be left out, e.g., $P_n(\underline{x}_n)$ will be written as P_n.

Although computing P_n directly from equation (11) is straightforward, it is advantageous to use the "odds in favor of condition B" ratio $R_n \equiv P_n/(1 - P_n)$, which is obtained directly from equation (11) as

$$R_n = [\bar{F}(n)]^{-1} \sum_{j=1}^{n} f(j)\Pi_{k=j}^{n} L(x_k). \tag{13}$$

This allows a recursive representation for R_n:

$$R_{n+1} = \frac{L(x_{n+1})}{\bar{F}(n+1)}[\bar{F}(n)R_n + f(n+1)], \tag{14}$$

which can be confirmed by substitution into equation (13). Equation (14) is an excellent way to compute $P_n = R_n/(1+R_n)$ since R_{n+1} is calculated from the previously obtained R_n after each new observation x_{n+1} by a simple addition and multiplication.

Equation (14) also clarifies the challenge of computing the expected time until P_n first equals or exceeds p^* or, equivalently, the first time R_n equals or exceeds the "odds threshold" $p^*/(1-p^*)$. In particular, when x_{n+1} is replaced by the r.v. X_{n+1}, we see that equation (14) 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 governed by the stochastic behavior of X_i, which in turn are governed by the p.d.f.s of equation (1).

5. Special Case: Geometric Time to Failure

The remainder of this paper assumes the operating time from G to B follows the geometric p.m.f. $f(t) = a(1-a)^{t-1}$, $t = 1, 2, 3, \ldots$, with cummulative mass function (c.m.f.)

$$F(t) = 1 - (1-a)^t, \quad t = 1, 2, 3 \ldots, \tag{15}$$

and expectation $E(T) = 1/a$. Using this distribution in equation (14), the random variables X_i produce the generating equation for the Markov Process R_n:

$$R_{n+1} = \frac{L(X_{n+1})}{1-a}[R_n + a].$$

In order to gain notational convenience and some advantage in computation, scaling and interpretation, the process R_n can be transformed into a new Markov Process $Z_n \equiv R_n/a$, $Z_n \in \mathbb{R}^+$. This new process has governing equation

$$Z_{n+1} = \ell(X_{n+1})[Z_n + 1], \tag{16}$$

where $\ell(X_n) \equiv L(X_n)/(1-a)$, and the process has an associated threshold

$$z^* = \frac{p^*}{a(1-p^*)}.$$

The absorption behavior (i.e., the distribution of time until Z_n first equals or exceeds z^*) of this process has a long and important history of study (see Shiryaev 1978), with computational (as contrasted to structural) results essentially constrained to limit theorems and approximations [e.g. Pollak (1985)]. Our emphasis is not on absorption, but on the steady-state behavior that results when the system is renewed after checking. This allows direct computation of the measures r_t and r_f from which we get δ, p_B and p_f (see Section 3).

6. No Information and Perfect Information

Before exploring how the performance measures and OCs might be calculated in general, it is instructive to examine two special situations. First, consider the limiting case where observations provide no information, a situation equivalent to $p(x_i) = q(x_i)$ or $L(x_i) = 1$, for $i = 1, 2 \ldots$. In this case, equation (14) reduces to

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

which is a deterministic difference equation with solution (given boundary condition $R_0 = 0$):

$$R_n = \frac{F(n)}{\bar{F}(n)}.$$

From the definition of R_n, this gives

$$P_n = F(n).$$

Clearly, observations of $\underline{x}_n$ have no effect on computing P_n, which is simply the cumulative distribution for the failure time T. This result holds for any $f(t)$.

To calculate the performance measures, we derive expressions for μ (expected number of false alarms per cycle), δ (expected response time) and the expected cycle time. The remaining performance measures follow immediately from their definitions.

The (deterministic) processing time to the first alarm, ϕ, is the smallest (integer) n such that $F(n) \geq p^*$, i.e., $\phi = \min_n\{n : F(n) \geq p^*\}$. Since there is a constant probability $F(\phi)$ that an alarm is a true alarm (which is the last alarm in a cycle), the expected number of false alarms per cycle is

$$\begin{aligned} \mu &= 0F(\phi) + 1(1 - F(\phi))F(\phi) + 2(1 - F(\phi))^2F(\phi) + \cdots \\ &= \frac{1 - F(\phi)}{F(\phi)} \end{aligned}$$

Computation of δ is slightly more complex. If $T \leq \phi$, then the time between T and when action is taken — the response time D — is

$$D = \phi - T.$$

If $k\phi < T \leq (k+1)\phi$ for $k = 1, 2, \ldots$, then there are k false alarms (at system times ϕ, $2\phi + c$, $3\phi + 2c$, ..., and $k\phi + (k-1)c$), followed by a "true detection" with response time

$$D = (k+1)\phi - T.$$

If the system were not renewed, then

$$\delta = \sum_{k=0}^{\infty} \sum_{t=k\phi+1}^{(k+1)\phi} [(k+1)\phi - t] f(t).$$

However, since $f(t)$ is geometric, the process renews after a false alarm, and

$$\delta = \frac{\sum_{t=1}^{\phi} (\phi - t) f(t)}{F(\phi)}. \tag{17}$$

From equation (15) for $F(t)$,

$$\phi = \min_n \{n : n \geq \ln(1-p^*)/\ln(1-a)\}.$$

Using this in equation (17), after some algebra, results in

$$\delta = \frac{(1-a)^\phi - 1 + a\phi}{a[1-(1-a)^\phi]}.$$

Given δ, μ, and $E(T) = 1/a$, the expected cycle time is

$$C(b,c) = 1/a + c\mu + \delta + b$$

which allows calculation of the remaining performance measures from their definitions:

$$\begin{aligned} r_f &= \mu/C(b,c), \\ r_t &= 1/C(b,c), \\ p_f &= \mu c/C(b,c), \\ p_B &= (\delta + b)/C(b,c). \end{aligned}$$

Figure 4 shows an example operating characteristic curve: plots of r versus δ as p^* varies, for $b = c = 1$ and selected values of a. Figure 5 contains the same information presented in "normalized" time units (i.e., in units of $E(T) = 1/a$) where $r' \equiv r/a$ is the normalized checking rate. Since these represent worst case results, any actual information gathering procedure should produce achievable points below the curves in Figures 4 and 5.

The limiting case of "perfect" information represents another special situation. In this case the supports of $p(x)$ and $q(x)$ are disjoint, so that $L(x_i) = 0$ for $i = 1, 2, \ldots T-1$, and $L(x_i) = \infty$ for $i = T, T+1 \ldots$. From equation (11), we see that *any* non-zero threshold is exceeded for all $n \geq T$, so that $\delta = 0$. Since the system is checked only when it fails, $\mu = 0$ and the true alarm rate is $r_t = 1/(E(T)+b) = a/(1+ab)$, the false alarm rate is $r_f = 0$, the fraction of time down is $p_B = b/(E(T)+b) = ab/(1+ab)$ and the fraction of time processing false alarms is $p_f = 0$.

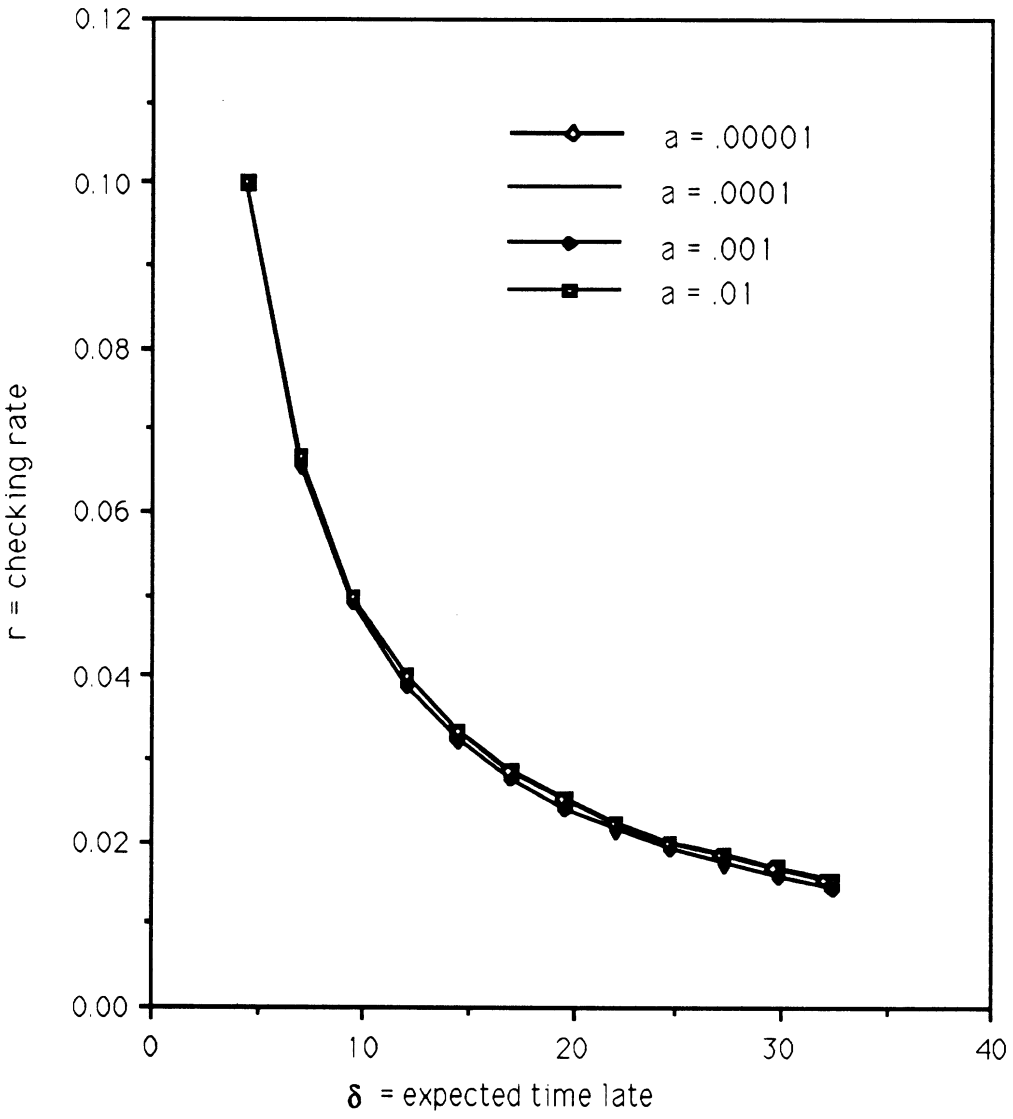

Figure 4: Operating Characteristic for "no information" case.

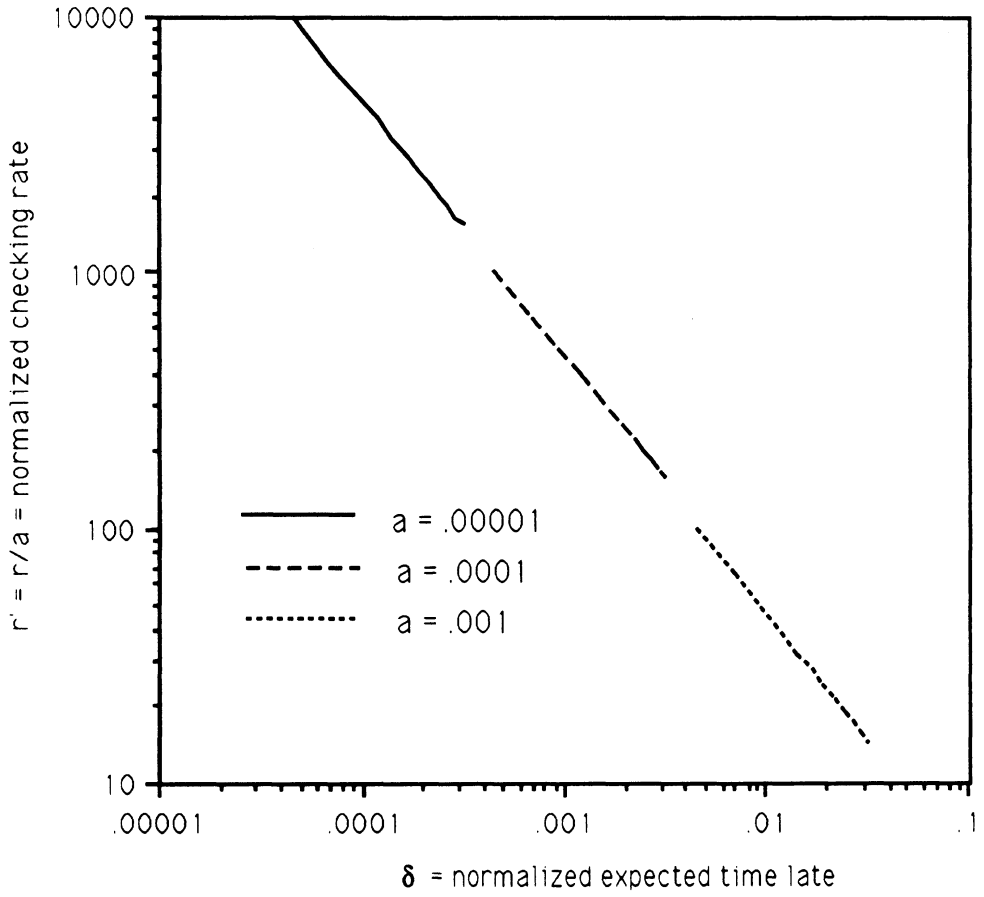

Figure 5: Normalized Operating Characteristic for "no information" case.

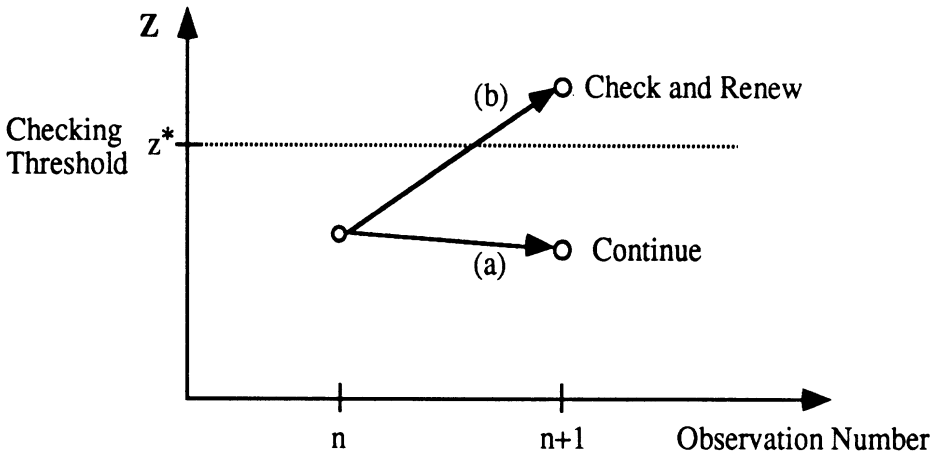

Figure 6: Evolution of the Z_n process for an observation that (b) leads to checking, and (a) does not lead to checking.

7. An Extension of the Process Z_n

When information (in the form of $\underline{x}_n$) is gathered, the process Z_n described by equation (16), along with the system condition, determines the behavior of the monitoring procedure. Figure 6 shows two possible results of computing Z_{n+1} from Z_n: an observation of x_{n+1} is made, and either

a) Z_{n+1} is less than the threshold z^*, and the process continues; or

b) Z_{n+1} equals or exceeds z^* and action (check and renew) is taken.

Recall that $z^* \equiv p^*/a(1-p^*)$ is the threshold value for Z_n, and that $Z_n \geq z^*$ means a decision is made to check; when this action is taken, the system is renewed (either because condition B is discovered, or because of the memoryless property of the geometric distribution if condition G is discovered).

The conditions for (a) and (b), respectively, from equation (16) are such that

a) $\ell(x_{n+1}) < z^*/(Z_n + 1)$, and

b) $\ell(x_{n+1}) \geq z^*/(Z_n + 1)$.

Recall that x_i are realizations of random variables X_i which are generated with p.d.f. $p(x)$ while the system is in G, and with p.d.f. $q(x)$ while the system is in B.

In order to directly compute the various performance measures, we characterize the checking/monitoring procedure by means of a mixed continuous-discrete state Markov Process (called, for convenience, "MPZ") based upon combining the process Z_n with the system condition C_n. The performance measures can be obtained from the appropriate stationary state probabilities of this MPZ.

The state of MPZ at the end of the n^{th} transition will be denoted as $S_n \in \mathcal{S}$, $n = 1, 2, \ldots$. The state space $\mathcal{S}$ is the union of five sub-spaces: three singletons and two mixed continuous-discrete. These sub-spaces are:

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

$\mathcal{S}_G^* \equiv \{(z^*, G)\}$, false alarm state: the state entered after checking while the system is in G and occupied until system renewal, i.e., when $Z_n \geq z^*$ and $C_n \equiv G$;

$\mathcal{S}_F^* \equiv \{(z^*, B)\}$, true alarm state: the state entered after checking while the system is in B and occupied until system renewal, i.e., when $Z_n \geq z^*$ and $C_n = B$;

$\mathcal{S}_G \equiv \{(z, G) : z \in (0, z^*)\}$, set of states where Z_n is between 0 and z^*, and $C_n = G$;

$\mathcal{S}_B \equiv \{(z, B) : z \in (0, z^*)\}$, set of states where Z_n is between 0 and z^*, and $C_n = B$.

The transition probabilities among the states in these sets are governed by the evolution of Z_n described by equation (16), the behavior of the random variable X_{n+1} given in equation (1), and the geometric failure time distribution of equation (15) which implies a probability a of the system condition going from G to B at each transition (except from $\mathcal{S}_G^*$ where this probability is zero).

In the following development we assume $b = c = 1$, since computing results for arbitrary b and c values (even for zero values) is straight forward — see Appendix C. The steady state equations for MPZ are derived in Appendix A. We note some of the properties of this Markov Process:

a) it is 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 0 is 1. Both represent one transition (since we assume $b = c = 1$) needed to check and renew the system (the former after a false alarm, the latter after a true alarm);

c) due to the geometric failure time distribution of equation (15), the single-step transition probability:

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

   ii) from the state 0 to the set $\mathcal{S}_B \cup \mathcal{S}_B^*$.

   is a.

The steady-state probabilities for the singleton states are defined as

$$\pi_0 \equiv \lim_{n\to\infty} \text{prob}\{S_n = 0\},$$
$$\pi_G^* \equiv \lim_{n\to\infty} \text{prob}\{S_n = \mathcal{S}_G^*\},$$
$$\pi_B^* \equiv \lim_{n\to\infty} \text{prob}\{S_n = \mathcal{S}_B^*\}.$$

In addition, we define the steady state cumulative distribution functions for the sets $\mathcal{S}_G$ and $\mathcal{S}_B$ as

$$\begin{aligned} \Pi_G(z) &= \lim_{n\to\infty} \text{prob}\{S_n \in (t,G) : t \le z\}, \quad 0 < z < z^*, \\ \Pi_B(z) &= \lim_{n\to\infty} \text{prob}\{S_n \in (t,B) : t \le z\}, \quad 0 < z < z^*, \end{aligned}$$

and set, for convience, $\Pi_G(0) = \Pi_B(0) = 0$, $\Pi_G(z^*) = \lim_{\epsilon\to 0} \Pi_G(z^* - \epsilon)$, and $\Pi_B(z^*) = \lim_{\epsilon\to 0} \Pi_B(z^* - \epsilon)$.

If these steady state probabilities and distributions can be computed from the equations in Appendix A, then the performance measures are easily obtained. In particular, by appealing to the ergodic theorem for Markov Processes, we know that

π_0 = expected fraction of time the process is in the renewed state.

Thus, $r = r_t + r_f = \pi_0$, since MPZ is in 0 for exactly one time unit per cycle.

Similiarly, since

π_G^* = expected fraction of time the process is in the false alarm state,

π_B^* = expected fraction of time the process is in the true alarm state,

we know that $r_f = \pi_G^*$ and $r_t = \pi_B^*$ (recall that c or b, the time to renew from either G or B, is one time unit). Given these values for r_f and r_t, we have from the analysis in Section 3 (with $b = c = 1$ and $E(T) = 1/a$),

$$\mu = \frac{\pi_G^*}{\pi_B^*}, \tag{18}$$

$$\delta = \frac{1 - \pi_G^*}{\pi_B^*} - \frac{1}{a} - 1, \tag{19}$$

$$p_B = 1 - \frac{\pi_B^*}{a} - \pi_G^*, \tag{20}$$

$$p_f = \pi_G^*. \tag{21}$$

Thus we can present OC plots (such as r_f v.s. p_B), given the steady-state probabilities π_B^* and π_G^*.

8. Markov Chain Approximation

The steady-state equations used to calculate π_B^* and π_G^* – (A4) and (A5) 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 the ARL measure ϕ_G (rather than the measures like r_f and p_B that we seek). However, as Pollak notes, a solution method is still lacking for even the simplest forms of $p(\cdot)$ and $q(\cdot)$. The computational literature is problem-specific and essentially suggests using variable transformations and discretization approximations tailored to the problem at hand. Following this approach, we now develop an approximation for some specific cases of $p(x)$ and $q(x)$. In the context of our problem this approximation method is equivalent to proposing that the process MPZ can be approximated by a Markov *Chain* ("MCZ") with *finite* state space.

To construct MCZ, we require values of Z_n that do not cover the interval $[0, z^*]$, but in fact are restricted to a finite set

$$0 \cup \mathcal{S}^1 \cup z^*$$

where

$$\mathcal{S}^1 \equiv \{z_1, z_2, \ldots z_{m-1}\}.$$

The key to this restriction is to find values of z_i that are numerous enough, and "cover" the interval $(0, z^*)$ in such a way that the sums over probabilities in $\mathcal{S}^1$ well approximate the integrals over $\mathcal{S}$ implicit in equations (A2) to (A5). Obtaining such a set of z-values is discussed below; at this point we will assume that $\mathcal{S}^1$ is available.

Given the finite elements of $\mathcal{S}^1$, we define a finite state space $\mathcal{Z}$ for MCZ

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

where the first three sub-spaces are the singletons (corresponding to the renewed, false alarm

and true alarm "states" of the system, respectively) and

$$\mathcal{Z}_G \equiv \{z_i : z_i \in \mathcal{S}^1,\ C_n = G,\ i = 1,2,\ldots m-1,\ n = 1,2,\ldots\}$$
$$\mathcal{Z}_B \equiv \{z_i : z_i \in \mathcal{S}^1,\ C_n = B,\ i = 1,2,\ldots m-1,\ n = 1,2,\ldots\}.$$

Thus $\mathcal{Z}_G$ represents a subset of Z_n values while the system is in G, and $\mathcal{Z}_B$ represents a subset of Z_n values when the system is in B.

A simple re-numbering of states now allows us to represent MCZ as a $(2m+1)$-state ergodic Markov Chain, which we will refer to as "MC," with state space $I_{2m+1} \equiv \{0,1,2,\ldots, 2m\}$, $\sigma_n \equiv$ the state of MC after the n^{th} transition, and transition matrix P with elements $[P]_{ij} = p_{ij} \equiv \text{prob}\{\sigma_n = j | \sigma_{n-1} = i\}$ for $i,j \in I_{2m+1}$, and $n = 1,2,\ldots$. Details of this representation of MCZ and MC are contained in Appendix B.

Given the elements of P developed in Appendix B, the steady-state probability vector $\pi \equiv \{\pi_0, \pi_1, \pi_2, \ldots, \pi_{2m}\}$, where

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

can be obtained by solving the set of linear equations:

$$\pi = \pi P,$$
$$\pi = \pi \underline{1},$$

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

The key performance measures of interest can be immediately obtained from these equations since $\pi^*_G = \pi_m$ and $\pi^*_B = \pi_{2m}$.

9. Special case: Bernoulli Observations

The general conditions under which the "discretization" of MPZ to MCZ is valid, and which allow an explicit determination of π_G^* and π_B^* from $p(x)$ and $q(x)$, are not addressed in this paper. We consider, instead, the following special case of Bernoulli observations, where $X_n = 0$ or $1, n = 1, 2, \ldots$, 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 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

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

then

$$\ell(x) = \begin{cases} w_0 & \text{if } x = 0, \\ w_1 & \text{if } x = 1, \end{cases}$$

so that equation (16) can be written

$$Z_{n+1} = \begin{cases} w_0(Z_n + 1) & \text{if } X_{n+1} = 0, \\ w_1(Z_n + 1) & \text{if } X_{n+1} = 1. \end{cases} \tag{22}$$

Before writing the steady-state equations for this case, it is convenient to note that, for a realistic problem,

a) the value of a (the probability that the system goes from G to B on any transition) is usually quite small;

b) α and β, the "misclassification" errors in a single observation of x, are relatively small compared to one (generally, $\alpha, \beta \leq .1$);

Under these conditions

$$0 < w_0 < 1 < w_1 < z^*,$$

and equations (A2) to (A6) can be written, after some reduction and algebra:

$$\Pi_G(z) = \begin{cases} 0 & \text{if } 0 \leq z < w_0, \\ (1-a)(1-\alpha)\left[\Pi_G(\frac{z}{w_0} - 1) + \pi_0\right] & \text{if } w_0 \leq z \leq w_1, \\ (1-a)\left[(1-\alpha)\Pi_G\left(\frac{z}{w_0} - 1\right) + \alpha\Pi_G\left(\frac{z}{w_1} - 1\right) + \pi_0\right] & \text{if } w_1 < z \leq z^*, \end{cases} \tag{23}$$

$$\Pi_B(z) = \begin{cases} 0 & \text{if } 0 \leq z < w_0, \\ \beta\left[a\Pi_G(\frac{z}{w_0} - 1) + \Pi_B(\frac{z}{w_0} - 1) + a\pi_0\right] & \text{if } w_0 \leq z \leq w_1, \\ \beta\left[a\Pi_G\left(\frac{z}{w_0} - 1\right) + \Pi_B\left(\frac{z}{w_0} - 1\right)\right] & \\ \quad +(1-\beta)\left[a\Pi_G\left(\frac{z}{w_1} - 1\right) + \Pi_B\left(\frac{z}{w_1} - 1\right)\right] + a\pi_0 & \text{if } w_1 < z \leq z^*, \end{cases} \tag{24}$$

$$\pi_G^* = (1-a)\alpha\left[\Pi_G(z^*) - \Pi_G\left(\frac{z^*}{w_1} - 1\right)\right], \tag{25}$$

$$\pi_B^* = a\beta\left[\Pi_G(z^*) - \Pi_G\left(\frac{z^*}{w_1} - 1\right)\right] + (1-\beta)\left[\Pi_B(z^*) - \Pi_B\left(\frac{z^*}{w_1} - 1\right)\right], \tag{26}$$

$$\pi_0 = \pi_G^* + \pi_B^*, \tag{27}$$

where $\Pi_G(0) = \Pi_B(0) = 0$, $\Pi_G(z) = \Pi_G(z^*)$ for $z > z^*$, and $\Pi_B(z) = \Pi_B(z^*)$ for $z > z^*$.

10. Markov Process Discretization

In order to create the MC version of equations (23) to (27), we now consider the evolution of the Z_n process of equation (22), when the process starts with $Z_n = 0$ (i.e., $R_0 = P_0 = 0$). The possible values of Z that can be generated after the first three observations, assuming none exceed z^*, are given in Table 1:

Observation Number n	Possible Z_n Values	Label
1	w_0	1
	w_1	2
2	$(1+w_0)w_0$	3
	$(1+w_0)w_1$	4
	$(1+w_1)w_0$	5
	$(1+w_1)w_1$	6
3	$(1+(1+w_0)w_0)w_0$	7
	$(1+(1+w_0)w_0)w_1$	8
	$(1+(1+w_0)w_1)w_0$	9
	$(1+(1+w_0)w_1)w_1$	10
	$(1+(1+w_1)w_0)w_0$	11
	$(1+(1+w_1)w_0)w_1$	12
	$(1+(1+w_1)w_1)w_0$	13
	$(1+(1+w_1)w_1)w_1$	14

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

Figure 7 shows all possible values of Z_n for $a = 0.01, \alpha = 0.2, \beta = 0.1$, and $n = 6$. Each distinct value of Z is assigned an arbitrary "label" number. After n observations

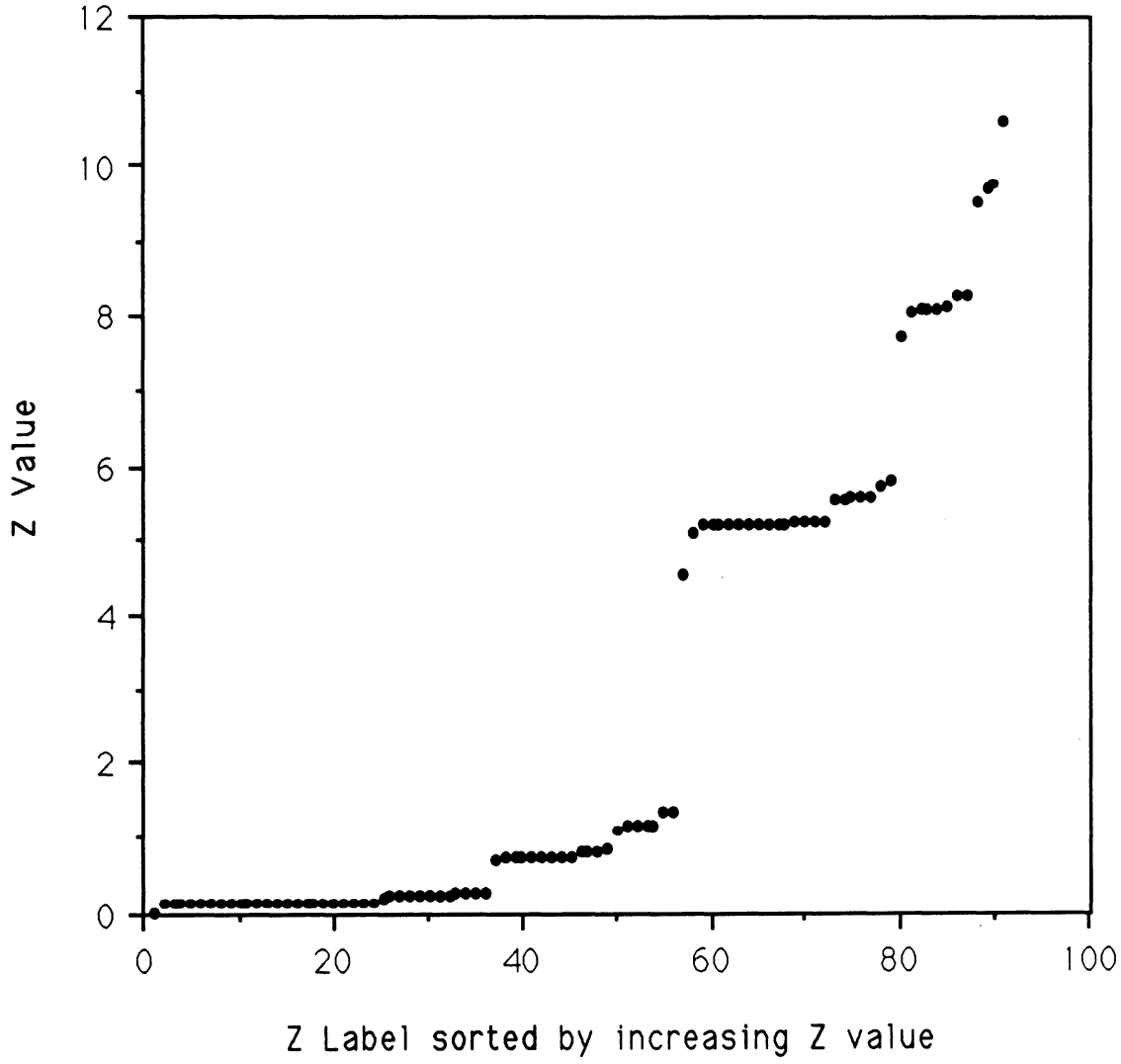

Figure 7: Possible Z_n values after 6 observations for $a = 0.01$, $\alpha = 0.2$, $\beta = 0.1$, and $p^* = .1$ (or $z^* = 11.11$). Since calculated Z values exceeding z^* are set to zero and not used subsequently, only 90 out of the 254 possible Z_n realizations after 6 observations are greater than zero.

the number of possible distinct values of Z, hence the number of different labels, is clearly $2^{n+1} - 2$. However, two important effects allow us to reduce this number, which in turn allows a practical means of discretizing MCZ to MC:

1. As n becomes large, some values of Z_n satisfy $Z_n \geq z^*$, in which case either state $\mathcal{S}_G^*$ or $\mathcal{S}_B^*$ occurs and Z_{n+1} becomes 0;

2. As n becomes large, values of Z_n less than z^* tend to "cluster" around a finite set W of $m-1$ points ($m-1 = |W|$).

Thus, we can approximately incorporate into a *single* discrete value z_i all those Z_n that are within some small range $\epsilon > 0$ of z_i. Then the entire evolution of the process can be represented by a MC with states 0, $\mathcal{S}_G^*$, $\mathcal{S}_B^*$ plus the $2m-2$ states from the cross product of the set W and the set of system conditions: $\{G, B\}$. The associated probability transition matrix P can be easily created (see Appendix B) from the two probability transition submatrices P^G and P^B corresponding to state transitions placing the system in G and B, respectively. Figure 8 depicts the composition of the transition matrix P using these submatrices:

$$\left[P^G\right]_{ij} = \begin{cases} 1-\alpha & \text{if } j = J_0(i),\ i \in \{0,1,2,\ldots,m-1\}, \\ \alpha & \text{if } j = J_1(i),\ i \in \{0,1,2,\ldots,m-1\}, \\ 0 & \text{other } i \in \{0,1,\ldots,m-1\},\ j \in \{0,1,\ldots,m\}, \end{cases} \tag{28}$$

$$\left[P^B\right]_{ij} = \begin{cases} \beta & \text{if } j = J_0(i)+1,\ i \in \{0,1,\ldots,m-1\}, \\ 1-\beta & \text{if } j = J_1(i)+1,\ i \in \{0,1,\ldots,m-1\}, \\ 0 & \text{other } i \in \{0,1,\ldots,m-1\},\ j \in \{1,2,\ldots,m\} \end{cases} \tag{29}$$

$$\tag{30}$$

where, after letting $W_0^* = \{0 \cup W \cup z^*\}$ and defining 0 as its 0^{th} element (i.e., $z_0 = 0$) and

z^* as its m^{th} element (i.e., $z_m = z^*$), respectively,

$$\begin{aligned} J_0(i) &= \text{the index of the closest element in } W_0^* \text{ to } Z_{n+1} \text{ given } Z_n = z_i \text{ and } x_{n+1} = 0 \\ &= \arg \min_{k \in W_0^*} \{|w_0(z_i + 1) - z_k|\} \end{aligned}$$

and

$$\begin{aligned} J_1(i) &= \text{the index of the closest element in } W_0^* \text{ to } Z_{n+1} \text{ given } Z_n = z_i \text{ and } x_{n+1} = 1 \\ &= \arg \min_{k \in W_0^*} \{|w_1(z_i + 1) - z_k|\}. \end{aligned}$$

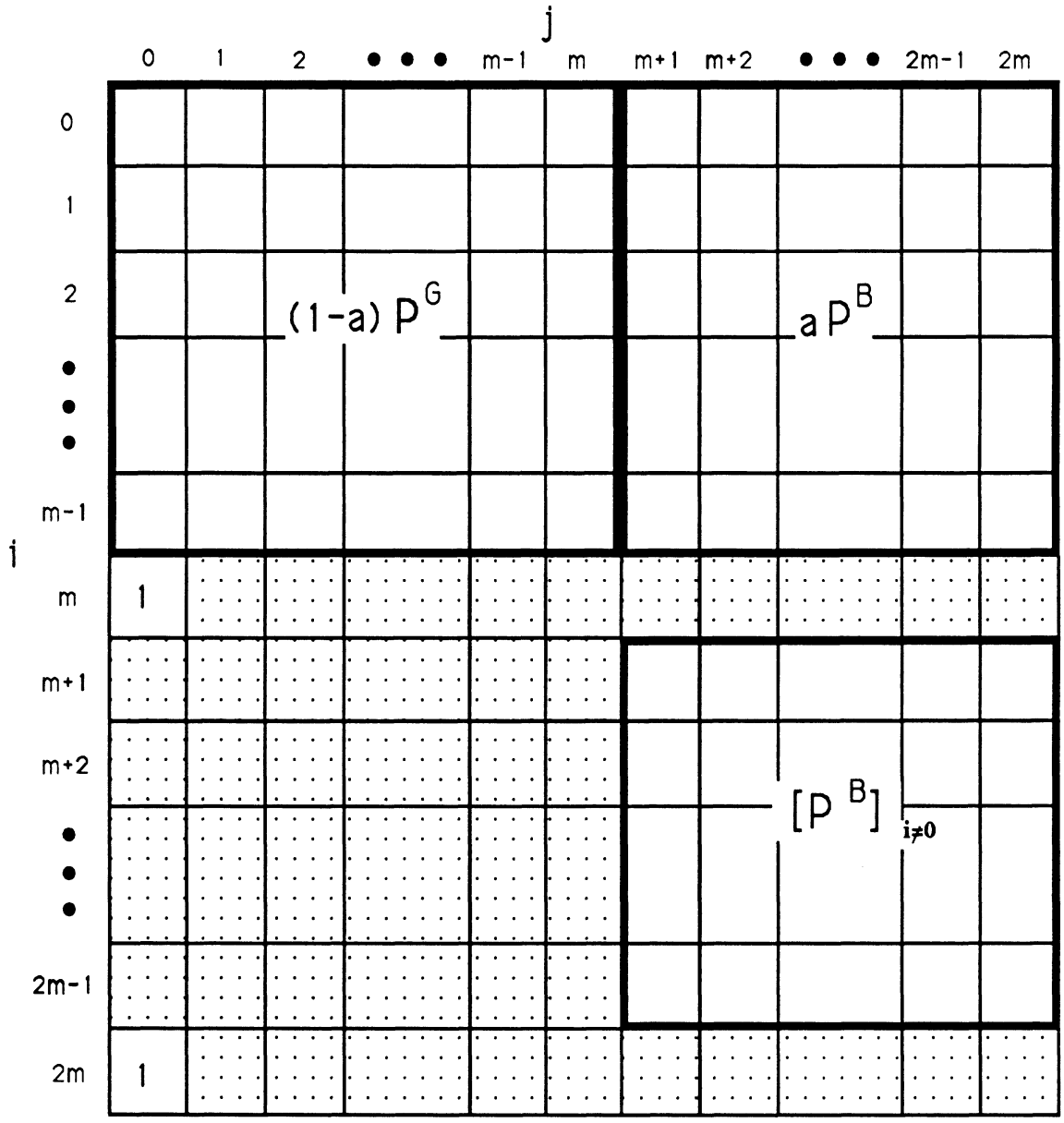

Figure 8: Schematic representation of the transition matrix P from state i to state j showing the use of submatrices P^G and P^B. $[M]_{i\neq 0}$ denotes the submatrix created by removing the row associated with $i = 0$ from M. Shaded regions represent zero transition probabilities.

11. Conclusion

The value of the Operating Characteristic as a tool for evaluation of monitoring procedures and policies has yet to be established. We believe, however, that it is an important and evocative tool for the comparison of policies and the comparison of alternative observation technologies. The structure set forth in this report provides a method for the computation of critical measures such as the expected detection time and total alarm rate (δ and r), needed to express various OCs. It is an open question, however, as to whether the approach outlined in section 10 will prove viable for computation in the case when there are Bernoulli observations. For the case of general sampling functions $p(x)$ and $q(x)$, Appendix A presents the equations that need to be solved to find key performance measures. Again, demonstrated numerical techniques for their solution are not yet available.

In a companion paper we explore the effectiveness of the discretization method outlined in section 10 as a means of obtaining numerical representation of useful OC's.

Acknowledgments

This research was supproted 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.

APPENDIX A: Steady State Properties of MPZ

In this Appendix, we show how the Chapman-Kolmogorov (C-K) equations can be used to write expressions for the steady-state probabilities and distributions for MPZ. Although the notation and details may seem formidable, the method is a straightforward extension of the use of C-K equations for finding the steady-state solutions to a finite state Markov chain. In the development that follows, it might be helpful to refer to the suggestive flow diagram of Figure A1. In this diagram, the "states" $\{Z_G\}$ and $\{Z_B\}$ refer to Z-values in the open interval $(0, z^*)$ while the system 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 7. The labels on the transition arrows represent governing probabilities (to singleton states) and densities (to $\{Z_G\}$ and $\{Z_B\}$), where $\bar{P}$ and $\bar{Q}$ are the complementary distribution functions associated with p and q.

Consider the computation of $\Pi_G(z)$. By definition

$$\begin{aligned}\Pi_G(z) &= \lim_{n\to\infty} \text{prob}\{S_n \in \{(t,G)\} : 0 < t \leq z\} \\ &= \lim_{n\to\infty} \text{prob}\{0 < Z_n \leq z \cap C_n = G\},\end{aligned}$$

which by conditioning on the value of Z_{n-1}, and noting that $C_n = G$ is only possible if $C_{n-1} = G$, gives

$$\Pi_G(z) = \int_{y=0}^{z^*} \lim_{n\to\infty} \text{prob}\{0 < Z_n \leq z \cap C_n = G | Z_{n-1} = y \cap C_{n-1} = G\}$$
$$\times \text{prob}\{Z_{n-1} \in (y, y+dy) \cap C_{n-1} = G\}. \qquad (A1)$$

The first probability in (A1) can be obtained by using:

a) equation (16) which governs the behavior of Z_n when $Z_{n-1} = y$, and

b) $\text{prob}\{C_n = G | C_{n-1} = B\} = 1 - a$ which is independent of the value of Z_{n-1}.

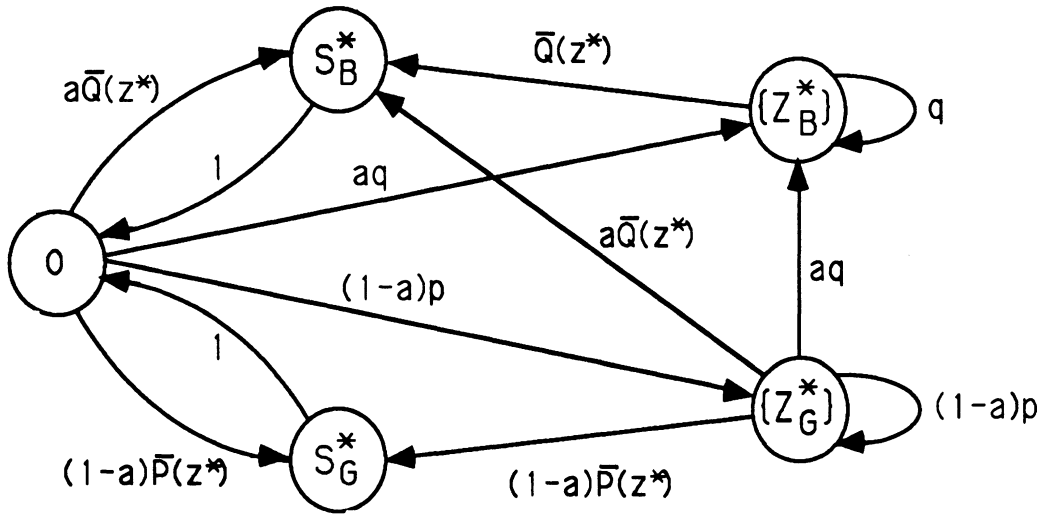

Figure A1: Schematic representation of transitions among the states in MPZ. Note that 0, S_B^* and S_G^* are singleton states, while $\{Z_G\}$ and $\{Z_B\}$ represent a continuum of states in the open interval $(0, z^*)$.

Thus

$$\Pi_G(z) = \int_{y=0}^{z^*} \lim_{n\to\infty}(1-a)\ \text{prob}\{0 < \ell(X_{n-1})(y+1) \le z | C_{n-1} = G\}$$
$$\times \text{prob}\{Z_{n-1} \in (y, y+dy) \cap C_{n-1} = G\}.$$

By definition

$$\lim_{n\to\infty}\ \text{prob}\{Z_{n-1} \in (y, y+dy) \cap C_{n-1} = G\} = d\Pi_G(y), \qquad 0 < y \le z^*.$$

Hence, after taking the limit as $n \to \infty$,

$$\begin{aligned}\Pi_G(z) \;=\; & \int_{y=0}^{z^*}(1-a)\ \text{prob}\{0 < \ell(X_{n-1}(y+1) \le z | C_{n-1} = G\} d\Pi_G(y) \\ & +(1-a)\pi_0\ \text{prob}\{0 < \ell(X_{n-1}) \le z | C_{n-1} = G\}.\end{aligned}$$

For convenience, we define the region $\mathcal{C}(z,t) \equiv \{x :\ \ell(x) < z/(1+t)\}$. Using this notation, $\mathcal{C}(z^*, x_n)$ is the set of "continuation" values of the observation x_n.

Finally, since the p.d.f. for X_{n-1}, given $C_{n-1} = G$, is $p(x)$,

$$\Pi_G(z) = (1-a)\int_{y=0}^{z^*}\int_{x\in\mathcal{C}(z,y)} p(x)\,dx\,d\Pi_G(y) + (1-a)\pi_0\int_{x\in\mathcal{C}(z,0)} p(x)\,dx. \qquad (A2)$$

Similarly, it can be shown that

$$\begin{aligned}\Pi_B(z) \;=\; & a\int_{y=0}^{z^*}\int_{x\in\mathcal{C}(z,y)} q(x)\,dx\,d\Pi_G(y) + a\pi_0\int_{x\in\mathcal{C}(z,0)} q(x)\,dx \\ & +\int_{y=0}^{z^*}\int_{x\notin\mathcal{C}(z^*,y)} q(x)dx d\Pi_B(y) \qquad (A3)\end{aligned}$$

$$\pi_G^* \;=\; \int_{y=0}^{z^*}\int_{x\notin\mathcal{C}(z^*,0)}(1-a)p(x)\,dx\,d\Pi_G(y) + \pi_0\int_{x\notin\mathcal{C}(z^*,0)}(1-a)p(x)\,dx. \qquad (A4)$$

$$\begin{aligned}\pi_B^* \;=\; & \int_{y=0}^{z^*}\int_{x\notin\mathcal{C}(z^*,0)} aq(x)\,dx\,d\Pi_G(y) + \pi_0\int_{x\notin\mathcal{C}(z^*,0)} aq(x)\,dx. \\ & +\int_{y=0}^{z^*}\int_{x\notin\mathcal{C}(z^*,y)} q(x)\,dx\,d\Pi_B(y) \qquad (A5)\end{aligned}$$

$$
\begin{aligned}
\pi_0 &= \pi_G^* + \pi_B^* \\
&= \int_{y=0}^{z^*} \int_{x \notin C(z^*,0)} [aq(x) + (1-a)p(x)]\, dx\, d\Pi_G(y) \\
&\quad + \int_{y=0}^{z^*} \int_{x \notin C(z^*,y)} q(x)\, dx\, d\Pi_B(y) + \pi_0 \int_{x \notin C(z^*,0)} [aq(x) + (1-a)p(x)]\, dx. \qquad \text{(A6)}
\end{aligned}
$$

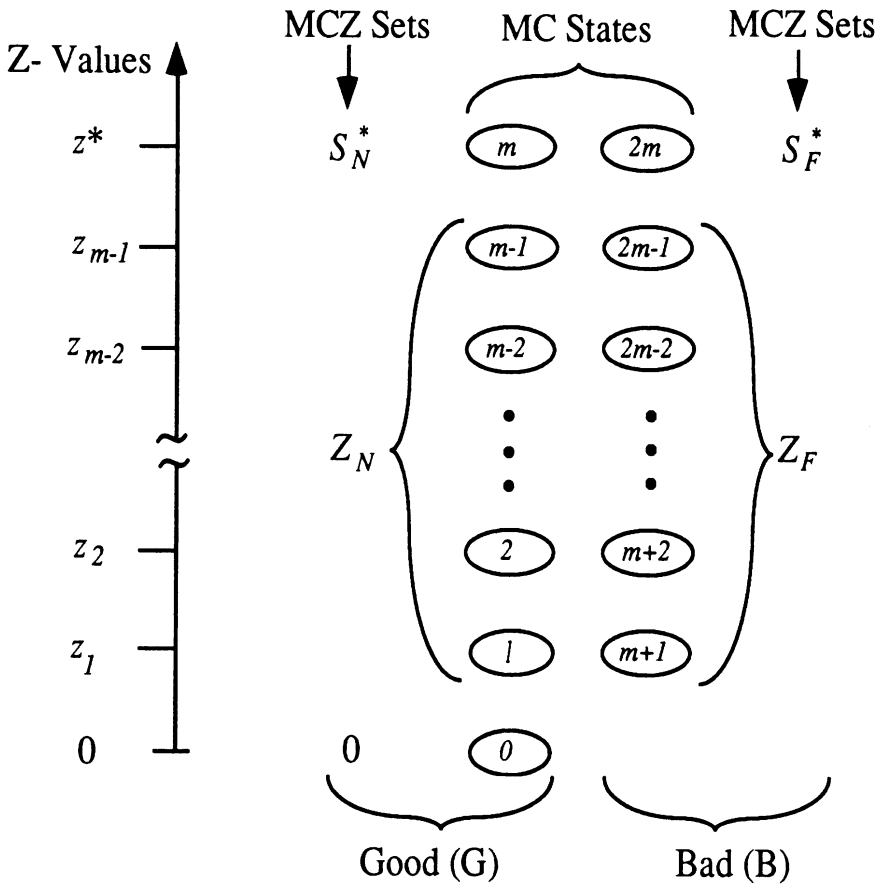

Figure B1: Schematic representation of the MC states, the MCZ sets, and the Z-values generated from equation (16). System condition transitions are governed by a geometric probability distribution from G to B.

APPENDIX B: Correspondance between Formulations MC and MCZ

The correspondance between elements of $I_{2m+1} \in \{0, 1, 2, \ldots, 2m\}$ of the chain MC and of the elements MCZ 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 $R_n = 0$ and $Z_n = 0$) it is also the "renewal" state, with system condition G; state m is the false alarm state, since $Z_n = z^*$ and the system condition is G; state $2m$ is the true alarm state, since $Z_n = z^*$ and the system condition is B. The set $\mathcal{Z}_G$ represent states

where Z_n lies between 0 and z^* when the system condition is G; $\mathcal{Z}_B$ represents states where Z_n lies between 0 and z^* and the system condition is B.

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

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

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

c) for any value of $Z_n < z^*$, if the system condition is G then with probability a condition B applies on the next transition;

d) the only transition 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

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

The elements of these submatrices are given by the nature of the distributions $p(\cdot)$ and $q(\cdot)$ of equation (1). In terms of these submatrices, the remainding elements of P are given by:

$$p_{ij} = (1-a)[P^G]_{ij} \qquad \text{for } i = 0,1,\ldots,m-1;\ j = 0,1,\ldots,m$$

$$p_{ij} = a[P^B]_{i,j-m} \qquad \text{for } i = 0,1,\ldots,m-1;\ j = m+1,\ldots,2m,$$

$$p_{ij} = [P^B]_{i-m,j-m} \qquad \text{for } i = m+1,m+2,\ldots,2m-1;\ j = m+1,m+2,\ldots,2m.$$

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 system is in B. Matrix P is shown schematically in Figure 8.

APPENDIX C: State Probabilities for Arbitrary b and c

In our development, we assumed renewal times of one time unit, i.e., $b = c = 1$. Calculating the steady-state probabilities for arbitrary b and c is described here.

Recall that $C(b,c)$ is the expected cycle time (between two consecutive renewals from B) given b and c are the renewal times after a checking action detects condition B or G, respectively. Let $\pi^*_G(b,c)$ be the associated steady-state probability the system is in the false alarm state given b and c. It is clear that $\pi^*_G(1,1)C(1,1)$ is the expected number of false alarms per cycle when $b = c = 1$. For arbitrary b and c, the cycle time is $C(1,1)$ plus $b-1$ (for the renewal while in B) plus $\pi^*_G(1,1)C(1,1)(c-1)$ (for the false alarm renewals). Hence, the expected cycle time for arbitrary b and c is

$$C(b,c) = C(1,1) + (b-1) + \pi^*_G(1,1)C(1,1)(c-1).$$

Note, that $\text{prob}(S|b = c = 1)C(1,1)$ is the expected time per cycle the system is in state S for any state $S \in \mathcal{S}$ and this time is independent of b and c for all states except the two renewal states, i.e., the false alarm and true alarm renewal states. The probability of any non-renewal state $S \in \mathcal{S}$ is then

$$\text{prob}(S) = \frac{\text{prob}\{S|b = c = 1\}C(1,1)}{C(b,c)}.$$

Using these results, the probability of the false alarm renewal state is

$$\pi^*_G(b,c) = \frac{\pi^*_G(1,1)C(1,1)c}{C(b,c)},$$

and the probabilty of the true alarm renewal state is

$$\pi^*_B(b,c) = \frac{b}{C(b,c)}.$$

References

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).

Girshik, 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).

Lorden, G., "Procedures for Reacting to a Change in Distribution," *Ann. Math. Stat.*, 1897-1908, (1971).

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).

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

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).

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

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).