Accelerated Testing of On-Board Diagnostics

Spencer Graves[1], Søren Bisgaard[2], Murat Kulahci[3], John Van Gilder[4], John James[5], Ken Marko[5], Hal Zatorski[6], Tom Ting[7], Cuiping Wu[8]

[1] *PDF Solutions, Inc., San Jose, CA 95110*

[2] *University of Massachusetts, Amherst, MA 01003*

[3] *Arizona State University, Tempe, AZ 85287*

[4] *General Motors Proving Ground, Milford, MI 48380*

[5] *Ford Research Labs, Dearborn, MI 48121*

[6] *DaimlerChrysler, Auburn Hills, MI 48326*

[7] *General Motors Research, Development and Planning, Warren, MI 48090*

[8] *DaimlerChrysler Proving Grounds, Chelsea, MI 48118*

Modern products frequently feature monitors designed to detect actual or impending malfunctions. False alarms (Type I errors) or excessive delays in detecting real malfunctions (Type II errors) can seriously reduce monitor utility. Sound engineering practice includes physical evaluation of error rates. Type II error rates are relatively easy to evaluate empirically. However, adequate evaluation of a low Type I error rate is difficult without using accelerated testing concepts, inducing false alarms using artificially low thresholds and then selecting production thresholds by appropriate extrapolation as outlined here. This acceleration methodology allows for informed determination of detection thresholds and confidence in monitor performance with substantial reductions over current alternatives in time and cost required for monitor development.

KEY WORDS: False alarms; Delay to detection; Detecting malfunctions in dynamic systems; Monitoring; Run length modeling; Emission controls.

Correspondence to: Murat Kulahci, Arizona State University, Dept. of Industrial Engineering, P. O. Box 875906, Tempe, AZ 85287
E-mail: Murat.Kulahci@asu.edu

1. INTRODUCTION

Many products today include on-board diagnostics (OBDs) designed to detect actual or impending malfunctions and to alert the user when maintenance is required; see Box et al. [1] and CARB [2]. Typically, monitoring algorithms consist of real time processing of a stream of incoming data, for example with a cumulative sum (Cusum) or an exponentially weighted moving average (EWMA), and setting an alarm when an appropriate statistic exceeds a threshold, h. A key issue in the design of monitors is, however, that diagnostics may fail to detect a real malfunction in a timely fashion, a Type II error, or report a malfunction when none exists, a false alarm or Type I error; see Box et al. [3].

Error rates can often be evaluated from theory. However, the assumptions required for the theory will always be violated to some degree. Sound engineering practice therefore calls for additional direct physical verification (*a*) to confirm that violations of assumptions are negligible and (*b*) in many cases to refine estimates of theoretical model parameters in hopes of compensating adequately for violations of assumptions. For Type II errors this is relatively simple: A malfunction is artificially produced and the monitor's response observed. Unless the monitor responds quickly and consistently, it must be modified. Thus empirical evaluation of Type II errors can be done relatively inexpensively and will not be discussed in this article.

Empirical evaluation of the Type I error rate is more challenging. Ideally multiple copies of the plant would be tested for the entire design life. (In this article, the system monitored will be called the "plant", for consistency with the control theory literature.) Such testing would far exceed any reasonable budget. Fortunately, in many cases appropriate accelerated testing could produce adequate data, and therefore also better monitors, for a reasonable budget.

Accelerated testing, whether for OBDs or other applications, is built on the principle that information is lost when a time to failure is replaced by the knowledge that the lifetime exceeds a certain censoring time such as time on test; if certain censored observations can be converted into observed failures by calibrated increases in stress, the resulting increases in information can yield more accurate predictions of lifetime distributions in less time for less money. This can be seen in the standard derivation of maximum likelihood estimation for exponentially distributed failure times with mean or characteristic life *R*, where the Fisher information for ln(*R*) is $J_{\ln(R)} = n[1 - \exp(-t_0/R)]$ with *n* observations using a censoring time t_0. Typical OBD testing involves running only one unit to its design life, so *n* = 1. Moreover, the censoring time t_0, while quite large, is often small relative to *R*. If this holds, then the first order Taylor expansion for e^x gives us roughly $J_{\ln(R)} \approx t_0/R < 1$. However, if accelerated testing makes *R* sufficiently small, we can also reduce t_0 enough to allow increasing *n* while still having *R* modest relative to the reduced t_0, which would also make the information per observation closer to 1. This would increase $J_{\ln(R)}$ from a number much smaller than 1 to something closer to *n*. (The Fisher information depends on the parameterization; since we anticipate manipulating *R* to generate early false alarms, we consider here the information for ln(*R*) rather than *R* or R^{-1} as $J_{\ln(R)}$ is dimensionless and roughly interpretable as squares of percentage change.)

Of course, the utility of this rests on the adequacy of the model relating the stress conditions to normal use. For example, in traditional accelerated life testing, an Arrhenius or Eyring relationship is used to model the effect of temperature on lifetime of many materials (see Nelson [4] and Meeker and Escobar [5]) with test results used to estimate critical parameter(s) of the Arrhenius or Eyring relationship. Deroune, Parmon and Lemos [6] describe applications of accelerated testing for different types of catalysts. As we will illustrate below, a relationship between parameters of the run length distribution and the process parameters can serve a role for accelerated testing of monitors similar to Arrhenius or Eyring relationships for testing materials.

We can often record the time to failure under stressed conditions with less time and money than would be required to observe a lifetime of normal use of a product. In this article we will describe methods for accelerated testing of monitoring systems with specific emphasis on threshold selection and false alarm rate estimation.

2. THRESHOLD SELECTION

As an introduction to our proposals for accelerated testing of monitoring systems, we will first discuss the selection of thresholds for triggering alarms. In general terms, alarm thresholds are selected to balance the probability of an excessive delay against the probability of a false alarm. Figure 1a shows the conceptual relationship between the threshold, *h*, the probability of a false alarm and an excessive delay for setting a valid alarm. As illustrated in this graph, increasing the threshold reduces the probability of a false alarm while increasing the probability of an excessive delay.

(Figure 1 about here)

Ideally the threshold is selected such that both error rates are sufficiently low. However, in practice complete separation as exemplified in Figure 1a is not always possible. Figure 1b represent a situation where there is no clear separation. For such cases a compromise must be made between the probabilities of a false alarm and an excessive delay.

Conceptually, threshold selection requires an evaluation of the consequences ("cost") of an excessive delay versus a false alarm. Different applications require different compromises. In the automotive context roughly 100 − 150 OBDs are running simultaneously all tied to a single malfunction indicator light (MIL). Thus a false alarm problem caused by one monitor may raise questions about all. To avoid "teaching" the driver to ignore the MIL, the false alarm rate for the individual monitor must be a small fraction of one percent in the design life of the plant. Thus this is an application where the Type I error rate must receive serious attention to avoid negating the original intent of the monitoring scheme.

An implantable defibrillator designed to detect and interrupt excessive heart rate (tachycardia) exemplifies the opposite extreme. A false alarm (Type I error) implies that the patient gets an unnecessary and uncomfortable but not life threatening electric shock to the heart. However, an excessive delay to detect a problem, a Type II error, may mean that tachycardia proceeds uninterrupted to patient death (Gunderson [7]). The point here is that the design targets for Types I and II error rates will vary with the application and will likely not be the same for an implantable defibrillator as for automotive emission controls.

3. A MATHEMATICAL BASIS FOR ACCELERATED TESTING

A major challenge associated with testing the performance of a monitor is how to reliably estimate the false alarm rate (Type I error) across the entire design life. Conceptually, the smaller the false alarm rate, the fewer actual events will likely occur in a given test period, and the longer the process must be observed to validate compliance with design targets. Further, the smaller the false alarm rate, the larger the relative uncertainty associated with any estimate of it.

Lai [8] reported that the distributions of run lengths to false alarms for generalized likelihood ratio (GLR) are often approximately exponential. This is consistent with our own unpublished simulation studies of algorithms like Cusums, EWMAs and 2-in-a-row of both independent and autocorrelated observations. Thus, for the examples considered in this article, we shall assume that the probability of a false alarm before time t is

$$\Pr\{T \le t\} = 1 - \exp(-t/R), \tag{1}$$

where R = mean time to a false alarm. We will denote the false alarm rate $\Pr\{T \le t\}$ with π.

Now suppose a theoretical relationship $R = R(h, \boldsymbol{\theta})$ between the average run length R, the threshold h, and a vector of process parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ ' is known. This relationship can then be the basis for developing accelerated testing schemes. Specifically, we can test a monitoring system under "stressed" conditions, where for example the threshold is lowered or process parameters are changed. The results of this testing can be used to estimate one or more components of $\boldsymbol{\theta}$, which are then used with the relationship $R = R(h, \boldsymbol{\theta})$ to predict the average run length under normal conditions.

Example 1. Suppose we want to develop a Cusum monitor for an automobile that receives a normally distributed independent input signal with mean μ and variance σ^2 once every mile. Further suppose the expected design life is $t_1 = 100{,}000$ miles. For the Cusum, Bagshaw and Johnson [9] provided an asymptotic approximation for *R*, later modified by Reynolds [10] and Siegmund [11] given by

$$R = R(h,\mu) = \frac{1}{2\mu^2}\left[\exp(-2\mu h_a) - 1 + 2\mu h_a\right], \tag{2}$$

where $h_a = h + 1.166$, *h* is the threshold, and μ is the mean of the observations Cusummed, rescaled to standard deviation $\sigma = 1$; for $\mu = 0$, $R(h, \mu)$ is defined by the appropriate limit. (Graves et al. [12] present refinements of $h_a = h_a(h,\mu)$ that make (2) more accurate.)

This relationship is shown graphically in Figure 2. Of particular importance for accelerated testing, notice that for $\mu < 0$ (i.e. "good" vehicles), small changes in *h* produce very large changes in *R*.

(Figure 2 about here)

Now suppose the data coming from a "good" prototype system has been rescaled to variance $\sigma^2 = 1$, which made the mean of the input signal $\mu = -2$. We also assume that for a "good" system, μ will not change over time. To achieve a design goal of a false alarm rate less than 1% in the t_1 = 100,000 mile design life of the plant, an engineer considers setting the threshold at $h = 3.4$. With one observation per mile, this implies using (2) that $R \approx 10{,}700{,}000$. Next, using (1) we find that the probability of a false alarm in t_1 = 100,000 miles with one observation per mile will be $\pi_1 \approx 0.0092$.

Thus, to the extent that these computations are accurate, if we tested 1,000 vehicles 100,000 miles, we might expect to see roughly 9 false alarms. However, that level of testing is not feasible, yet without some similar level of testing, it can be practically impossible to obtain a reasonable upper confidence limit for the probability of such rare events.

However, suppose we artificially lowered the threshold to $h = 1$ during a test period. Then $R \approx 724$ and the probability of a false alarm in a test period of duration t_0 = 1,000 miles would be $\pi_0 \approx 0.75$. False alarms would then happen so frequently that a much shorter test period would produce data to support substantially improved prediction by extrapolation of the performance of the monitor with more realistic production thresholds around $h = 3.4$. Although we assumed in (2) that we could rescale the data to σ = 1, our procedure includes provisions for estimating a value for σ different from 1 in a way that will hopefully compensate adequately for violations of assumptions such as dependence between observations and nonnormality. The resulting $\hat{\sigma}$ would not be the standard deviation of the dependent, nonnormal observations we have but of hypothetical, independent normal observations for which the monitor would have essentially the same run length distribution.

4. A CONCEPTUAL FRAMEWORK FOR ACCELERATED TESTING

As indicated above, accelerated testing of monitors for empirical estimation of a false alarm rate is based on data from the actual system under “stressed” conditions. The base for the acceleration is knowledge of an approximate relationship $R = R(h, \boldsymbol{\theta})$ between the average run length, the threshold h and the process parameters $\boldsymbol{\theta}$ so we can predict what the average run length will be under both normal and accelerated conditions.

Typically the process parameters θ will be the mean and the variance of the data generating process.

These considerations lead to three different modes of extrapolation that can be used in accelerated testing for OBD:

1. *Threshold*: Extrapolate from artificially low thresholds required to induce false alarms during a test period, h_0, to production thresholds, h_1, at which few "good" plants will generate false alarms in the design life.
2. *Condition*: Adjust for anticipated variations in the condition of the plant to account for unit-to-unit variations at the end of the production line and for deterioration in the condition of the plant over its design life assuming these conditions remain within limits of what is considered "good", e.g., from mean μ_0 during testing to μ_1 for some relevant portion of the design life.
3. *Time*: Adjust for the longer time period in the design life, t_1, relative to the test period t_0.

Traditional accelerated testing increases stress (here threshold) to reduce time. We add *condition* to reduce the gap between the beautiful models and the messy realities with which all practical engineers must work.

We will now elaborate further on these three modes of acceleration.

4.1 Variations in Detection Thresholds

As mentioned with (1), the exponential distribution provides a reasonable approximation to the run length distributions of many monitors even with serially autocorrelated observations under constant good conditions of the plant. Certain patterns of deterioration might give this run length distribution an increasing hazard rate, possibly making the Weibull a better model than the exponential, but we have not explored that possibility.

With automobiles, observations will usually be serially dependent, non-normally distributed and have non-constant variance. Research by Bagshaw and Johnson [9] suggests that the average run length for the Cusum of *autocorrelated* normal observations can be adequately modeled by assuming independent, normally distributed observations by appropriately adjusting σ. The work of Siegmund [11] seems to suggest that further adjustments of σ might similarly compensate for nonnormality. Also, transformations can often reduce simultaneously inhomogeneity of variance and nonnormality.

Based on these considerations it seems reasonable to assume for a broad class of problems that we can obtain an approximate relationship $R = R(h, \theta)$ between the average run length *R*, the threshold *h* and process parameters θ. The process parameters will typically be the mean μ and the standard deviation σ so that $R = R(h, \mu, \sigma)$. The probability of a false alarm for μ = a constant good condition in a time period *t* is then approximately $\Pr\{T \leq t\} = 1 - \exp[\,-t / R(h, \mu, \sigma)\,]$. This expression then summarizes for a given σ the effect on the probability of an alarm for h = threshold, μ = condition of the plant, and *t* = time. For accelerated testing of monitors we want to extrapolate $\Pr\{T \leq t \mid h, \mu, \sigma\}$ from $t = t_0$ to $t = t_1$, from test thresholds h_0 to production thresholds h_1, and from the actual condition μ_0 during testing, etc.

Whatever run length distribution we use, we are primarily interested in the lower tail of the lifetime distribution and are largely unconcerned with whether the upper tail of the distribution is accurately characterized. This helps justify estimating *R* from censored run lengths – provided we observe enough false alarms to support reasonable estimation of the parameter(s) of the distribution used. Indeed, a run length distribution that was accurate *except in the lower tail* would be completely useless for our purposes; moreover, this discrepancy could only be detected with substantial observations in that lower tail, and these would be difficult to obtain prior to commercialization without accelerated testing and goodness of fit tests using likelihood for censored observations.

4.2 Variation in the Condition of the Plant

In the OBD context, it is frequently reasonable to assume that the condition of the plant is adequately characterized by the mean μ of that condition and estimated by a short-term average of observations for that parameter. This condition can vary from good to bad as illustrated in Figure 3. The scale in Figure 3 has two special points marked "worst acceptable (w.a.)" and "best unacceptable (b.u.)". Conditions w.a. or better are considered good, while conditions b.u. or worse are considered bad. Design objectives for a monitor are naturally stated in terms of the timeliness of detecting a bad condition and the maximum proportion of good units expected to generate a false alarm in the design life of the plant. The difference between these requirements means that there must be separation between w.a. and b.u. Otherwise, no feasible solution will exist.

(Figure 3 about here)

In the automotive context, the condition μ exhibits two types of variability. First, different units at the end of the production line will differ due to production variability. Second, a given unit will also exhibit wear over its entire useful life. Moreover, there will typically be substantial variation between units in the pattern of deterioration over time. In Figure 4 we illustrate symbolically the life trajectories for four different vehicles: one constant as new (a.n.), one constant w.a., one deteriorating linearly from a.n. to w.a., and the fourth deteriorating abruptly from a.n. to w.a. at 20,000 miles. [The term "new" here is used in two different senses: It is used to denote the time period immediately following completion of production of the unit and preceding any detectable deterioration or damage. It is also used in the phrase "as new (a.n.)" to denote a typical condition for units in this time period.]

From the manufacturers' and regulators' points of view, the issue is not necessarily the performance of the OBD on an individual vehicle. The primary concern is with the performance of entire fleets.

Now suppose sufficient prior data are available to develop models of the stochastic nature of the input signals across a large population of products and over time. Such models can then be used in conjunction with $R = R(h, \mu, \sigma)$ and data obtained from a smaller test fleet to estimate the false alarm rate for a larger population of products.

(Figure 4 about here)

4.3 Probabilities of a False Alarm for Varying Durations of Time

As above, (t_i, h_i, μ_i) = time, threshold, and condition of the plant during testing (i = 0) and production (i = 1). The required extrapolation can be accomplished with the knowledge of $R = R(h, \mathbf{\theta})$ and the cumulative probability distribution function (cdf) of the time to a false alarm, $\Pr\{T \leq t\}$, as a function of t.

A key issue is the selection of the test period t_0. Meeker and Escobar [5] note that reducing t_0 will tend to (a) increase the sensitivity of the analysis to deficiencies in the model used for extrapolation, and (b) amplify the effect of sampling variability. They quote Evans [13], who suggested that acceleration factors of 10 “are not unreasonable”, but factors much larger may involve excessive risks in extrapolation. This advice in the OBD context would have $t_0 = 10{,}000 = 0.1\ t_1$.

In the next section we will describe more specifically how accelerated testing for OBD is organized.

5. ORGANIZING ACCELERATED TESTING FOR OBD

Figure 5 summarizes hypothetical 10,000 mile testing at six different thresholds, h = 0.5, 1.0, 1.5, 2.0, 2.5, and 3, at which 28, 16, 7, 3, 1 and 0 false alarms were recorded at the mileages indicated in the plot. These data could be collected in at least two different ways. Either the raw data from a single unit could be processed in parallel against several different thresholds or several units could be tested with different thresholds. In either case, each time a monitor exceeds its threshold, appropriate data are stored and the monitor is reinitialized.

(Figure 5 about here)

In the next section, we illustrate our methodology assuming for simplicity that the only data available were the top three lines of Figure 5, for h = 2.0, 2.5, and 3.0. These are shown in Figure 6. We see that at h = 3.0, no false alarms were observed in the 10,000 miles. Thus, this one observation was *censored* at 10,000 miles. We write this as $t_{3,1}^{c}$ = 10,000 miles where the superscript “c” indicates that we observed a censoring time of 10,000 miles, and the run length exceeded that. The two subscripts “3,1” denote the threshold, 3, and index the observations at that threshold, 1.

(Figure 6 about here)

At h = 2.5, one false alarm was observed at 6,528 miles. The monitor was reset, and the vehicle went the remaining 3,472 of the 10,000 miles without another false alarm. We denote this as $t_{2.5,1}$ = 6,528 and $t^c_{2.5,2}$ = 10,000 – 6,528 = 3,472. (If a false alarm was set at 6,528 miles, but the monitor was not reset until 6,537, then $t^c_{2.5,2}$ = 3,463 not 3,472.)

Similarly, at h = 2.0, false alarms were observed at 2,760, 5,768, and 9,272 miles. We denote these $t_{2,1}$ = 2,760, $t_{2,2}$ = 5,768 – 2,760 = 3,008, and $t_{2,3}$ = 9,272 – 5.768 = 3,504. No failures were observed in the remaining 728 of the 10,000 mile testing, which we denote as $t^c_{2,4}$ = 728.

We will assume that these run lengths are all statistically independent. This assumption would be violated if the run lengths to different thresholds were obtained from monitors using different thresholds processing data from the same test unit. For example, two runs of 100 observations each with a threshold of 1 might be generated from the same observations that produced a run of 200 observations against a threshold of 2. Clearly, there would be statistical dependence between these three run lengths, which will not be considered in the present article; we assume instead that each test unit employs only one threshold. (Of course, if a manufacturer had the capability to process all the data simultaneously against different thresholds, we would encourage that practice. If the effects of this dependency were found to be material, we would need to develop a more appropriate analysis methodology, possibly using Monte Carlo, e.g., Robert and Casella [14] and Carlin and Louis [15]. We would be negatively impressed with any statistician who would tell engineers to collect less data, just because they didn’t know how to model the dependence!)

A more subtle issue is that high serial dependence between successive observations could generate serial dependence between successive run lengths computed from those observations. This problem would be greater with lower thresholds; successive runs to higher thresholds would necessarily employ observations that are farther apart and therefore presumably less serially dependent. For the present discussion, we assume that this effect is negligible.

Another issue is that even modest levels of serial dependence between successive observations can seriously distort the run lengths, and average run lengths in particular, of monitors (e.g., Cusums, EWMAs) computed from those numbers. In the present discussion, we assume that an alternative choice for the standard deviation of these observations can adequately compensate the existing level of serial dependence, as mentioned above in section 4.1. As previously stated, if any of these assumptions are considered inappropriate, the present methodology could be revised to simulate via Monte Carlo assumptions considered more appropriate.

6. A WORKED EXAMPLE

Monitors that trigger on the second observation in a row exceeding a threshold are sometimes legally prescribed and used in the automotive context. We shall assume that the run lengths in Figure 5 are run lengths to the second observation in a row exceeding a threshold. We do this largely to simplify the discussion of accelerated testing principles, because the theoretical properties for *k*-in-a-row monitors are better known and more easily described (see below) than for other monitors such as Cusums or EWMAs that may make more efficient use of data typically encountered in engineering applications. As an example we will apply the accelerated testing methodology to a 2-in-a-row monitoring rule assuming one observation is processed per mile and the design life of the vehicle is $t_1 = 100{,}000$ miles using the simulated data in Figure 6.

In the automotive context often as many as 150 monitors run simultaneously tied to the same malfunction indicator light (MIL), with each testing for a different type of malfunction. If 50 of them were statistically independent with false alarm rates for each of 2% in the design life of the vehicle (while the false alarm rate for the other 100 were negligible), this would generate on average one false alarm per vehicle design life. To avoid this, we will therefore select a much smaller target false alarm rate of 0.0005 = 0.05%.

However, in applying this design target, we need somehow to consider variations in the condition of the plant, both between vehicles and for a given vehicle between the time that it is new to mature use. Many monitors exhibit behavior qualitatively similar to the Cusum approximation (2), where *R* decreases roughly exponentially as the condition of the plant μ becomes less negative, moving from "as new" towards "worst acceptable". Because of this, we shall assume that only about half of the units spend any appreciable amounts of time near "worst acceptable", and that the probability of a false alarm for the other half can be ignored. Thus we will set the design target for the false alarm rate as $\pi_1 = (0.000\,5)/0.5 = 0.001$.

We will assume that the input signals to the monitor are independent normally distributed observations with mean μ and standard deviation σ. The probability that a single independent observation will exceed the threshold *h* is therefore

$$p = \Phi[(\mu - h)/\sigma], \tag{3}$$

where Φ[.] is the cumulative distribution function (cdf) for the normal distribution.

For an alarm that triggers on the *k*th observation in a row exceeding a threshold, Feller [16, p. 324] gives the following expressions:

$$\begin{aligned} R(h,\mu,\sigma) = E\{T\} &= \frac{(1-p^k)}{(1-p)p^k} \\ &= (1+p)/p^2\text{, if } k=2\text{,} \end{aligned} \tag{4}$$

and

$$\begin{aligned} \operatorname{var}(T) &= \frac{1}{\left[(1-p)p^k\right]^2} - \frac{2k+1}{(1-p)p^k} - \frac{p}{(1-p)^2} \\ &\approx \frac{1}{\left[(1-p)p^k\right]^2}\text{, if } p \text{ is small,} \end{aligned}$$

for p from (3). We include the formula for the variance of the run length, T, because it shows that when p is small, i.e., when the condition of the plant is good, the standard deviation of T is approximately equal to the mean. This is a property of the exponential distribution and helps justify its use in this context. Thus for good plants (small p's), we will assume that the run length distribution is approximately exponential.

As noted above, we require the false alarm rate to be at most $\pi_1 = 0.001$. Since run length distributions for good plants are approximately exponential, this means that Pr{false alarm in t_1} = $\Pr\{T \leq t_1\}$ = $\left[1-\exp(-t_1/R)\right] \leq \pi_1$. After some algebra this inequality can be written as

$$R \geq \{t_1 / [-\ln(1-\pi_1)]\}.$$

When π_1 is small, which will be the case for this application, $\ln(1-\pi_1) \cong (-\pi_1)$. Thus this inequality is virtually equivalent to

$$R \geq (t_1 / \pi_1). \tag{5}$$

Note that expression (5) applies to any monitor that has approximately an exponential run length distribution including the Cusum, the EWMA, and *k*-in-a-row. Note further that for our example with t_1 = 100,000 miles and π_1 = 0.001 the lower bound on the theoretical R is 100,000,000 miles. (The absurd magnitude of this number underscores the need for accelerated testing.)

We will now apply (5) to the 2-in-a-row example to find the threshold h. From (4), we have that $R = R(h, \mu, \sigma) = (1+p)/p^2$. If R must be large, p must be small. Hence $R(h, \mu, \sigma) \approx 1/p^2$. We combine this with (3) and (5) to get

$$1 / \{ \Phi[(\mu - h)/\sigma] \}^2 \geq (t_1 / \pi_1).$$

After a little algebra we get

$$\mu - \sigma \ \Phi^{-1}\left\{\sqrt{\pi_1/t_1}\right\} \leq h. \qquad (6)$$

To solve (6) for h we need estimates of μ and σ. We will assume for simplicity that the data are scaled to have mean $\mu = 0$. As noted above, we will estimate σ for the observations indirectly from its impact, via (4) and (3), on the average run length data in Figure 6. We do this using maximum likelihood. For an exponential distribution with mean R, the probability density for a run of length t is $\exp(-t/R)/R$ and the probability that the run length exceeds a censoring time t^c is given by $\Pr\{T > t^c\} = \exp(-t^c/R)$. Hence the log(likelihood) is

$$l(\sigma) = \log[L(\sigma)] = \sum_{\substack{\text{all} \\ \text{runs}, i}} [-t_i / R(h_i, \mu, \sigma)] - \sum_{\substack{\text{censored} \\ \text{runs}, j}} \log[R(h_j, \mu, \sigma)] \ . \qquad (7)$$

where $R(h_i, \mu, \sigma)$ is as given above. (Note that we have dropped the superscript "*c*" used to denote censoring times in Figure 6 and collapsed the two subscripts into one.) For the numbers in Figure 6, we found that the likelihood was maximized at $\hat{\sigma}$ = 0.977.

We can get an approximate 95% confidence interval as

$$\{ \ \sigma \mid [-2 \log(\text{likelihood}(\sigma))] \leq \chi^2_{1,.05} + [-2 \log(\text{likelihood}(\hat{\sigma}))]\} \qquad (8)$$

(McCullagh and Nelder [17, sec. 7.4.1]; Cox and Hinkley [18, 343]). For this example, log(likelihood($\hat{\sigma}$)) = – 38.27 and $\chi^2_{1,.05}$ = 3.84. Hence we need values of σ for which the log(likelihood) exceeds (– 38.27 $-3.84/2$) = -40.19. The easiest way to obtain a confidence interval is to make a profile plot of the log likelihood function around the maximum, see Meeker and Escobar [5, sec. 8.3.2]. Using this approach we find that an approximate 95% confidence interval for σ is (0.885, 1.064).

When setting thresholds we want to be on the safe side providing an upper bound for the false alarm rate. Thus because *R* decreases with σ , we will use the upper bound for σ to obtain a lower confidence bound on the threshold required *h*. Hence we will use $\hat{\sigma} = 1.064$ in the following calculations. (If we apply the techniques described here to the standard problem of the mean of normally distributed observations with a known variance and use only the upper bound, we get the standard 97.5% one-sided confidence bound for that problem. However, since the likelihood (7) is generally not symmetric, it is not obvious how to evaluate the actual confidence level of this one-sided limit for σ other than to observe that it exceeds 95%, assuming the use of the chi-square approximation here is adequate. A more precise confidence bound could be obtained with bootstrapping, as described by Meeker and Escobar [5, ch. 9] and Chernick 19, ch. 3].)

Using $\hat{\sigma}$ = 1.064 and t_1 = 100,000 miles we get $\pi_1 / t_1 = 10^{-8}$, so $\sqrt{\pi_1/t_1} = 10^{-4}$. Then from (6),

$$h \approx 0 - 1.064\Phi^{-1}\left\{\sqrt{\pi_1/t_1}\right\} = -1.064\,(-3.72) = 3.96.$$

A threshold calculation like this would be based on data from test vehicles that presumably were all "good" with $\mu = 0$ throughout the test period. In practice, there will be variability between vehicles coming off the production line causing vehicle-to-vehicle variability. In addition, vehicles age over time without necessarily crossing the worst acceptable (w.a.) line of Figure 3. To take this into account we will consider a w.a. vehicle one with $\mu = 0.5$. For the 2-in-a-row rule considered here, the design goal π_1 for the false alarm rate should therefore apply for such a vehicle and the lower bound for the threshold should be $h \approx 0.5 - (1.064) \times (-3.72) = 4.46$.

This lower bound for *h* adjusts for certain issues but not others. We explicitly considered the random variability in the data for the vehicles tested, but it may also be desirable to make a further adjustment for the between-vehicle component of variance. Other modifications to our procedure would be possible if data were available to model how the mean and standard deviation of the observations change with age. We have assumed in the above that (*a*) the measurement standard deviation remains constant throughout the product life and (*b*) only the mean changes with age. For certain automotive catalyst monitors that rely on counts of "rich-to-lean" and "lean-to-rich" transitions in the exhaust gas downstream of the catalytic converter, the standard deviation also increases with age: When the catalyst is new, this number and its standard deviation are both quite small. As the catalyst ages, both the mean and the standard deviation increase. If this pattern could be modeled, more refined estimates of the false alarm rate could be developed. A simple step in this direction might be to monitor the square roots of switch counts, recalling that the square root is the traditional variance stabilizing transformation for Poisson counts. Even if these counts were not Poisson, this transformation might still reduce the sensitivity of the standard deviation to changes in the mean.

We have considered a procedure for estimating and controlling the false alarm rate. A separate analysis is needed to determine if the monitor considered in this example with a threshold of 4.46 will signal sufficiently fast for a best unacceptable (b.u.) plant. We will not discuss this issue, except to note that it may not be appropriate to assume that the run length distribution is exponential for bad plants.

6. DISCUSSION

Several major automobile manufacturers currently spend large sums every year verifying that new vehicles will not have a major false alarm problem. Unfortunately, even with these substantial investments, they still have difficulties obtaining reasonable estimates of the false alarm rates for their monitors. This article has described a methodology that will allow organizations to evaluate false alarm rates more precisely than in the past, in less time for less money, while simultaneously doing a better job of establishing detection thresholds.

This reduction in cost and time to market requires the user to develop models to support the extrapolations discussed above: Adjusting for the fact that prototypes are generally better than worst acceptable (w.a.), adjusting for the difference in duration between the test period and the design life, and extrapolating from test to production thresholds. The methodology was described, and a hypothetical example was presented.

While this theory can be used to support a substantial reduction in the 100,000 mile testing of new vehicles performed by some automobile manufacturers, it would not be wise to eliminate completely the 100,000 mile testing. This is because an extrapolation methodology that works in some contexts may be inadequate in others. For example, the theory of statistical confidence intervals in expression (8) adjusts for errors assuming the model is correct, but may not provide the anticipated protection if the model is seriously deficient.

Decisions regarding the desired magnitudes of both accelerated testing and follow-on model verification can be performed by following two recommendations of Meeker and Escobar [5]: (*a*) Simulate data collection and analysis for a wide range of plausible situations for reality and for the test plan. (*b*) Use Bayesian reliability theory to pool information from similar monitors across a variety of products to obtain tighter estimates of the false alarm rate from limited testing. This will help decision makers formally balance costs and risks of a variety of procedures.

Finally, we acknowledge that changes to existing practices in an organization often require substantial justification. In some organizations, part of this justification may include the fact that certain types of accelerated testing are already legally mandated for emission control on new automobiles (Mondt [20, p. 37]).

7. ACKNOWLEDGEMENTS

This research was supported by contracts between (*a*) the Low Emissions Partnership of Ford, DaimlerChrysler, and General Motors and (*b*) the University of Wisconsin, the University of St. Gallen, and Søren Bisgaard and Associates.

REFERENCES

1. Box G, Graves S, Bisgaard S, Van Gilder J, Marko K, James J, Seifer M, Poublon M, Fodale F. Detecting Malfunctions in Dynamic Systems. *Transactions of the*

Society of Automotive Engineers: Electronic Engine Controls 2000, *SAE Technical Paper Series* 2000-01-0363, pp. 1-11.

2. CARB. Malfunction and Diagnostic System Requirements--1994 and Subsequent Model-Year Passenger Cars, Light-Duty Trucks, and Medium-Duty Vehicles and Engines (OBD II), with modifications effective as of September 25, 1997, sec. 1986.1 of Title 13,California Code of Regulations. Sacramento, CA: California Air Resources Board; available from http://www.arb.ca.gov/msprog/obdprog/obdprog.htm, accessed Nov. 14, 2004.
3. Box G, Bisgaard S, Graves S, Kulahci M, Marko K, James J, Van Gilder J, Ting T, Zatorski H, Wu C. Performance Evaluation of Dynamic Monitoring Systems: The Waterfall Chart. *Quality Engineering*, 2003 **16**: 183-191.
4. Nelson W. *Accelerated Testing*; Wiley: NY 1990.
5. Meeker WQ, Escobar LA. *Statistical Methods for Reliability Data*. Wiley: NY 1998.
6. Deroune EG, Parmon V, Lemos F. *Principles and Methods for Accelerated Catalyst Design & Testing*; Kluwer: NY 2002.
7. Steiner SH, Cook RJ, Farewell VT, Treasure T. Monitoring Surgical Performance Using Risk Adjusted Cumulative Sum Charts. *Biostatistics* 2000 **1**: 441-452.
8. Lai TL. Sequential Change Point detection in Quality Control and Dynamic Systems. *Journal of the Royal Statistical Society, Series B* 1995 **57**: 613-658.
9. Bagshaw M, Johnson RA. The Effect of Serial Correlation on the Performance of Cusum Tests II. *Technometrics*, 1975 **17**: 73-80.

10. Reynolds MR Jr. Approximations to the Average Run Length in Cumulative sum Control Charts. *Technometrics* 1975 **17**: 65-71.
11. Siegmund D. *Sequential Analysis*; Springer Verlag: NY 1985.
12. Graves SB, Kulahci M, Bisgaard S, James J, Marko K, Ting T, Van Gilder J, Wu C, Zatorski H A New Approximation for the Average Run Length of a Cusum ... 2004
13. Evans RA Accelerated Testing. *IEEE Transactions on Reliability* 1991 **R-40**: 497 (quoted from Meeker and Escobar 1998, p. 489).
14. Robert CP, Casella G. *Monte Carlo Statistical Methods*; Springer Verlag: NY 1999.
15. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*; Chapman and Hall: NY 1996.
16. Feller W) *An Introduction to Probability Theory and Its Applications*, vol. I, 3rd ed., rev. printing; Wiley: NY 1970.
17. McCullagh P, Nelder JA. *Generalized Linear Models*, 2nd ed.; Chapman and Hall: NY 1989.
18. Cox DR, Hinkley DV. *Theoretical Statistics*; Chapman and Hall: London 1974.
19. Chernick MR. *Bootstrap Methods*; Wiley: NY 1999.
20. Mondt JR *Cleaner Cars*; Society of Automotive Engineers: Warrendale, PA 2000.

Author's Biographies

Spencer Graves is a Senior Development Engineer with PDF Solutions in San José, CA focusing on the statistics of yield improvement in wafer fab. He holds Bachelors and

Masters degrees in Aerospace and Industrial Engineering, respectively, and a PhD in Statistics. His email address is spencer.graves@prodsyse.com.

Søren Bisgaard is the Eugene M. Isenberg Professor of Technology Management at the Eugene M. Isenberg School of Management at University of Massachusetts. His email address is Bisgaard@som.umass.edu.

Murat Kulahci is an Assistant Professor at the Department of Industrial Engineering at Arizona State University. His email address is kulahci@asu.edu.

John Van Gilder
Powertrain Control Center, GM Proving Grounds
31E Powertrain
Milford, MI 48380-3726
phone: 248-685-5845
e-mail: john.vangilder@gm.com

John James
Advanced Diagnostics, Ford Research Labs, MD-1170
PO Box 2053, 20 000 Rotunda Dr.
Dearborn, MI 48121-2053
phone: 313-323-1039
e-mail: jjames3@ford.com

Ken Marko
Advanced Diagnostics, Ford Research Labs, MD-1170
PO Box 2053, 20 000 Rotunda Dr.
Dearborn, MI 48121-2053
phone: 313- 390-1379
e-mail: kmarko@ford.com

Hal Zatorski
On-Board Diagnostics Program Manager
DaimlerChrysler Corporation, CIMS: 482-01-07
800 Chrysler Dr.
Auburn Hills, MI 48326-2757
phone: 248-576-5006
e-mail: hz2@daimlerchrysler.com

Tom Ting
General Motors Research, Development and Planning
30500 Mound Rd. , MC 480-106-390
Warren, MI 48090
phone: 810-986-3356
e-mail: Tom_Ting@gmrnotes3.gmr.com

Cuiping Wu
Reliability Engineer
DaimlerChrysler Chelsea Proving Grounds, CIMS 422-01-10
3700 S. M-52
Chelsea, MI 48118-9600
phone: 734-475-5668
e-mail: CW59@daimlerchrysler.com

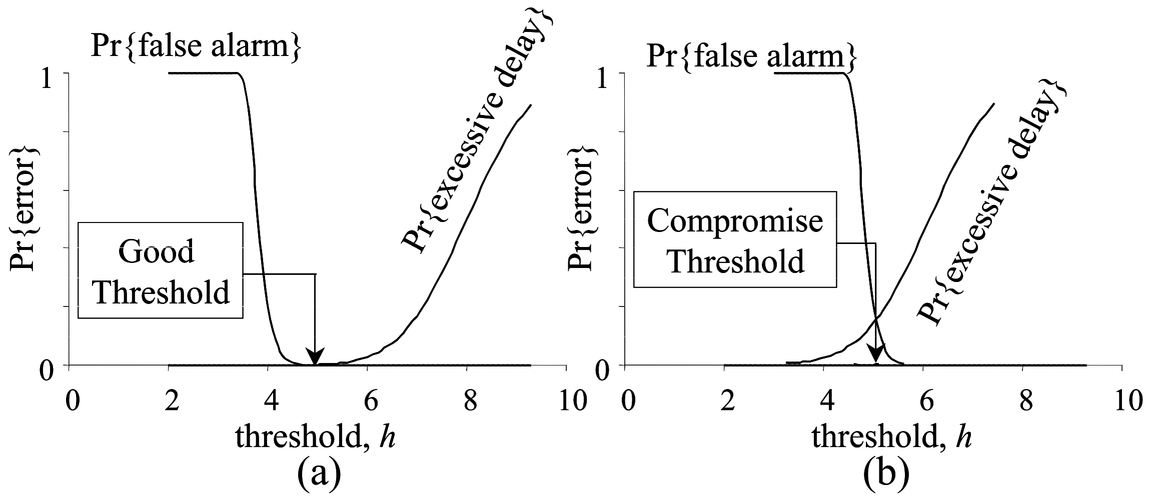

Figure 1. The probabilities of false alarms and excessive delays as functions of the threshold, *h*. For threshold selection we need to balance these two probabilities. (a) Monitor with clear separation and (b) without clear separation.

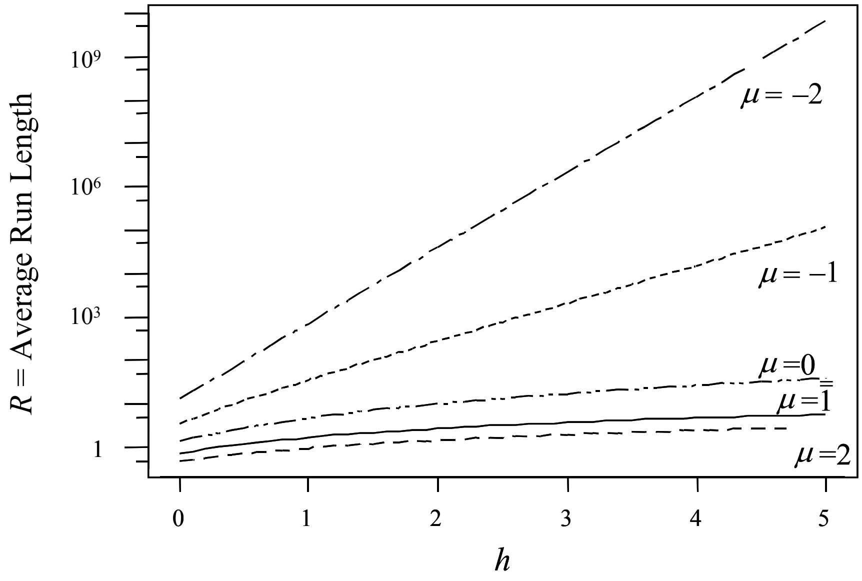

Figure 2. The average run length given by (2) as a function of the threshold *h*.

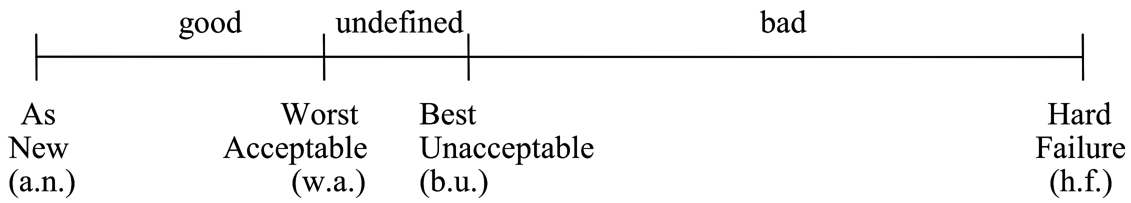

Figure 3. The terminology typically used in the OBD context for the range of conditions of the plant.

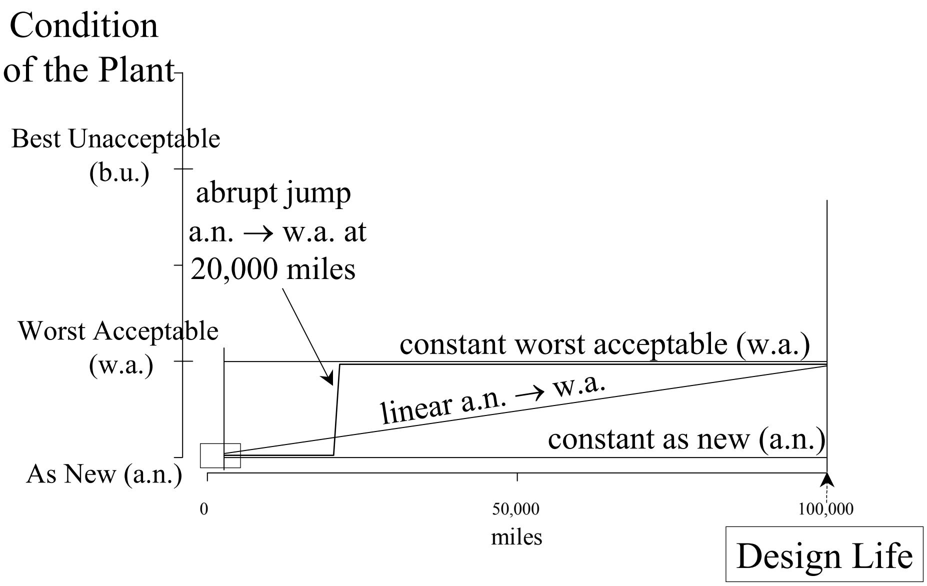

Figure 4. Variations in the Condition of the Plant

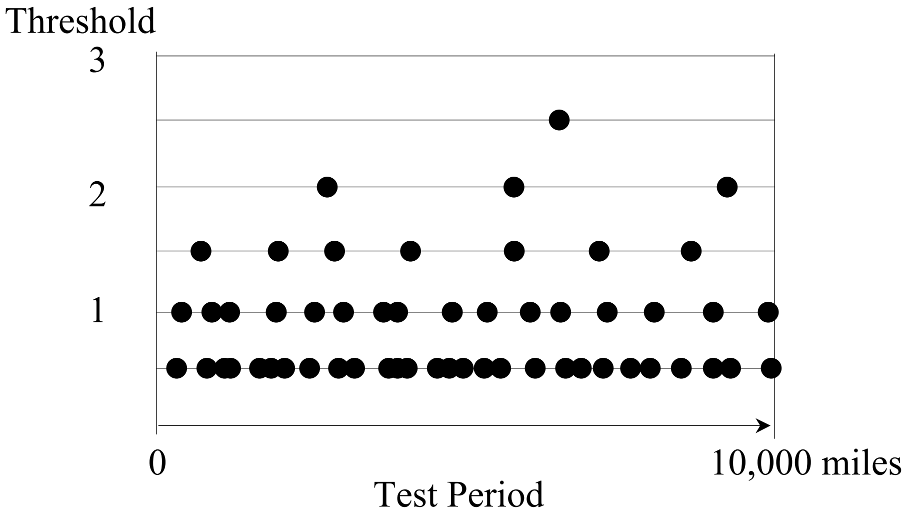

Figure 5. Testing 10,000 Miles at Multiple Thresholds

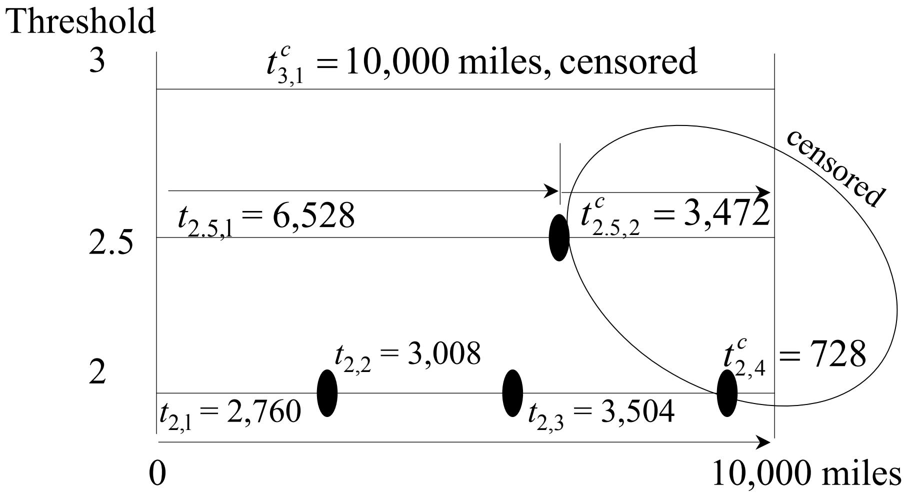

Figure 6. Time to False Alarms from 10,000 Mile Testing at Multiple Thresholds

Figure Captions

Figure 1. The probabilities of false alarms and excessive delays as functions of the threshold, *h*. For threshold selection we need to balance these two probabilities. (a) Monitor with clear separation and (b) without clear separation.

Figure 2. The average run length given by (2) as a function of the threshold *h*.

Figure 3. The terminology typically used in the OBD context for the range of conditions of the plant.

Figure 4. Variations in the Condition of the Plant

Figure 5. Testing 10,000 Miles at Multiple Thresholds

Figure 6. Time to False Alarms from 10,000 Mile Testing at Multiple Thresholds