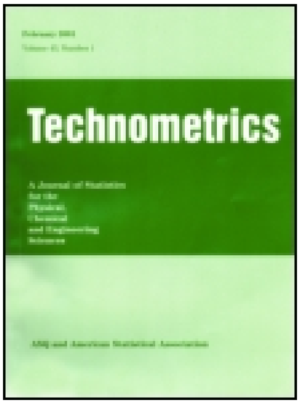

Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

A Random Onset Model for Degradation of High-Reliability Systems

Scott Vander Wiel[a], Alyson Wilson[b], Todd Graves[c] & Shane Reese[d]

[a] Los Alamos National Laboratory, Los Alamos, NM 87545

[b] Department of Statistics, Iowa State University, Ames, IA 50011

[c] Los Alamos National Laboratory, Los Alamos, NM 87545

[d] Department of Statistics, Brigham Young University, Provo, UT 84602

Published online: 01 Jan 2012.

A Random Onset Model for Degradation of High-Reliability Systems

Scott VANDER WIEL

Los Alamos National Laboratory
Los Alamos, NM 87545
(scottv@lanl.gov)

Alyson WILSON

Department of Statistics
Iowa State University
Ames, IA 50011
(agw@iastate.edu)

Todd GRAVES

Los Alamos National Laboratory
Los Alamos, NM 87545
(tgraves@lanl.gov)

Shane REESE

Department of Statistics
Brigham Young University
Provo, UT 84602
(reese@stat.byu.edu)

Weapons stockpiles are expected to have high reliability over time, but prudence demands regular testing to detect detrimental aging effects and maintain confidence that reliability is high. We present a model, called RADAR, in which a stockpile has high initial reliability that may begin declining at any time. RADAR provides a framework for answering questions about how confidence in continued high reliability can change as a result of reduced sampling, discovery of failed units, and information about when a unit failed.

Supplemental materials (available on the *Technometrics* web site) provide lemmas used in the proof of Theorem 1, details of the Markov chain Monte Carlo algorithm, and additional examples.

KEY WORDS: Aging; Bayesian; Reliability uncertainty; Sampling; Surveillance planning.

1. INTRODUCTION

Both the U.S. Department of Defense (DoD) and Department of Energy (DOE) maintain weapons stockpiles: items like bullets, missiles, and bombs that have already been produced and are being stored until needed. Ideally, these weapons stockpiles maintain high reliability over time. To assess stockpile reliability, a *surveillance* program is typically implemented, where units are removed from the stockpile and tested periodically. The most definitive tests typically destroy the weapons so a given unit is tested only once. Questions of primary interest to surveillance managers include how many units should be tested, how often should they be tested, what tests should be done, and how can the resulting data be used to estimate the stockpile's current and future reliability. These questions are particularly critical from a planning perspective: what is an appropriate and cost-effective surveillance program?

The evaluation of military and civilian surveillance programs has been of interest for over fifty years. For example, Derman and Solomon (1958) evaluated surveillance programs where units demonstrate step function or exponential decay; Hillier (1962) assumed that units decay with a common known exponential distribution and that the time until use is also a known exponential distribution. Valdez-Flores and Feldman (1989) provided a general review of preventative maintenance models, with a relevant discussion of inspection models. In general, there must be a sufficient supply of usable units available when they are needed; due to the deterioration of the units over time, a systematic surveillance program must be conducted.

The RADAR model (a Rationale to Assess Degradation Arriving at Random), defined in Section 2, describes a stockpile that will likely maintain high reliability from one year to the next under a small, but ever-present, possibility that reliability could begin to decline at any time. In particular, suppose that:

- reliability in the year of manufacture is high;
- reliability in one year is typically the same as the previous year, but certainly no better;
- in any given year there is a small chance that reliability will begin to degrade;
- once degradation starts, it continues at a fixed but uncertain (geometric) rate;
- samples are routinely drawn from the stockpile for destructive testing; individual units either pass or fail.

Under these conditions, it is clear that ongoing sampling and testing are imperative to maintain confidence that a stockpile has high reliability. The RADAR model provides a means for quantitatively assessing confidence in stockpile reliability where the potential impact of aging explicitly contributes to uncertainty. We do not explicitly model differences in reliability between different units in the stockpile because we assume that testing destroys a unit and therefore between-unit variations in reliability cannot be distinguished with available pass-fail data.

Figure 1, based on results of methodology presented in Section 2 and later, is a preview of RADAR's main characteristic: *Uncertainty grows large once you stop looking.* The figure plots median reliability estimates (black lines at 1.0) and various uncertainty bounds (boundaries of shaded regions) that would be

TECHNOMETRICS, MAY 2011, VOL. 53, NO. 2
DOI 10.1198/TECH.2011.09119

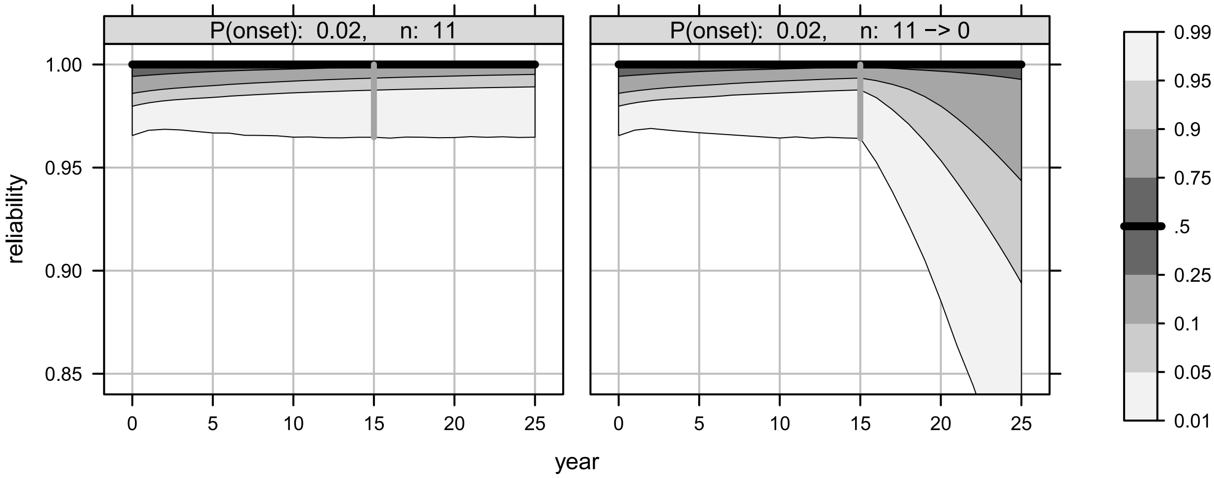

Figure 1. Reliability estimates over time when no failures are found in surveilled units. In both panels 11 units per year are surveilled from year zero through year 15. This continues through year 25 in the left panel but in the right panel no further sampling is done after year 15. Curtailment of sampling does not impact the median estimate (thick black line at reliability 1.0) but dramatically affects the lower uncertainty bands (shaded regions). Shading shows quantiles of the reliability distribution as indicated by the key. Quantiles above the median have reliability 1.0 and are therefore not visible.

obtained under a nominal sampling plan of 11 units per year starting in year zero (left panel) and in a situation when sampling ceases completely after year 15 (right panel). All units sampled are assumed to pass when tested. Input specifications used to produce the figure match the illustrative example given in Section 2, but can be summarized as follows: (a) initial reliability of the stockpile is most likely greater than 99%; (b) detrimental aging could occur at any time but is not expected for 50 years (i.e., each year, the probability of onset of degradation is 0.02); and (c) once degradation begins, it will continue at an unknown rate that is expected to be around 1% of the current reliability level each year. Complete curtailment of the surveillance plan has no impact on the median estimate of reliability (which remains at 1.0 throughout the 25-year period) but it has an immediate and progressive impact on the uncertainty about reliability. If no units are sampled after the 15th year uncertainty bounds begin to drop at rates in the neighborhood of 1% per year. Note that the lack of complete smoothness on the lowest quantile is due to Monte Carlo error.

1.1 Planning for the Unexpected

"Unknown unknowns" is military parlance for circumstances and effects that cannot be anticipated but, nevertheless, must be guarded against. (Donald Rumsfeld, U.S. Secretary of Defense, popularized the phrase by using it to describe a chaotic situation in Afghanistan in a press briefing on February 12, 2002.) In the context of stockpile surveillance, unknown unknowns are unexpected manufacturing issues or aging phenomena that have a detrimental effect on the stockpile but have not yet been discovered or envisaged. Surveillance tests naturally focus on gathering information about known aging phenomena such as chemical and mechanical processes that can be anticipated by engineers and scientists. An important side-benefit of these tests, however, is that they afford opportunities to discover potential issues of which no one has yet conceived. Prudence demands that we look for such issues, but it is difficult to include the critical value of discovering the unknown when planning future surveillance activities, especially when tight budgets press to reduce unnecessary testing. After all, if decades of tests on a component show no trends that cause concern, why should expensive surveillance continue?

RADAR provides a probabilistic description of the uncertainties that accumulate if surveillance is curtailed on a highly reliable system. The outputs depend on a prior distribution for reliability degradation over time. Following is a discussion of two different contexts in which we have used the RADAR model with emphasis on how the prior was chosen in each case.

One application of RADAR was with a group of engineers and scientists responsible for a component of a weapon system. In this case we elicited a prior distribution for future reliability trends based on the opinions of experts with decades of experience. Expert scientists are naturally concerned about unknown unknowns and relate anecdotes of confronting unexpected mechanisms in their past experiences. Not surprisingly, however, they found it difficult to assess probabilities of reliability degradation caused by aging phenomena that have not yet been understood or even conceived. Eventually they did agree on a prior and were then satisfied that the RADAR model output was a good representation of their collective concerns about the impact of reductions in surveillance rates that were being contemplated at the time.

In the second application of RADAR, a "scorecard" was being designed as a tool for high-level reporting on uncertainties in stockpile health. Posterior distributions of reliability trends were used for one type of uncertainty in the scorecard. In this case a prior distribution was determined by a consensus opinion about what represented a conservative set of degradation trajectories that a surveillance program should guard against for the purpose of discovering detrimental unknown unknowns. Again, the participating engineers agreed that posteriors from the RADAR model appropriately reflect the range of plausible degradation scenarios given the sequence of successful test results over many years. In the scorecard application, if surveillance of a component is curtailed, the health score for that component is reduced as a result of the increased risk of undiscovered aging issues.

These two applications demonstrate that the RADAR model is valuable as a tool to quantify the relationship between surveillance sampling rates and confidence that current and future reliability remains high for the difficult problem of mitigating risk from unknown unknowns. We claim that the broad features of the model are adequate for this purpose, even if the detailed mathematical description of degradation trajectories is not derived from specific physical models of aging phenomena—lack of a tight justification for a particular probabilistic model is the nature of the epistemic uncertainty of unknown unknowns.

1.2 Stockpile Surveillance

Actual surveillance programs are costly, broad, and deep, especially in the DOE, where the U.S. nuclear weapons surveillance program must "ensure, through various tests, that the reliability of nuclear weapons is maintained" in the absence of full-system testing (General Accounting Office 1996). The DOE program consists primarily of three types of tests: nonnuclear flight tests, that involve the actual dropping or launching of a weapon from which the nuclear components have been removed; systems laboratory tests, which detect changes or defects due to aging, manufacturing, and design of the weapons; and new material laboratory tests, which pull newly produced units from the manufacturing line and test them as a quality check. Fully integrated analysis of the suite of nuclear weapons surveillance data is an ongoing area of research (Wilson et al. 2007). The RADAR model captures plausible high-level features of stockpile reliability over time and can be used to answer broad *policy* questions about the surveillance programs. Our intention is to provide a framework that generates tractable answers that integrate expert knowledge and coarse (pass/fail) summaries of surveillance data to allow decision-making about appropriate trade-offs between the cost of data and the accuracy and precision of stockpile reliability estimates.

Weapon systems surveillance plans have traditionally been designed to provide a specified level of confidence in detecting problems that affect certain proportions of the stockpile within a certain period of time. For example a "90/10" plan assesses 22 units per year with a justification that a sample of 22 provides a 90% chance of detecting a problem that affects 10% of the stockpile.

The 90/10 rationale is a simple justification for a certain sample size for each year of a surveillance plan, but the rationale does not attempt to anticipate when stockpile reliability might begin to degrade or how quickly aging might progress. When paired with classical binomial confidence bounds on reliability, the result is a system built on an assumption that reliability does not change with time, at least over a given analysis window. An implicitly static model of a stockpile ignores the potential that aging effects will eventually degrade reliability and gives rise to an argument that if no defects have been found after many years of testing, then a few more years of data will do little to improve a lower credible bound on reliability.

A more realistic method acknowledges degradation in reliability as a stockpile ages and, as a result, demonstrates that uncertainty on reliability estimates grows if surveillance sampling is curtailed for a few years or ramped down to lower levels.

The RADAR model is presented in Section 2 and its posterior reliability distribution is derived in Section 3 for the case when no units have failed testing and certain parameters are known. Example calculations in Appendix B in the supplemental materials (available on the *Technometrics* web site) show how quickly a lower confidence bound can degrade if routine sampling is curtailed or if expectations about degradation are modified to reflect anomalous test results that have not resulted in failures. Markov chain Monte Carlo methods are derived in Section 4 to simulate the posterior reliability in cases where some units have failed testing and some model parameters are given hierarchical prior distributions. Several examples are discussed in Section 5. Finally, possible extensions to the model and practical aspects of using it are discussed in Section 6.

2. A RANDOM ONSET DEGRADATION MODEL

We seek an a priori description of the various possible ways that reliability can evolve over time. The following random onset model for reliability degradation is useful for planning surveillance activities because it can describe the typical situation in which, with high probability, no reliability degradation will occur throughout the design life of a stockpile and yet with each new year a nonzero probability exists that reliability may begin to decline.

As a reflection of the phenomenology above, let π_t denote the reliability of a stockpile in year t,

$$\pi_t = \Pr(\text{a random unit will function when tested in year } t),$$

and model the evolution of π_t according to the following Markov process:

$$\pi_0 = \rho_0,$$
$$\pi_t = \rho_t \pi_{t-1}, \qquad t = 1, 2, \ldots,$$

where the annual (geometric) degradation rates ρ_t evolve as

$$\rho_t = r_t, \qquad t = 0, 1, \tag{1.a}$$
$$\rho_t = r_t \rho_{t-1}, \qquad t = 2, 3, \ldots, \tag{1.b}$$

and where r_t for $t = 0, 1, 2, \ldots$ are a series of independent shocks with prior distribution,

$$r_t \sim \text{OneBeta}(p_t, a_t, b_t)$$
$$\equiv \begin{cases} 1, & \text{with probability } p_t \\ \text{Beta}(a_t, b_t), & \text{otherwise.} \end{cases} \tag{1.c}$$

If p_t is near 1 or $a_t \gg b_t$, then r_t is at or near 1 with high probability. If all the p_t are near 1, then it is likely that π_0 is 1 and the reliability π_t remains at 1 for many periods, until the onset time, τ, of reliability degradation,

$$\tau \equiv \min\{t : r_t < 1, t = 1, 2, \ldots\}.$$

Thereafter, the reliability degrades at a geometric rate r_t until the next time, t', with $r_{t'} < 1$ and then degradation progresses more rapidly. The shock r_0 is the initial reliability; r_1 determines the initial decay rate ($r_1 = 1$ produces no initial decay) and subsequent shocks accelerate the decay. For our initial development, we treat the model parameters a_t, b_t, and p_t as given; in Section 4 we consider a hierarchical specification.

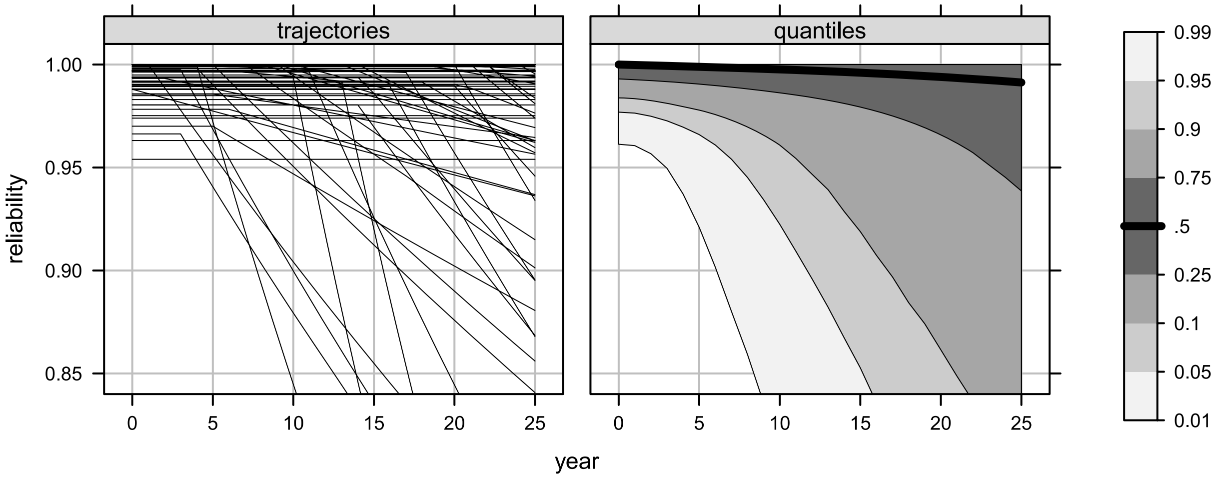

Figure 2. One hundred simulated reliability trajectories (left) and quantiles (right) of the reliability distribution from the random onset model.

Figure 2 shows 100 sample paths (left panel) of π_t over a 25-year period with the parameters set as $(a_0, b_0, p_0) = (99, 1, 0.5)$ and $(a_t, b_t, p_t) = (99, 1, 0.98)$ for $t \geq 1$. Also shown (right panel) are various quantiles of the trajectory distribution from 200,000 simulated paths. Each sample path in the left panel represents a possible trajectory for the probability that a random unit will function when tested or fielded.

In this case, with $p_t = 0.98$ $(t \geq 1)$, the mean time to onset of degradation is $1/(1 - 0.98) = 50$ years. With $a_t = 99$ and $b_t = 1$, once degradation begins, it will progress, on average, by $1/(99 + 1) = 1\%$ of the previous year's reliability until the next shock. The exception, $p_0 = 0.5$, means that initial reliability in year zero is 1.0 with probability 0.5 and otherwise has an expectation of $a_0/(a_0 + b_0) = 99\%$.

3. POSTERIOR ANALYSIS WITH KNOWN ONEBETA PARAMETERS WHEN ALL UNITS PASS TESTING

The parameters p_t, a_t, b_t $(t \geq 0)$ determine a prior distribution for the evolution of π_t, including the time of onset of degradation and its initial rate. RADAR allows for a variety of reliability degradation patterns including early or late onset and slow or fast progression. With parameters set as in Figure 2, most realizations maintain reliability above 99% for at least 10 years, but in a few cases, it begins to degrade rather steeply within that period. Fresh tests with no failures are required to provide confidence that reliability remains high.

Suppose that in each of years $t = 0, 1, 2, \ldots, T$, a random sample of n_t units is drawn from the stockpile for testing. If all units sampled in each year work successfully, what is the posterior of the system reliability?

It is most convenient to work with the posterior distribution of the shocks r_t and use this to determine the posterior of the reliabilities π_t through (1). Define, for $t, T = 0, 1, \ldots,$

$X_t =$ number of units that pass testing in year t,

$S_T = \{X_t = n_t : t = 0, \ldots, T\} =$ all successes through year T,

and

$$f_t(r) = p_t\delta(r - 1) + (1 - p_t)B(r, a_t, b_t),$$

where $\delta(x)$ is Dirac's delta function and $B(r, a, b) = r^{a-1}(1 - r)^{b-1}/B(a, b)$ is the Beta density function with $B(a, b)$ denoting the beta function, defined as

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)},$$

where

$$\Gamma(a) = \int_0^\infty t^{(a-1)} e^{-t}\, dt.$$

Thus, $f_t(r)$ is the prior probability "density" function of r_t, a mixture of a point mass at unity and a beta distribution. Then, given no failures through year T, the posterior of the r_t variables is

$$\begin{aligned} f(r_0, \ldots, r_T | S_T) &\propto \prod_{t=0}^{T} f_t(r_t)\pi_t^{n_t} = \prod_{t=0}^{T} f_t(r_t)(\rho_0 \cdots \rho_t)^{n_t} \\ &= f_0(r_0) r_0^{n_0} \prod_{t=1}^{T} f_t(r_t)(r_0 r_1^t \cdots r_t^1)^{n_t} \\ &= f_0(r_0) r_0^{N_{0:T}} \prod_{t=1}^{T} f_t(r_t) r_t^{N_{t:T}}, \end{aligned}$$

where $N_{0:T} \equiv \sum_{i=0}^{T} n_i$ and (for $t = 1, \ldots, T$) $N_{t:T} \equiv \sum_{i=t}^{T} (i - t + 1)n_i$. The final expression shows that the r_t are independent given S_T, with posterior densities proportional to

$$f_t(r_t) r_t^{N_{t:T}} = p_t\delta(r_t - 1) + (1 - p_t)B(r_t, a_t, b_t) r_t^{N_{t:T}}.$$

The normalizing constant is $p_t + (1 - p_t)B(a_t + N_{t:T}, b_t)/B(a_t, b_t)$. Therefore the explicit posterior pdf of $[r_t|S_T]$ is

$$f(r_t|S_T) = p'_{t,T}\delta(r_t - 1) + (1 - p'_{t,T})B(r_t, a_t + N_{t:T}, b_t),$$

where

$$p'_{t,T} \equiv \frac{p_t B(a_t, b_t)}{p_t B(a_t, b_t) + (1 - p_t)B(a_t + N_{t:T}, b_t)}.$$

In summary, the posterior distribution of the r_t variables $(t = 0, \ldots, T)$ is

$$[r_t|S_T] \overset{ind}{\sim} \text{OneBeta}(p'_{t,T}, a_t + N_{t:T}, b_t).$$

This provides a simple mechanism to simulate the posterior system reliabilities $\pi_t = r_0 \prod_{i=1}^{t} r_i^{t-i+1}$ given that all $n_0, \ldots, n_T$ units have tested successfully. For $t > T$, r_t is independent of S_T and therefore its posterior is identical to its prior. We will see later that the r_t are no longer independent if at least one unit fails.

It is clear that, under model 1, the reliability will eventually degrade to zero: $\pi_T \to 0$ almost surely as $T \to \infty$. But what is the limit of the posterior $[\pi_T|S_T]$? In other words, if testing continues for a long time with no failures, what does this imply about the current reliability? The following theorem provides a partial answer.

Theorem 1. Let π_T evolve according to the random onset model with constant parameters, $a_t = a$, $b_t = b$, $p_t = p$, and $n_t = n$ for $t = 0, 1, \ldots$. If $0 < p < 1$ and $b > 1/2$, then $[\pi_T|S_T]$ converges to a nondegenerate random variable almost surely. If $p = 0$ or $b \le 1/2$, then $[\pi_T|S_T] \to 0$ almost surely. If $p = 1$, then $[\pi_T|S_T] = 1$ almost surely.

Proof. See Appendix A.

It is interesting that if $\Pr(r_t = 1) = 0$ (i.e., if $p = 0$), then, even with a perfect test record, the posterior distribution on reliability will eventually converge to zero regardless of the size of the annual sample, n. That is, with $p = 0$, the prior so strongly favors degradation that it eventually overcomes the constant stream of success data (and a long stream of successes becomes increasingly unlikely). Asymptotic convergence to zero may not be problematic if the posterior distribution on reliability is reasonable throughout the design life of the system. Nevertheless, we recommend using priors with $p > 0$ and $b > 1/2$ for most applications because in this case a long string of successes can exactly counterbalance the downward pressure that the prior imposes on reliability and this possibility of equilibrium seems more appropriate than certain convergence to zero.

Appendix B in the supplemental materials (available on the *Technometrics* web site) gives quantitative examples of posterior analysis when all units pass and the OneBeta parameters are known.

4. POSTERIOR ANALYSIS WITH A PRIOR FOR p_t AND WHEN SOME UNITS MAY FAIL

In this section, we consider examples where failures may be observed. Instead of having a closed-form result, these analyses use Markov chain Monte Carlo (MCMC) to compute posterior estimates and intervals. Section 5 calculates posteriors for several interesting scenarios in which we obtain both data and additional knowledge about mechanisms affecting the stockpile.

Within the MCMC framework we can also admit uncertainty in the p_t parameters by giving them a prior distribution. A decreasing sequence p_t corresponds to a stockpile with increasing propensity for aging to begin affecting reliability. Although this effect can be achieved by fixing a known p_t sequence, we choose to introduce additional uncertainty by modeling p_t as a logit-linear sequence with two uncertain parameters and a high probability of decreasing with t.

4.1 A Baseline Scenario

As a baseline scenario for considering posterior analysis when units fail, we will always assume that the system has been surveilled for 10 years, with 11 samples per year, and that no failures have been observed. For these examples, the prior specification is changed slightly compared to the prior used in Section 2 and Appendix B (in the supplemental materials available on the *Technometrics* web site). We still have $(a_0, b_0, p_0) = (99, 1, 0.5)$ and $(a_t, b_t) = (99, 1)$ for $t \ge 1$. However, now p_t changes over time, with $\text{logit}(p_t) = c + dt$.

Let τ denote the onset time of reliability degradation. Suppose that we would also like to specify a distribution for c and d such that the marginal distribution of τ has 0.15 and 0.85 quantiles equal to 20 and 50 years, respectively. In addition, we would like to specify that the conditional probability $\Pr(\tau > 20|c, d) = [p_1 \cdots p_{20}|c, d]$ has mean 0.85 and standard deviation 0.025 and $\Pr(\tau > 50|c, d) = [p_1 \cdots p_{50}|c, d]$ has mean 0.15 and standard deviation 0.025.

These specifications are achieved with the following multivariate normal prior distribution on (c, d):

$$\begin{pmatrix} c \\ d \end{pmatrix} = \text{Normal}\left(\begin{pmatrix} 5.7 \\ -0.076 \end{pmatrix}, 10^{-3} \begin{pmatrix} 78 & -2.2 \\ -2.2 & 0.066 \end{pmatrix} \right).$$

With this choice d is almost certainly negative and thus p_t declines over time, meaning that the annual probability of degradation onset increases with age. The Gaussian form is used for convenience.

Under the prior used in Section 2 and Appendix B (in the supplemental materials available on the *Technometrics* web site), where $p_t = 0.98$ for $t > 0$, the 0.15 and 0.85 quantiles for the onset time of reliability degradation are 9 and 94 years. Thus, although the p_t sequence is now modeled as uncertain, the induced distribution on τ actually has less variation because negative values of d reflect increasing hazard for age-related reliability degradation as the stockpile becomes older.

Figure 3 plots the median reliability estimates (black line at 1.0) and uncertainty bounds (boundaries of shaded regions) obtained using this prior distribution and observing 10 years of 11 units with no failures. The results beyond 10 years are predictions made after observing the year-10 data. The computations

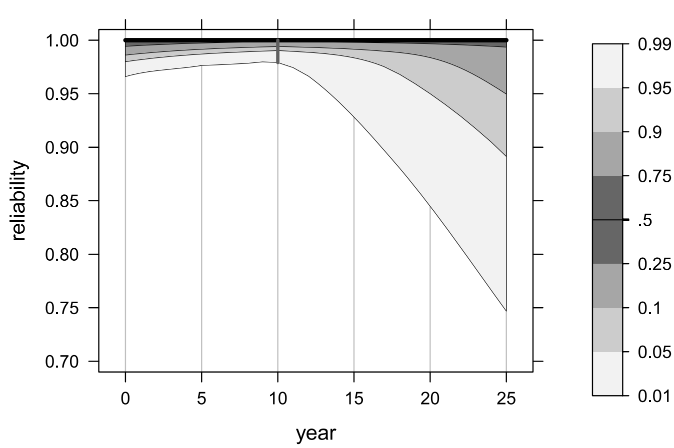

Figure 3. Reliability estimates when 11 units are surveilled each year with no failures for 10 years. Data collection stops after year 10 and the k-step-forecast posterior is plotted as explained in Section 4.3.

used to produce Figure 3 are discussed in the following two subsections. Figure 3 is similar to the right panel of Figure 1 except that curtailed sampling begins in year 11 instead of year 16 and the quantities p_t now follow the uncertain logistic pattern with respect to time. Figure 3 will be discussed further in Section 5 as we compare it to other scenarios.

4.2 Markov Chain Monte Carlo Computations

The posterior density function of the unknown parameter $\theta = (c, d, r_0, r_1, \ldots, r_T)$ given test results $X_0, X_1, \ldots, X_t$ is proportional to the product of the Binomial likelihood and the prior density for θ. X_0 corresponds to the number of passing units at the quality inspection immediately following manufacture. The index $T \geq t$ is taken to be the latest year for which analysis of the posterior distribution of reliability is desired; $T = 25$ for the analysis of Figure 3. The likelihood is

$$L(\theta|X_0, X_1, \ldots, X_t) \propto \prod_{i=0}^{t} \pi_i^{X_i}(1-\pi_i)^{n_i - X_i},$$

where each π_t is a function of θ through its dependence on $r_0, r_1, \ldots, r_t$. The prior distribution on θ is given by

$$[r_t|a_t, b_t, p_t] \overset{ind}{\sim} \text{OneBeta}(a_t, b_t, p_t)$$

(for $t = 0, 1, \ldots, T$) with a_t and b_t known and the distribution of p_t determined through bivariate normality of c and d along with $\text{logit}(p_t) = c + dt$ for $t > 0$. The parameter p_0 is fixed at 0.5.

To sample from the posterior distribution of θ we use a variable-at-a-time Metropolis–Hastings algorithm with reversible jump (Green 1995) steps used to sample the r_t's. Details are given in Appendix D in the supplemental materials (available on the *Technometrics* web site). The algorithm was implemented using YADAS (Graves 2003, 2007), a set of Java classes to help users develop MCMC algorithms in nontrivial problems. It includes support for reversible jump MCMC.

4.3 Time Evolution of the Posterior Distribution

A single run of the MCMC algorithm produces a sample from the posterior distribution $[\theta|X_0, \ldots, X_t]$, conditioned on data up to a given time t. Recall that θ includes innovations $r_0, \ldots, r_T$ in the analysis window $T \geq t$. Let $\pi_i(\theta)$ denote the reliability in year $i \in \{0, \ldots, T\}$, as determined by $r_0, \ldots, r_i$ and define the vector $\pi(\theta) = (\pi_0(\theta), \ldots, \pi_T(\theta))$. A single MCMC run produces an empirical distribution of $[\pi(\theta)|X_0, \ldots, X_t]$ in which reliabilities for all years $(0, \ldots, T)$ are conditioned on test results through a fixed year t.

Multiple MCMC runs, conditioning on progressively more data, produce a set of posterior distributions $\{[\pi(\theta)|X_0, \ldots, X_t]\}_{t=0}^{T}$. Figure 3 plots selected portions of these posterior distributions over an analysis window up to year $T = 25$. For years up to and including the vertical line segment at year 10, the shaded regions summarize what we call the *progressive* posteriors $\{[\pi_t(\theta)|X_0, \ldots, X_t]\}_{t=0}^{10}$. These are current estimates of reliability based on data up to each given year from 0 to 10. From year 11 forward the plot shows what we term the *k-step-forecast* posteriors based on test results through year 10, $\{[\pi_t(\theta)|X_0, \ldots, X_{10}]\}_{t=11}^{25}$. This format of plotting is used in the following section to demonstrate how current and future assessment of reliability changes as various kinds of new data are obtained.

5. FAILURE SCENARIOS

Recall the baseline scenario of 10 years with 11 units successfully surveilled, as illustrated in Figure 3, with the vertical line segment at year 10 demarcating the switch from progressive to k-step-forecast reliability estimates. With this as the starting point, we illustrate the evolution of the posterior when a unit fails testing and then as additional information is discovered about whether the failure resulted from a manufacturing defect or whether an aging mechanism has begun to reduce reliability. These examples are representative of a variety of scenarios that are of interest to policy-makers. In particular, interest focuses both on how reliability estimates change given a range of new information and on how the uncertainty changes. We consider observing a failure that is determined to come from a manufacturing defect. We then consider a failure concomitant with information that aging has begun to reduce reliability.

Suppose that following the baseline of 10 years with no trouble found, at year 11 a single failure is observed in 11 units tested. Figure 4 shows posterior summaries based on this information alone, while Figures 5 and 6 incorporate further information. Figure 4(a) plots reliability posteriors through year 25,

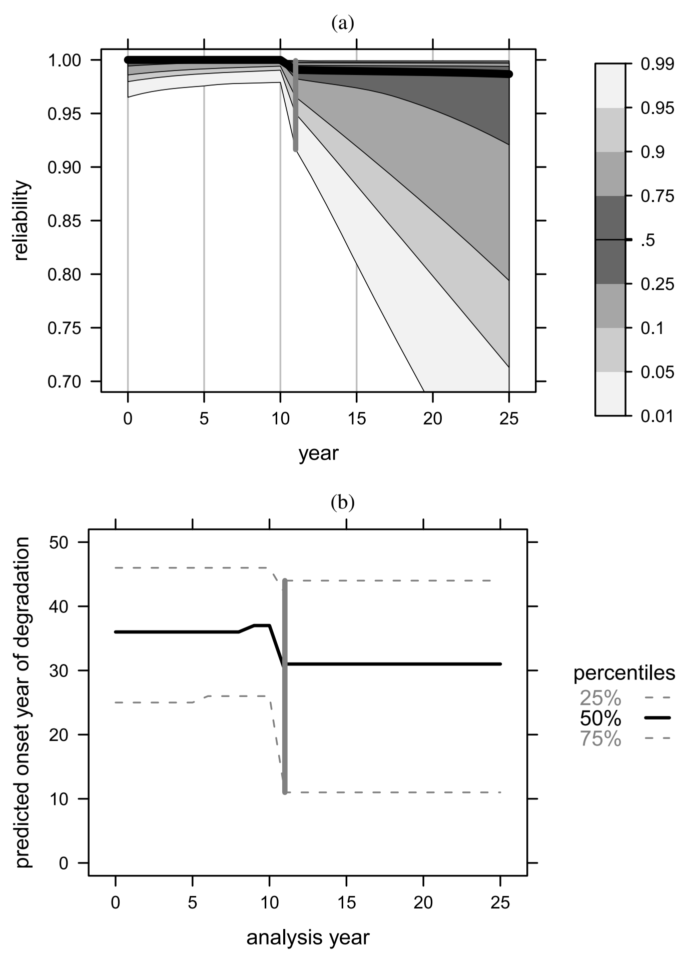

Figure 4. Estimates when 11 units are surveilled each year with no failures for 10 years and then one of 11 units fails in year 11: (a) Reliability. (b) Year of onset of degradation. For each year t up to year 11, posterior distributions are based on all data collected up to year t. Reliability estimates after year 11 are based on data only through year 11 (i.e., they could be computed at year 11). See Section 4.3.

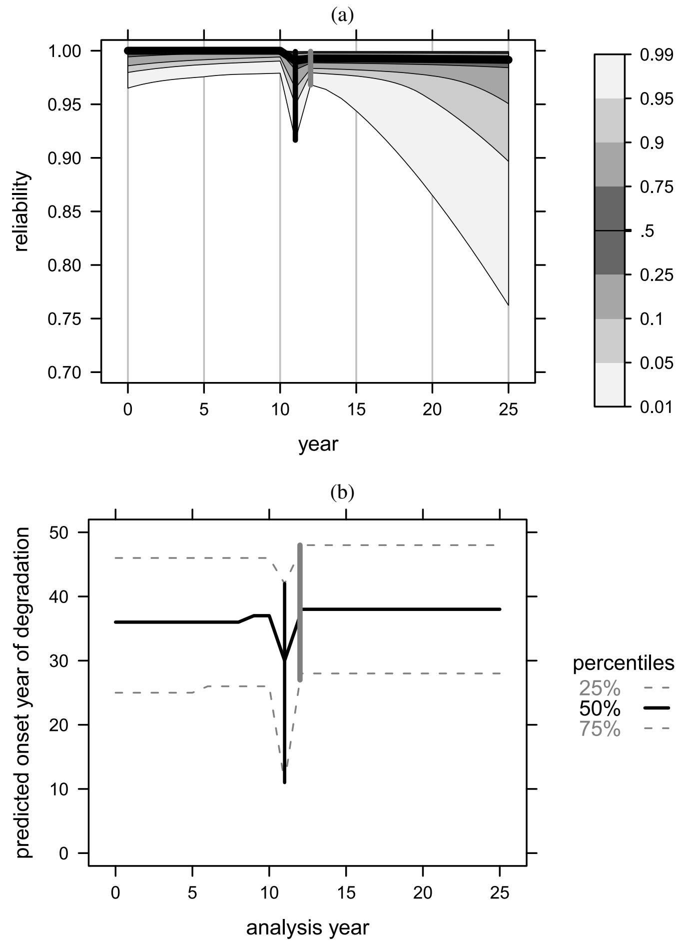

Figure 5. Estimates when 11 units are surveilled each year with no failures for 10 years and then one of 11 units fails in year 11 due to a manufacturing issue and 11 more units are tested in year 12 with no failures: (a) Reliability. (b) Year of onset of degradation.

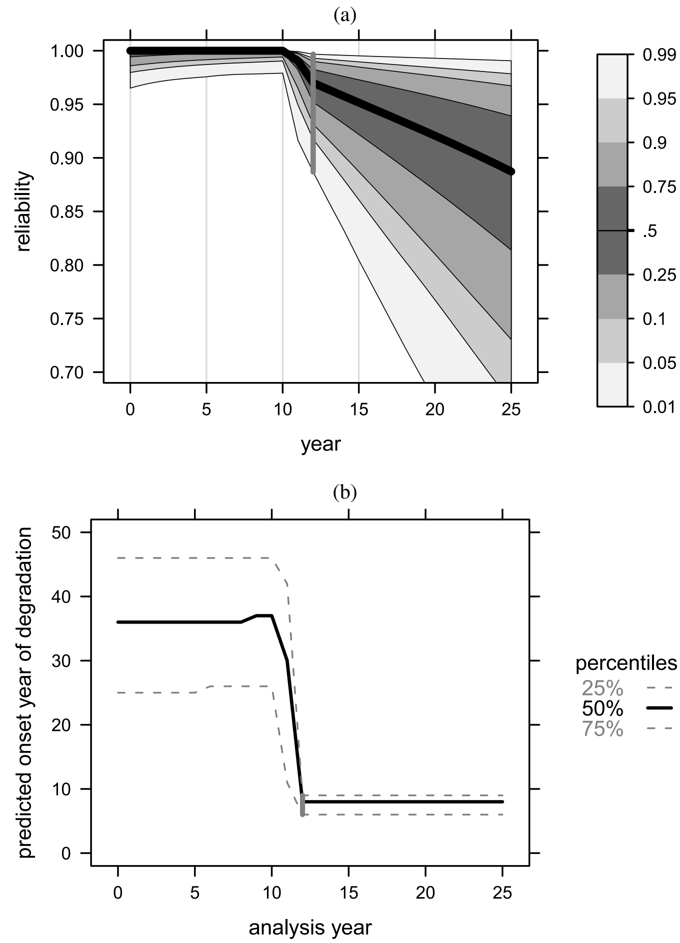

Figure 6. Estimates when 11 units are surveilled each year with no failures for 10 years and then one of 11 units fails in year 11, and it is established that aging has begun: (a) Reliability. (b) Year of onset of degradation.

and Figure 4(b) plots posteriors of the onset of degradation, τ. Notice that a single observed failure dramatically changes the lower tails of the distributions as compared to Figure 3, before the failure was observed. The median of the k-step-forecast posterior of reliability drops to about 0.99 but is nearly flat through year 25, consistent with age-related degradation of reliability remaining relatively unlikely, even though a defective unit has been observed. Note that 0.99 is approximately the median of the Beta$(99 + 131, 1 + 1)$ distribution that is the posterior for a simple conjugate beta-binomial analysis with prior distribution Beta$(99, 1)$ and data consisting of a single failure among 12 samples of 11. This demonstrates that the dynamic effects of aging in the random onset model have limited impact on the median of the k-step-forecast posterior distribution as compared to the simple standard beta-binomial analysis method. On the other hand, the lower tails degrade at rates between 1% and 3% per year, consistent with the possibility that the defective test result may be indicative of age-related reliability degradation, even if such degradation has not explicitly been demonstrated by understanding the root cause of the failure. The 25th percentile line in Figure 4(b) shows that, for example, at year 11 the probability that degradation has already begun is (about) 25%, while the 75th percentile line shows that there is about a 25% chance that degradation will not even have begun by year 45.

Next, suppose that further investigation of the observed failure reveals that it was a manufacturing issue and does not indicate the onset of age-related degradation. Formally, this new information is included in the posterior analysis as follows. The likelihood contribution for a successful test in year t is $L(\pi|\text{pass}) \propto \pi_t$, and the likelihood contribution for a failed test, *absent any attribution* of whether the failed unit was defective at the time of manufacture, is $L(\pi|\text{fail}) \propto 1 - \pi_t$. If, however, attribution information is available, then we regard this information as data and extend the Bernoulli model for the data as follows:

$$\begin{aligned}&\Pr(\text{unit passes test in year } t|\\&\qquad \pi, \text{unit was not defective at manufacture}) = \pi_t/\pi_0,\\&\Pr(\text{unit passes test in year } t|\\&\qquad \pi, \text{unit was defective at manufacture}) = 0.\end{aligned}$$

This extended model is consistent with the original marginal model for the pass-fail data and implies that a failure in year t that is known to be defective at manufacture has likelihood contribution $L(\pi|\text{fail, defective at manufacture}) \propto 1 - \pi_0$, just as if the unit had been tested at year zero instead of year t.

In Figure 5(a) the single failure in year 11 is determined in year 12 to be caused from a manufacturing defect and not from age-related degradation. In addition, 11 more units are tested in year 12 with no failures. The progressive posterior distribution of reliability in year 12 is much improved from the previous year because the failed unit no longer heightens concern that progressive reliability degradation has already begun. This is also reflected in Figure 5(b) showing the posterior distribution for predicted year of degradation onset. Notice that the lower tail in year 12 and beyond is improved from the assessment in year 11.

Compare Figure 5 to Figure 3. While the updated predictions do acknowledge that the defect was present at manufacture, there is no evidence that the onset of degradation is sooner than expected, and the lower tails of the k-step-forecast posteriors behave quite comparably following the vertical line segments at the current year in each figure.

Now suppose that instead of discovering that the single failure in year 11 was caused by a manufacturing defect, the follow-up investigation during year 11 reveals that a degradation mechanism impacting reliability began at year 10 or before. This new information is incorporated into the analysis in Figure 6, beginning in year 12, the current year as indicated by the vertical line segments. The portions of Figure 6 from year 12 forward are obtained by subsetting the MCMC results to include only those Markov chain iterations in which at least one of $r_1, \ldots, r_{10}$ is less than 1 so that the subset reflects the conditional distribution of θ given that aging begins by year 10.

Figure 6(a) plots the reliability through year 25, and Figure 6(b) plots the predicted year of the onset of degradation. Knowledge that degradation has begun to impact reliability dramatically reduces the median, upper, and lower tails of both plots. All quantiles of the k-step-forecast posterior distribution of reliability trend downward, in contrast to Figures 3–5 where the median and upper quantiles are relatively constant.

Figures 3–6 illustrate various ways that the distribution of reliability might evolve as various pieces of information become available over time: first a failure and then information that the failed unit was defective at manufacture or information that reliability of the population is degrading. When used in this way, the model is a flexible tool to plan surveillance sampling under various contingencies. We have used similar plots to show the impacts of reduced sampling rates, obtaining information that age-related reliability issues are likely to occur sooner than originally expected (altering the prior assumed for c and d), and discovering that a known aging issue impacts only a small fraction of the population, rather than affecting all units. Each of these scenarios addresses a contingency that engineers need to consider in their plans for future stockpile surveillance.

6. DISCUSSION AND POSSIBLE EXTENSIONS

We have described a random onset model for stockpile reliability and shown how "what-if" scenarios can be used to demonstrate how confidence is affected by changes in sampling rate, failures in testing, and updated information about whether reliability degradation is known to be under way. For our problem, we have assumed that sampling is done annually; clearly one could also consider varying the sampling intervals as well as the sample size. We argue that this simple model captures the main features of uncertainty in stockpile reliability and should be used to answer broad *policy* questions about the surveillance programs.

Specifying appropriate prior information is an important part of using the RADAR model. Obtaining consensus on parameter values, as discussed for two applications in Section 1.1, is time-consuming and may not even be possible, as different experts have different opinions about future risks. To some extent, it is impossible to defend any prior distribution of reliability trajectories as an adequate description of the possible aging issues that no person has yet conceived and yet could impact a stockpile in the future. There is, nevertheless, an uncomfortable necessity to acknowledge the possibility of unknown unknowns and do one's best to measure uncertainty associated with the risk of not discovering aging issues because of limited surveillance over time. Although RADAR results for different choices of prior parameters and sampling rates are *qualitatively* similar, decisions are often made using quantitative results, and these depend heavily on specific parameter values. The model provides a framework for planning and policy, but is not a substitute for more detailed, physically based, reliability analyses when knowledge of specific physical aging processes is available.

It is clear that ongoing sampling and testing are imperative to maintain confidence that a stockpile has high reliability. The random onset model described herein provides a means for quantitatively assessing confidence in stockpile reliability for the purpose of planning surveillance programs. In addition, this framework is flexible enough to allow for extensions to more complex situations—including, perhaps, remedial actions and more complex types of data.

APPENDIX A: PROOF OF THEOREM 1

The reliability in year T can be expanded as

$$\pi_T = r_0 \prod_{t=1}^{T} r_t^{T-t+1} \quad \Longrightarrow$$

$$\log(\pi_T) = \log(r_0) + \sum_{t=1}^{T} (T-t+1)\log(r_t),$$

and therefore

$$[-\log(\pi_T)|S_T] \sim Y_{0,T} + \sum_{t=1}^{T} (T-t+1) Y_{t,T},$$

where

$$Y_{t,T} \overset{ind}{\sim} -\log(\text{OneBeta}(p'_{t,T}, a_t + N_{t:T}, b_t))$$

for $t = 0, 1, \ldots$.

We study convergence based on the two series

$$\sum_{t=1}^{T} (T-t+1)\mathbf{E}Y_{t,T} \qquad \text{and} \qquad \sum_{t=1}^{T} (T-t+1)^2 \operatorname{Var} Y_{t,T}.$$

Lemma 1 (lemmas are collected in Appendix C of the supplemental materials available on the *Technometrics* web site) implies that if both the expectation series and the variance series are bounded as $T \to \infty$, then $[\pi_T | S_T]$ converges a.s. to a nondegenerate random variable. Furthermore, if the expectation series is infinite and the variance series is bounded, then $[\pi_T | S_T]$ converges to 0 a.s.

Although the theorem is stated for the case of constant parameters, the following proof can easily be extended to a case with parameters constant only for large t. With constant parameters, $N_{t:T} = n(T-t+1)(T-t+2)/2$ (for $t = 1, \ldots, T$) and we set $k = T - t + 1$ and write

$$N_k \equiv nk(k+1)/2 \qquad \text{and}$$

$$p'_k \equiv \frac{pB(a,b)}{pB(a,b) + (1-p)B(a+N_k, b)}$$

for $k = 1, \ldots, T$.

Case $p = 0$: Lemma 2 gives $\mathbf{E}Y_{t:T} = g_b(a + N_{t:T})$ where $g_b(x) = \psi(x+b) - \psi(x)$ and $\psi(\cdot)$ is the digamma function. Thus, the mean series is

$$\begin{aligned}\sum_{t=1}^{T}(T-t+1)\mathbf{E}Y_{t,T} &= \sum_{t=1}^{T}(T-t+1)g_b(a+N_{t:T}) \\ &= \sum_{k=1}^{T} k g_b(a+N_k) \\ &= \sum_{k=1}^{T}\frac{k}{N_k}[N_k g_b(a+N_k)]\end{aligned}$$

and this becomes infinite as $T \to \infty$ because the final bracketed factor converges to $b > 0$ by Lemma 3, and the remaining factor, k/N_k, is asymptotic to $1/k$, which diverges as a series.

Lemma 2 also gives $\operatorname{Var} Y_{t,T} = h_b(a + N_{t:T})$ where $h_b(x) = \psi_1(x) - \psi_1(x+b)$ and $\psi_1(\cdot)$ is the trigamma function. Thus, the variance series is

$$\begin{aligned}\sum_{t=1}^{T}(T-t+1)^2 \operatorname{Var} Y_{t,T} &= \sum_{t=1}^{T}(T-t+1)^2 h_b(a+N_{t:T}) \\ &= \sum_{k=1}^{T} k^2 h_b(a+N_k) \\ &= \sum_{k=1}^{T}\frac{k^2}{N_k^2}[N_k^2 h_b(a+N_k)]\end{aligned}$$

and this converges as $T \to \infty$ because the final bracketed factor is bounded by Lemma 3, and the remaining factor, k^2/N_k^2, is asymptotic to $1/k^2$, which converges as a series.

Thus Lemma 1 implies that $[-\log(\pi_T)|S_T] \to \infty$ a.s. and therefore, $[\pi_T|S_T] \to 0$ a.s.

Case $0 < p < 1$: Combining the definition of $Y_{t,T}$ with Lemma 2 gives $\mathbf{E}Y_{t,T} = (1 - p'_{t,T})g_b(a + N_{t:T})$ and it is straightforward to show that $(1 - p'_k) \le (1-p)B(a+N_k, b)/(pB(a,b))$. Therefore, the mean series is

$$\begin{aligned}\sum_{t=1}^{T}(T-t+1)\mathbf{E}Y_{t,T} &= \sum_{t=1}^{T}(T-t+1)(1-p'_{t,T})g_b(a+N_{t:T}) \\ &= \sum_{k=1}^{T} k(1-p'_k)g_b(a+N_k) \\ &= \sum_{k=1}^{T}\frac{k}{N_k}(1-p'_k)(N_k g_b(a+N_k)) \\ &\le \frac{1-p}{pB(a,b)}\sum_{k=1}^{T}\left[\frac{B(a+N_k,b)}{N_k^{1/2}}\right] \\ &\quad \times \left[\left(\frac{2k}{k+1}\right)^{1/2} N_k g_b(a+N_k)\right].\end{aligned}$$

The final bracketed factor converges to $b\sqrt{2} > 0$ by Lemma 3 and thus the series converges if and only if $\sum n^{-1/2}B(a+n,b)$ converges, which is equivalent to $b > 1/2$ by Lemma 4.

For the variance series, the definition of $Y_{t,T}$ and Lemma 2 give

$$\begin{aligned}\operatorname{Var} Y_{t,T} &= \mathbf{E}Y_{t,T}^2 - (\mathbf{E}Y_{t,T})^2 \\ &= (1-p'_{t,T})[h_b(a+N_{t:T}) + p'_{t,T}g_b^2(a+N_{t:T})]\end{aligned}$$

and thus

$$\begin{aligned}&\sum_{t=1}^{T}(T-t+1)^2 \operatorname{Var} Y_{t,T} \\ &\quad = \sum_{k=1}^{T} k^2(1-p'_k)[h_b(a+N_k) + p'_k g_b^2(a+N_k)] \\ &\quad = \sum_{k=1}^{T}\frac{k^2}{N_k^2}\big[(1-p'_k)N_k^2(h_b(a+N_k) + g_b^2(a+N_k))\big].\end{aligned}$$

The series converges because the final bracketed factor is bounded by Lemma 3 and the remaining factor, k^2/N_k^2, is asymptotic to $1/k^2$, which converges as a series.

Lemma 1 now implies that the posterior reliability $[\pi_T|S_T]$ converges a.s. to a nondegenerate distribution if $b > 1/2$ and to 0 otherwise.

The case $p = 1$ is trivial.

SUPPLEMENTARY MATERIALS

Examples, lemmas and details: Appendix B: Additional examples with no failures. Appendix C: Lemmas used in proof of Theorem 1. Appendix D: Details of Markov chain Monte Carlo algorithm. (supplemental.final.pdf)

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and three referees for many constructive comments and suggestions that significantly improved the quality of the article. The authors also thank Cheryll Faust, Mike Hamada, Charles Hills, Harry Martz, and William Meeker for helpful discussions.

[Received July 2009. Revised September 2010.]

REFERENCES

Derman, C., and Solomon, H. (1958), "Development and Evaluation of Surveillance Sampling Plans," *Management Science*, 5 (1), 72–88. [163]

General Accounting Office (1996), "Report to the Committee on Armed Services, U.S. Senate. Nuclear Weapons: Improvements Needed to DOE's Nuclear Weapons Stockpile Surveillance Program," Report GAO/RCED-96-216, Washington, DC. [165]

Graves, T. (2003), "An Introduction to YADAS," available at *http://yadas.lanl.gov*. [168]

——— (2007), "Design Ideas for Markov Chain Monte Carlo Software," *Journal of Computational and Graphical Statistics*, 16 (1), 24–43. [168]

Green, P. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82 (4), 711–732. [168]

Hillier, F. (1962), "Surveillance Programs for Lots in Storage," *Technometrics*, 4 (4), 515–530. [163]

Valdez-Flores, C., and Feldman, R. (1989), "A Survey of Preventative Maintenance Models for Stochastically Deteriorating Single-Unit Systems," *Naval Research Logistics*, 36 (4), 419–446. [163]

Wilson, A., Hamada, M., Anderson-Cook, C., Graves, T., Vander Wiel, S., Sims, B., and Wilson, G. (2007), "Statistical Models for Stockpile Health Assessment," in *ADTSC Nuclear Weapons Highlights 2007*, Los Alamos National Laboratory, LA-LP-041, available at *http://www.lanl.gov/source/orgs/adtsc/nw_highlights_2007.php*. [165]