Abstract

We use a Bayesian hierarchical model to assess the reliability of the Joint Light Tactical Vehicle (JLTV), which is a family of vehicles. The proposed model effectively combines information across three phases of testing and across common vehicle components. The analysis yields estimates of failure rates for specific failure modes and vehicles as well as an overall estimate of the failure rate for the family of vehicles. We are also able to obtain estimates of how well vehicle modifications between test phases improve failure rates. In addition to using all data to improve on current assessments of reliability and reliability growth, we illustrate how to leverage the information learned from the three phases to determine appropriate specifications for subsequent testing that will demonstrate if the reliability meets a given reliability threshold.

1 Introduction

Reliability is a high priority in the testing of military systems. A common interest is *reliability growth*, which tracks the change, and ideally improvement, in the reliability of a system as it moves through testing phases and undergoes periodic corrective actions. Ideally, plans for future test events should be based on what is observed in previous testing.

In this paper, we look at combining information across multiple test phases with the objective of planning a future test. When combining information, there is no omnibus solution. Rather, models need to be carefully considered and evaluated to ensure that they accurately reflect the data and the underlying physical processes. Our analysis is motivated by reliability data from the Joint Light Tactical Vehicle (JLTV), which is a family of vehicles designed to replace one-third of the legacy high-mobility multipurpose wheeled vehicle (Humvee) fleet. There are four unique types of vehicles, and within types there are many possible variants. The primary mission of the family of vehicles is to provide ground mobility that is deployable worldwide and capable of operating across the range of military roles including combat, sustainment, police action, peacekeeping, and security patrol, in all weather and terrain conditions.

The JLTV has been through a series of test events as it has been developed. Engineering and Manufacturing Development (EMD) included three phases of testing, where fixes occurred only during a set Corrective Action Period (CAP) between each phase. For every vehicle, each failure encountered during testing was recorded and attributed to a specific failure mode. Failures discovered during mission execution that result in an abort or termination of a mission in progress are scored as Operational Mission Failures (OMF), and failures of mission essential components are scored as Essential Function Failures (EFF). The EFFs tend to include a large portion of the failure modes that drive maintenance costs and reduce system availability. While requirements are typically written in terms of OMFs, because OMFs are by definition EFFs that occurred at critical times, combining the failures provides a more robust reliability estimate. Since fixes were delayed to CAPs, analysis generally shows a distinct jump in the system reliability between test phases.

We propose a model that can effectively combine information across the three phases of testing and across common vehicle components. We use a Bayesian hierarchical model

to assess the reliability of the family of vehicles. The model accounts for the commonality across vehicle types and allows for uncertainty quantification. Inferential objectives from the proposed model include an estimate of the mean miles between failure (MMBF) for each vehicle, failure mode, and phase of test, and an assessment of the effectiveness of the corrections completed in the CAPs. Additionally, because future tests are being planned, we propose methodology to leverage the information learned from the three phases to determine the length of testing needed to demonstrate a given reliability threshold in subsequent testing.

2 Data

Eight vehicle prototypes were used for reliability, availability, and maintainability (RAM) testing. There were four two-seat utility vehicles, two four-seat general purpose vehicles, a two-seat heavy guns carrier, and a two-seat close combat weapons carrier. The close combat weapons carrier, one general purpose, and two utility vehicles were tested in Aberdeen, Maryland, where it is much colder in the winter months but the terrain is not as harsh, and the heavy weapons carrier, one general purpose, and two utility vehicles were tested in Yuma, Arizona, where it is warmer but with more challenging terrain, to attain testing in different environmental and terrain conditions. The vehicles were driven roughly 3000 to about 5000 miles in Phase 1, between 4500 and 7500 miles in Phase 2, and between 5000 and 6700 miles in Phase 3.[1]

There are many similarities between vehicle types; in fact, the most dissimilar vehicles have over 80% common parts. Consequently, some failure modes, such as brakes and radios, will be common to all eight vehicles. Other failure modes will be related, but not identical, as with hydraulics, frame, and body.

There are 91 OMFs and 1321 EFFs attributed to 26 failure modes recorded across all eight vehicles and all three test phases. In any given phase of test, every vehicle is not guaranteed to have a failure of all 26 failure modes, and each test phase does not necessarily end with a failure. Therefore, we have 624 right censored observations, accounting for miles driven without observed failures.

3 Methodology

3.1 Modeling Reliability

A standard reliability analysis employed by the Department of Defense (DoD) test community considers each test phase independently and uses the exponential distribution to model the miles between failure [1]. The traditional analysis is overly simplistic, relies on correct modeling assumptions and ignores valuable information learned about the individual vehicles and their failure modes.

[1]The exact mileage has been altered to protect proprietary information.

In this paper we propose an alternative approach using a Bayesian hierarchical model. We will begin by introducing a hierarchical structure that models the relationships in the data across all test phases and incorporate known similarities between vehicle failure modes. Next, we will consider different modeling and distributional assumptions and their implications. We will aslo discuss model diagnostics to consider when choosing a final model. Then we will use the JLTV data to illustrate this process and discuss system reliability results. This will lead us into the next section on using these results to improve the planning of furture test.

3.1.1 Model Structure

In the first phase of testing, we assume the vehicle miles at failure, y, follow a lifetime distribution Y with an unknown set of parameters that include a rate parameter λ_{ij}. Introducing notation,

$$y_{ijk} \mid \lambda_{ij} \sim Y(\lambda_{ij}, ...), \quad i = 1, 2, ..., v \quad j = 1, 2, ..., s \quad k = 1, 2, ..., n_{ij} \tag{1}$$

where y_{ijk} are the miles between failure for vehicle i failure mode j, v is the number of vehicles, s is the number of failure modes, and n_{ij} are the number of failures of vehicle i failure mode j. The number of failure modes is assumed fixed and known *a priori*.

The prior distribution on failure rate parameter, λ_{ij}, depends on whether failure mode j is considered to be common across vehicles or related but not identical. For the related failure modes, we place a prior distribution on the collection of λ_{ij}; in other words, we assume each vehicle has a distinct failure rate in failure mode j but they arise from a common distribution.

If failure mode j is considered common across vehicles, the collection of failure rates is collapsed to a single parameter, $\lambda_{ij} = \lambda_j$. As with the related failure modes, a prior distribution is placed on the single failure rate. The prior distributions are independent across failure mode, and can potentially have different hyperparameters.

The Phase 1 analysis yields an estimate of the failure rate λ_{ij} for each of the vehicles for failure modes that are related. We are assuming the vehicles are conditionally independent, therefore the failure rate estimate for the family of vehicles for such failure modes can be found by $\sum_i \lambda_{ij}$. For failure modes that are common across vehicles the Phase 1 analysis yields a λ_j, which is the failure rate for the family of vehicles. If we assume the miles at failure come from an exponential distribution the overall failure rate across all failure modes can be easily calculated by $\sum_j \lambda_j$. The exponential distribution is commonly used in reliability analysis because of its convenience, but as we will discuss in later sections it can be unrealistic in many real-world situations.

3.1.2 Fix Effectiveness

After the first CAP, Test Phase 2 begins with the repaired vehicles. To capture these revisions, the PM2 reliability growth model [7] is often used. This model explicitly captures testing phases, choices about which failure modes to correct, and the potential of not completely eliminating a failure upon repair. One of the downsides of PM2 is that many

parameters of potential interest, such as the Fix Effectiveness Factor (FEF), which measures how much repairs improve failure rates, are typically fixed. A common value for FEF is 0.70. We follow the premise of this type of model, but allow a more flexible and data-driven result that is less dependent on hard-coded assumptions.

One normal assumption used in reliability growth modeling is nondecreasing failure rates; that is either the fixes were effective or had no effect, but did not degrade the family of vehicles. This should generally be the case, but because we are dealing with complex systems and testing conditions are rarely completely consistent, we will sometimes see decreases in failure rates after adjustments are made. Therefore for the Phase 2 data, we write the rate parameters as a function of the rate parameters found in Phase 1. In particular, we define $\lambda_{ij}^{P2} = (\rho_j)\lambda_{ij}^{P1}$ where ρ_j represents the between phase change in failure mode j. Given this definition of λ_{ij}^{P2}, we again model the miles to failure for a given vehicle and failure mode using the same lifetime distribution from Test Phase 1. We will choose a prior distribution for the parameter ρ_j that has positive support. If ρ_j is less than one, this represents an improvement in reliability. After Phase 2 we can look again at failure rates across failure modes and vehicles and obtain an overall estimate of the rate for the family of vehicles. The analysis of Test Phase 3 follows the same pattern as that shown in Phase 2. At the end of Phase 3, we can look at failure rates across failure modes and vehicles and obtain an overall estimate of the rate for the family of vehicles. Future tests will be planned based on the inferences of Phase 3.

3.1.3 Distributional Assumption

The next modeling choice that must be made is slecting an appropriate probability model This includes a slecting a sampling distribution for the data and prior distributions for each parameter on which the sampling distribution depends. For lifetime data there are a number sampling distribution to choose from. Some common choices are exponential, Weibull, lognormal, gamma, inverse Gaussian, and normal failure time models. Because all of these distributions have rate parameter (sometimes called scale = 1/rate) they can be applied to the proposed modeling structure. For this paper we will focus on the exponential and Weibull distributions.

First the exponential model, Test Phase 1,

$$f(y_{ijk}|\lambda_{ij}) = \lambda_{ij} \exp(-\lambda_{ij} y_{ijk}) \tag{2}$$

Test Phase 2,

$$f(y_{ijk}|\lambda_{ij}\rho_{1,j}) = \lambda_{ij}\rho_{1,j} \exp(-\lambda_{ij}\rho_{1,j} y_{ijk}) \tag{3}$$

Test Phase 3,

$$f(y_{ijk}|\lambda_{ij}\rho 1_j \rho 2_j) = \lambda_{ij}\rho 1_j \rho 2_j \exp(-\lambda_{ij}\rho 1_j \rho 2_j y_{ijk}) \tag{4}$$

This is by far the most common parametric distribution used in reliability modeling because of its desirable math mathematical properties and simple interpretations. One of these properties is that the failure rate is constant. For the JLTV setting this would assume that a given component is equally likely to fail in the first mile of testing as it is in the

thousandth. This may be reasonable for some reliability settings, but in general we expect vehicle components to wear down and become more likely to fail the longer they are driven.

Despite its common uses the assumption of a constant failure rate over time is rarely justifiable. It has been well documented (Statistics, Testing, and Defense Acquisition: Background Papers chapter - 2 http://www.nap.edu/catalog/9655.html) the issues that can arise when this assumption is violated. We will now consider the same hierarchical model structure while using the Weibull distribution. Now we assume that the failure times y_{ijk} each follow a $Weibull(\lambda, \gamma)$ distribution with scale parameter λ and shape parameter γ, with probability density function for Test Phase 1,

$$f(y_{ijk}|\lambda_{ij}, \gamma_i) = \lambda_{ij}\gamma_i y_{ijk}^{\gamma_i - 1} \exp(-\lambda_{ij} y_{ijk}^{\gamma_i}) \tag{5}$$

The Weibull distribution is a more flexible model with both a rate parameter λ_{ij} and a shape parameter γ_j. The exponential is a special case of the Weibull, when $\gamma = 1$.

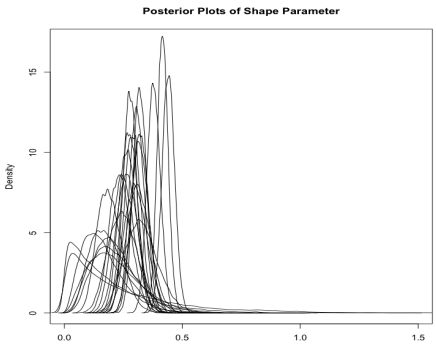
It should be noted that we are currently indexing both ρ and γ, only by j and not i, in other words we are assuming a single shape and between phase adjustment parameter for each failure mode across vehicles. All of these parameters could also be indexed by i and modeled hierarchical or with some combination of both. This is something we will omit for this paper but hope to explore in the future. One special case we will discuss is when a single Weibull shape parameter can be assumed across all failure modes. When this is justifiable it can greatly simplify test planning calculations, as we will show in the assurance testing section.

3.1.4 Model Diagnostics

So far we have discussed a number of different models and the assumptions that accompany them. We will now discuss the process of choosing a final model. The decision of which model to use can have a drastic impact on reliability assessment and future test planning. We will use the JLTV dataset to demonstrate the process.

The first model selection question we will explore is the parametric form, exponential versus Weibull. In statistical modeling we often face the trade-off between fit and interpretability, and this situation is no different. Because the Weibull is a more flexible model, it will always, in a sense, fit the data better, but this comes at a price. The exponential's convenient form makes both computation and interpretation straightforward. When the Weibull's shape parameter is introduced this advantage is lost. Thus, when the overall fit is close to the same between the exponential and the Weibull we will default to using the exponential.

The first check we used to decide between the exponential and Weibull models is to fit the Weibull model and look at the posterior distributions of the shape parameters and determine if one is a reasonable value. Looking at the plot below from the JLTV data, we see that for the 26 components it appears that the value of one falls in the extreme tails of the distributions. This is our first clue that the exponential model will not be a good fit to this data.



This is not a surprising result, because the exponential model assumes constant failure rate. For most vehicle components we would expect the failure rate to increase as the distance driven grows larger, and this corresponds with a shape parameter between zero and one. On the other hand, when the shape parameter is larger than 1, the failure rate will decrease with time/distance.

Now a more formal diagnostic tool used for model selection is the Deviance information criterion (DIC). This is a popular method for comparing the goodness of fit of multiple models. The DIC method gives a slight penalty for larger numbers of parameters. The DIC value is a unitless measure with lower values indicating a better fit to the data. In the table below we show the DIC results for the different distribution and structure combinations we considered with the JLTV dataset.

Goodness of Fit

Distribution	Structure	DIC
Exponential	Single Rate	23258
	Hierarchical Rate	23022
Weibull	Single Rate	18750
	Hierarchical Rate	**18556**
	Hierarchical Rate (One shape)	18677

The single rate structure represents the non-hierarchical model, using one rate parameter all 8 vehicles for a given failure mode. The hierarchical rate structure uses the common gamma distribution for the failure modes that are common across vehicles. The final entry in the table is the model that uses one shape parameter across all 26 failure modes.

While we have shown that the hierarchical Weibull model fits much better than the exponential, these test still do not tell us that this model is a good fit for the data. The last method we will present is called posterior predictive checking ***reference***. Here we are given important features of the data. In the JLTV case we were interested in the total failure counts for each phase and the number of time the miles between failures was less than 140. We then used the final model to simulate 5,000 new datasets. We then plot histograms for each of the 8 vehicles for all 3 phases. In Figure 1 are examples of the histograms produced, with the line showing where the value of the true dataset fell. For this method we dont expect all of the true values to fall in the center of the distribution. Values in the tales are to be expected in a random processes like this one. We will only be concerned if we find true values falling in the extreme tails or if we see a common bias to the high or low side of the

(a) Plot 1 (b) Plot 2

Figure 1: Posterior Predictive Plots

distributions.

3.1.5 JLTV Reliability Results

3.2 Assurance Test Planning

In this section inferences learned from the first three phases of testing will be used to develop a test plan for the next phase. Here the objective is to demonstrate that at a desired level of confidence, the system will meet or exceed a specified requirement. The methods to be employed will be similar to the assurance testing as discussed by Hamada et al. [2]. Bayesian assurance tests are used to insure that the reliability of an item meets or exceeds a specified requirement with a desired probability. Although practitioners often use "assure" and "demonstrate" synonymously, Meeker and Escobar [3] distinguish between reliability demonstration and reliability assurance testing. A *reliability demonstration test* is essentially a classical hypothesis test, which uses only the data from the test to assess whether the reliability-related quantity of interest meets or exceeds the requirement. A *reliability assurance test*, however, uses supplementary data and information in order to testing as efficient as possible.

In both classical and Bayesian test planning scenario we are interested in controlling error rates while minimizing the resources required for testing. In the DoD acquisition process the error rates will be referenced in terms of risk. *Consumer risk* which considers the event of purchasing a product that does not meet reliability requirements. *Producer risk* which considers the event of a product with acceptable reliability failing a given test and not being considered for purchase. Introducing the notation for this section, In the JLTV case study we will be determining how many miles on test T, are required and how many system failures to allow c, before the product is considered unacceptable. We will define $W(t)$ as a random variable that represents the total number of system failures after t miles. Suppose that π

denotes some quantity of interest that is related to system reliability at a given time. It is common to base both classical and Bayesian test plans on two specified levels of π: π_0, an *acceptable reliability level* (ARL), and π_1, a *rejectable reliability level* (RRL), where $\pi_1 \leq \pi_0$. Although the precise definition of ARL and RRL differ between the classical and Bayesian test criteria, we use them in an equivalent way.

3.2.1 Classical Approach

It is quite common to use two criteria in determining classical test plans. The *producer's risk* is the probability of failing the test when $\pi = \pi_0$, whereas the *consumer's risk* is the probability of passing the test when $\pi = \pi_1$. Suppose that we specify a maximum value, α, of the producer's risk and a maximum value, β, of the consumer's risk. For test planning, these criteria become

$$\begin{aligned} \text{Producer's Risk} &= P(\text{Test Is Failed } |\pi_0) \\ &= P(W(t) > c|\pi_0) \leq \alpha \end{aligned}$$

and

$$\begin{aligned} \text{Consumer's Risk} &= P(\text{Test Is Passed } |\pi_1) \\ &= P(W(t) \leq c|\pi_1) \leq \beta \end{aligned}$$

To choose a test plan for specified values of $(\alpha, \pi_0, \beta, \pi_1)$, we assume a distributional form that defines the relationship between the number of system failyres $W(t)$ and the reliability π. Then we can simply find the combinations of these probabilities by simultaneously solving these two equations. In the case where the component failure times are assumed to be exponentially distributed with rate parameter λ_i the failure counts for each component come from a homogeneous Poisson process $W_i(t) \sim \text{Poisson}(\lambda_i t)$, and the system failure counts $W(t) \sim \text{Poisson}(\lambda_S t)$ Numerous textbooks provide additional details of this purely classical approach.

3.2.2 Bayesian Approach

Traditional methods for determining the testing proceedure only rely on distributional assumptions and/or asymptoic results. With the Bayesian approach we incorperate the supplementary data from the previous testing phases with the hopes of minimizing the resorces needed for testing.

We now consider fully Bayesian posterior risks that convey a completely different outlook from the corresponding classical risks. While the classical provide assurance that satisfactory devices will pass the test and that unsatisfactory devices will fail it, posterior risks provide precisely the assurance that practitioners often desire: if the test is passed, then the consumer desires a maximum probability β that $\pi \leq \pi_1$. On the other hand, if the test is failed, then the producer desires a maximum probability α that $\pi \geq \pi_0$. Unlike the average risks, these posterior risks are fully Bayesian in the sense that they are subjective probability statements about π.

For a test that fails, the *posterior producer's risk* is the probability that $\pi \geq \pi_0$, or $P(\pi \geq \pi_0|\text{Test Is Failed})$. Notice that this is simply the posterior probability that $\pi \geq \pi_0$ given that we have observed more than c failures. In the exponential case, if we let π be the system failure rate λ_S, then using Bayes' Theorem, and assuming a maximum allowable posterior producer's risk α, an expression for the posterior producer's risk for the exponential test plan (T, c) is

$$\begin{aligned} P(\lambda_S \geq \lambda_0 \mid \text{Test Is Failed}) &= P(\lambda_S \geq \lambda_0 \mid W > c) \\ &= \int_{\lambda_0}^{\infty} p(\lambda_S \mid W > c) d\lambda_S \\ &= \int_{\lambda_0}^{\infty} \frac{f(W > c|\lambda_S)}{\int_0 \infty f(W > c|\lambda_S) d\lambda_S} d\lambda_S \\ &= \frac{\int_{\lambda_0}^{\infty} [\sum_{W=0}^{c} \frac{(\lambda_S T)^W exp(-\lambda_S T)}{W!}] p(\lambda_S) d\lambda_S}{\int_0^{\infty} [\sum_{W=0}^{c} \frac{(\lambda_S T)^W exp(-\lambda_S T)}{W!}] p(\lambda_S) d\lambda_S} \leq \alpha \end{aligned}$$

For simplicity we if fix c to be zero and then perform Monte Carlo integration using N posterior draws $\lambda_S^{(j)}$

$$\begin{aligned} P(\lambda_S \geq \lambda_0 \mid W = 0) &= \frac{\int_{\lambda_0}^{\infty} exp(-\lambda_S T) p(\lambda_S) d\lambda_S}{\int_0^{\infty} exp(-\lambda_S T) p(\lambda_S) d\lambda_S} \\ &\approx \frac{\sum_{j=1}^{N} exp(-\lambda_S^{(j)} T) I(\lambda_S^{(j)} \geq \lambda_0)}{\sum_{j=1}^{N} exp(-\lambda_S^{(j)} T)} \end{aligned}$$

Similarly, given that the test is passed, the *posterior consumer's risk* is the probability that $\pi \leq \pi_1$, or $P(\pi \leq \pi_0|\text{Test Is Passed})$. Notice that this is simply the posterior probability that $\pi \leq \pi_1$ given that we have observed no more than c failures. In the exponential case, if we let π be the system failure rate λ_S, then using Bayes' Theorem, and assuming a maximum allowable posterior consumer's risk β, an expression for the posterior producer's risk for the exponential test plan (T, c) is

$$\begin{aligned} P(\lambda_S \leq \lambda_1 \mid \text{Test Is Passed}) &= P(\lambda_S \leq \lambda_1 \mid W \leq c) \\ &= \int_{0}^{\lambda_1} p(\lambda_S \mid W > c) d\lambda_S \\ &= \int_{0}^{\lambda_1} \frac{f(W > c|\lambda_S)}{\int_0 \infty f(W > c|\lambda_S) d\lambda_S} d\lambda_S \\ &= \frac{\int_{0}^{\lambda_1} [\sum_{W=0}^{c} \frac{(\lambda_S T)^W exp(-\lambda_S T)}{W!}] p(\lambda_S) d\lambda_S}{\int_0^{\infty} [\sum_{W=0}^{c} \frac{(\lambda_S T)^W exp(-\lambda_S T)}{W!}] p(\lambda_S) d\lambda_S} \leq \beta \end{aligned}$$

For simplicity we if fix c to be zero and then perform Monte Carlo integration using N posterior draws $\lambda_S^{(j)}$

$$P(\lambda_S \leq \lambda_1 \mid W = 0) = \frac{\int_0^{\lambda_1} exp(-\lambda_S T) p(\lambda_S) d\lambda_S}{\int_0^{\infty} exp(-\lambda_S T) p(\lambda_S) d\lambda_S}$$
$$\approx \frac{\sum_{j=1}^{N} exp(-\lambda_S^{(j)} T) I(\lambda_S^{(j)} \leq \lambda_1)}{\sum_{j=1}^{N} exp(-\lambda_S^{(j)} T)}$$

3.2.3 Weibull Case

In the exponential case we defined our quantity of interest that is related to system reliability at a given time π to be equal to the system failure rate λ_S. This gave us a convenient distributional form for our system failure counts with $W(t) \sim \text{Poisson}(\lambda_S t)$

3.2.4 JLTV Test Plan Results

References

[1] Director, Defense Test and Evaluation. Test and evaluation of system reliability availability maintainability - A primer; Third Edition (1982).

[2] M. S. Hamada, A. G. Wilson, C. S. Reese and H. F. Martz. Bayesian Reliability. Springer (2008).

[3] W. Q. Meeker and L. A. Escobar. Reliability: the other dimension of quailty. *Quality Technology and Quantitative Management* 1, 1-25 (2004).

[4] M. S. Hamada, A. G. Wilson, B. P. Weaver, R. W. Griffiths and H. F. Martz. Bayesian binomial assurance tests for system reliability using component data. *Journal of Quality Technology* 46, 24-32 (2014).

[5] V. E. Johnson, T. L. Graves, M. S. Hamada, and C. S. Reese. A hierarchical model for estimating the reliability of complex systems. In Bayesian Statistics, Vol. 7, J. M. Bernardo, M. J. Bayerri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West eds., Oxford, UK: Oxford University Press (2003).

[6] M. S. Hamada, H. F. Martz, C. S. Reese, T. L. Graves, V. E. Johnson, and A. G. Wilson. A fully Bayesian approach for combining multilevel failure information in fault tree quantification and optimal follow-on resource allocation. *Reliability Engineering and System Safety* 86, 397-405 (2004).

[7] P. M. Ellner and J. B. Hall, An approach to reliability growth planning based on failure mode discovery and correction using AMSAA projection methodology. *IEEE Proceedings of the Annual Reliability and Maintainability Symposium* (2006).