How to Improve Maintenance Using Serial Number Tracking Data

Leopoldo E. Soto Arriagada • S. Craig Goodwyn

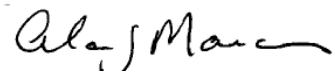
CRM D0022141.A1/Final
February 2010

CNA
ANALYSIS & SOLUTIONS

Photo credit line

U.S. Navy/Photographer's Mate Airman Javier Capella. Aviation Electronics Technician 3rd Class Robert Ponce and Aviation Electronics Technician Airman James Armstrong troubleshoot the power converter of a radar target data processor from an FA-18C Hornet aboard the Nimitz-class aircraft carrier USS Theodore Roosevelt (CVN 71).

Approved for distribution: February 2010



Alan J. Marcus, Director
Infrastructure and Resource Management Team
Resource Analysis Division

This document represents the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the Department of the Navy.

Approved for Public Release; Distribution Unlimited. Specific authority: N00014-05-D-0500.
Copies of this document can be obtained through the Defense Technical Information Center at www.dtic.mil
or contact CNA Document Control and Distribution Section at 703-824-2123.

Contents

Executive summary

We use the maintenance histories of APG-65 radar assemblies to show that analysis at the individual part level can provide valuable information about the reliability of parts. While traditional analysis focuses on *average* failure rates, in this study we show that it may be advantageous for the Navy to look at *individual* failure rates.

We focus on service time between failures, which defines a repair cycle. We follow part histories from birth, over a series of repair cycles throughout the service life of a part. Our analysis shows that:

1. The reliability of new parts drops quickly over the first few repair cycles and stabilizes thereafter. This implies that it is possible to set optimal retirement rules for parts.
2. Seemingly identical parts can have different levels of reliability—some parts turn out to be *lemons*. Analysis of the reliability of individual parts over a life cycle shows that some parts are inherently less reliable than others, even after accounting for a comprehensive set of factors that affect reliability.
3. We can identify lemons early in the life cycle, and we propose alternatives to deal with them. Lemons provide less service hours before failure and consume more maintenance resources than average parts; early identification will help free up maintenance resources.

The analysis for the APG-65 radar assembly presented in this study is a proof of concept that can be used for the analysis of other parts. Moreover, additional information about the cost of new parts and the cost of repairing parts would help the Navy define optimal retirement rules for average parts and early retirement rules for lemons. Our analysis also shows that it is possible to segregate individual parts according to their reliability. This information would allow the Navy to always have the best parts in operation.

This page intentionally left blank.

Maintenance histories and serial number tracking

Introduction

Traditionally, spare part failure rate analysis has focused on *average* failure rates. However, we believe that looking at failure rates at the *individual* level may be advantageous to the Navy. As a proof of concept, we used serial number tracking to analyze individual parts and their repair cycles. We looked specifically at APG-65 radars, and we focused on service time between failures, which is what defines a repair cycle. In our analysis, we found three important results:

1. The service time between repairs drops quickly and then stabilizes after a few repair cycles. Newer parts are more reliable than older parts, which suggests the possibility of optimal retirement policies.

2. Not all parts are created equal. While we found several observable factors that affect reliability, we also found that some parts are just inherently worse than others. These less reliable parts, or *lemons,* require more repairs and provide less hours of service than the average part.

3. We can identify lemons early in their life cycles, and we propose alternatives to deal with lemons. This could help the Navy free up maintenance resources.

We define our metric for reliability as the time of service between failures. In particular, we measure Flying Hours (FH) between failure. We use this metric to evaluate the reliability of parts throughout their life cycles.

We are able to follow a part over its service life by using serial number tracking.

We note that while the Department of Defense (DoD) and the services are implementing new technology to be able to track parts transparently through the system, current data are far from perfect. However, despite the problems with the data, we were able to construct a sample of complete maintenance histories, which allowed us to conduct our analysis.

We analyze the data in three ways.

1. We use descriptive statistics to see how reliability changes over the course of the lifetime of a part.
2. We also construct an event history model to test the effect of different factors such as age, employment (FH and sorties), cannibalizations and removals, depot repairs, and other variables on the reliability of a part. Moreover, we use this model to test for the presence of unexplained heterodoxy in the reliability of parts.
3. We use cluster analysis to classify parts into different levels of reliability and for early identification of lemons.

In the next few sections, we present the details of our analysis. To provide some background into this type of work, we start with a general description of the maintenance process of spare parts. That discussion is followed by a concise literature review. Later sections detail the analysis and results.

The maintenance process

Our analysis requires that we follow a spare part throughout its service life. We use serial numbers to keep track of the part as it moves from supply to aircraft, to the maintenance system at failure, and back to supply after repair. To better understand the life cycle of a spare part, we present a concise description of the maintenance process below.

The life cycle of parts

In general terms, the life of a spare part is defined as a cycle of operation, failure, and rejuvenation in the form of repairs. From induction into the supply system, a spare part goes through several of these cycles; after each repair, the part is brought back to a working state and re-issued for installation and operation.

A new Aviation Depot Level Repairable (AVDLR) installed in an aircraft will remain there until failure.[1] After the failed part is removed from the aircraft, the Aviation Intermediate Maintenance Department (AIMD) has three options:

1. Repair the part and return it to the supply system,
2. Declare the part Beyond the Capability of Maintenance (BCM) and send the part to the depot for repair, or

1. In some cases, components are removed from one aircraft and installed on another to allow the second aircraft to perform its flight operations. This practice is commonly referred to as *cannibalization*. CNA has conducted a variety of studies on the effects of cannibalization on reliability. See [1] and [2], for example. This study also analyzes this issue in the context of its effects on the reliability of an individual part.

3. Leave the part in its current state until repairs can be completed or, on rare occasions, the part is purged from the supply system.[2]

This same scenario is repeated for all parts (not consumables) many times during a life cycle.

While this defines the life history of a part, it can also help us define its reliability. For example, a spare part that can be used for several FH before failure would be considered more reliable than a part that fails after very few FH. In fact, in this study we focus precisely on the amount of service FH a part can last before failure as a measure of its reliability.

2. There are several reasons why a part may not be repaired soon after failure. One of the most likely reasons is that there is enough inventory of working parts to satisfy demand. Another reason is that there could be personnel and/or budget constraints that prevent the part from being repaired. Finally, certain parts experience decreasing demand when the aircraft that use those parts are retired.

Analysis of maintenance data

Traditional analysis

Traditional reliability analysis aims to forecast the average Mean Time Between Failures (MTBF) for parts of a specific type. Our approach differs from this type of analysis in that we focus on the individual part rather than averages. However, we do note that looking at averages is particularly helpful in identifying factors that affect the average reliability of parts. Therefore, it is useful to review the main results found in this literature as many of these results may apply at the single part level. Furthermore, in reviewing these studies, we are able to solidify our modeling strategy.

We divide this section into four subsections: (1) aging, (2) number of FH and number of sorties, (3) cannibalizations, and (4) other factors that affect reliability.

Aging

Several CNA studies have analyzed the effect of aging on reliability. Most of these studies are aggregate in nature and look at the effects of aging on the fleet of Navy aircraft. The results from these studies indicate that age decreases the average reliability of aircraft.

For example, Kleinman [3] looks at the effects of aircraft conversion on reliability.[3] However, he also looks at the effects of age on reliability before and after conversion. He finds that conversion erases much of the effects of aging on the aircraft. However, the aging process continues after conversion. This points to the possibility of rejuvenation of aircraft as a result of modification.

3. An aircraft conversion is when an aircraft is modified to extend its life, change its mission, or increase its capabilities. A conversion usually entails the upgrade of the aircraft series.

Jondrow et al. [4] analyze the effects of aging on depot level repairs of aircraft components. They use data for several years and several Type-Model-Series (TMS) and find a direct relationship between the aircraft's age and the number of repairs. Their analysis is at the aggregate level, and they find that an increase in the average age of the Navy's fleet of FA-18Cs (with an average age of 8.2 years in 2002) would increase the number of BCM repairs per FH by 9.2 percent. They pay particular attention to the budget and reliability effects associated with this relationship.

Boning and Brown [5] find further support for the negative relationship between airframe age and reliability. However, even though they do find a strong correlation between these two, they point out that structural changes in the Navy may have unaccounted for effects on the relationship between the average age of the aircraft fleet and its average reliability.

Boning, Soto Arriagada, and Goodwyn [6] examine many of the factors that affect reliability. They use squadron level data for the FA-18 fleet and find evidence consistent with previous research. They also include a section of component level analysis. While they are able to provide descriptive statistics at the part level, the data do not support further analysis. However, the authors do show preliminary evidence of age effects at the part level in the form of a declining number of sorties after each repair.

Number of FH and number of sorties

There is a long-standing belief of a differential effect of the number of FH and sorties on reliability. The general idea is that take-offs and landings have different effects on reliability than hours in the air do.

In a seminal CNA paper, from 1986, Levy [7] generalizes the standard model that relates reliability to FH by introducing a sortie effect. In the resulting model, the probability of failure consists of two underlying processes, one related to flight hours and another to sorties. The basic implication of this theoretical model is that these two effects are different.

Several more recent papers have tested this hypothesis empirically and have found differential effects. Results by Boning, Soto Arriagada, and Goodwyn [6], for example, support the general consensus, and they report that FH decrease reliability at a decreasing rate, while the number of sorties increases both the number and the rate of failure.

Cannibalizations

In general practice, parts are removed after failure, and they are replaced by a working part. However, in many instances a part is removed, or cannibalized, from an aircraft to allow a second aircraft to perform a flight mission. Evidence shows that there are negative effects of cannibalization on reliability.

Levy [1] analyzes the failure rates of FA-14A systems and subsystems for the 1986 deployments of the USS Enterprise and the USS America. The results show that cannibalizations decrease reliability for many but not all parts.

In 1989, Levy [2] updates his previous research using FA-18 avionic components and confirms the finding that cannibalizations have detrimental effects on reliability. However, this research also shows that these effects are not constant across systems or air wings. The effects extend to the aircraft that is cannibalized as well as to the aircraft that receives the part.

Other factors that affect reliability

Reliability is influenced by several factor other than those listed above. Differences in reliability may arise from (1) employment of aircraft, (2) where the part was last repaired, (3) the environment in which the part operates, (4) modifications to the part, or (5) differences in the parts themselves.

Boning, Soto Arriagada, and Goodwyn [6] present a comprehensive examination of many of these issues. They report that there are positive employment effects for aircraft deployed in theater. However, this result may be driven by low tolerance for failure during combat deployments and pre-deployment “grooming.”

They also examined location effects. These effects capture both the environment in which the aircraft operates and where the repairs where executed. The variables included in the model further confound the effects of the environment in which the aircraft operates and where parts were repaired. This makes these two effects impossible to disentangle without further information. However, without them, results from the model would be biased. So, while the interpretation of these variables would vary with context, they serve an important modeling purpose.

When comparing reliability across types of parts, the complexity of the part may affect reliability. In a 1952 study, Boodman [8] shows that increased complexity negatively affects reliability, and he proposes methods for improving it. His results point to heterogeneity as a source of differences in reliability.

Follmann and Goldberg's [9] 1986 study looks at the effect of heterogeneity on reliability, focusing on parts of the same type. The main assumption is that different copies of the same machine, or unit, may have different failure rates.

They report that ignoring this heterogeneity may induce bias in the results. They propose a model in which time to failure is assumed to follow a Weibull distribution and heterogeneity varies randomly under a gamma distribution. Note that despite the age of this reference (i.e. 1986), gamma distributions are still frequently used in the modeling of heterogeneity in survival studies.

In the next section, we show how we can use the ideas presented above in the modeling of failure rates at the single part level.

Modeling strategy

Reliability of the single part

As mentioned in the introduction, we examine the reliability of a part through its service life. We follow each individual part in our sample from first induction into the supply system to the last repair that appears in our data. These data, together with the assumption that each individual part has different levels of reliability even after accounting for other effects, drive our modeling strategy.

Our main goal is to find those factors that explain how many FH a part can operate between failures. We build upon the existing literature to produce a model that accounts for the life histories of individual parts as well as heterogeneity in reliability.

To achieve our goals, we use an event history model of repeated events in which the previous history of a part may influence current reliability. We can also use the model to test for otherwise unexplained heterogeneity in the reliability of parts.

Explaining maintenance histories

For our analysis, we use a sample of parts with complete maintenance histories. For each part, we observe the number of FH the part operates between repair cycles. We also observe a number of explanatory variables that we use to explain the number of FH a part can operate between failures.

The simplest way to model these data would be to use regression analysis (Ordinary Least Squares (OLS) regression) to explain FH per repair cycle as a linear function of several explanatory variables. While previous research uses OLS for the analysis of data, the aggregate nature of the data and the questions under examination, make OLS appropriate for those cases.

In this case, however, two common characteristics in the data would severely bias OLS results. These are:

1. Event dependence: the fact that previous failures and repairs may influence the number of FH a part will last before its next repair.

2. Heterogeneity: there may be unobserved heterogeneity that makes some parts more susceptible to failure than other parts—no matter what the event history of the part is.

Event dependence and heterogeneity violate OLS regression assumptions. Other characteristics of our data, such as censoring,[4] also makes OLS estimation inappropriate for this type of problem.

Therefore, in the following sub-sections, we explore other modeling strategies. To guide our modeling choices, we look at how these assumptions are violated and what we can do to account for these effects.

The effects of event dependence and heterogeneity

Event dependence and heterogeneity create within-subject correlation, which violates the OLS assumption that observations are independent and identically distributed (i.i.d) random variables. This is unlikely in the case of time-to-failure data.

Moreover, analyzing this type of data using OLS would also violate the normality assumption for the distribution of the error term. While the consequences of violating the normality assumption are fairly benign and often correctable, violation of the i.i.d. assumption will cause OLS estimation to be biased and maybe severely so.

4. Censoring occurs when we cannot observe a variable below a certain threshold. In this case, we cannot observe parts before they are first inducted in the supply system. In our sample, we only use observations that we can observe from this point on. However, censoring also occurs because we do not know what happens to the part after our last observation.

In the case of event dependence, for example, a bad repair will cause the number of FH a part can operate in its next cycle to be shorter than it would normally be. So, a previous event helps explain how long the part will be operational before its next failure. In fact, if that bad repair cannot be corrected in future repairs, its negative effects on reliability will last for the life of the part. OLS requires a *previous* history, which, in turn, provides *no* information about future outcomes.

In other words, the even dependence caused by the bad repair makes each repair cycle shorter than it would otherwise be. This, in turn, makes the timing of failures likely to be correlated for a given part, and it defines the within-subject serial correlation and renders OLS biased. Moreover, OLS would not be able to provide a good picture of the dynamics of the maintenance process.

Heterogeneity acts in a similar way. The main difference is that the serial correlation is not caused by any particular event. In this case, the serial correlation is caused by inherent characteristics of the part that, for some reason, we are not able to capture directly in our model. There is much anecdotal evidence of parts that fail more often than others without clear explanation. In fact, there is even a name for aircraft that are thought to be less reliable than others; they are often called “hangar queens.”

Problems with this type of serial correlation also affect models specifically designed for event data. The Cox model [10] or proportional hazards model is the most widely used event history model.

This model assumes that event times are independent conditional on the explanatory variables included in the model. This means that if there is any within-subject serial correlation in the data, the explanatory variables would have to capture it, or the estimates from this model would be biased. So, any remaining correlation, such as that induced by event dependence or heterogeneity, if not captured in the explanatory variables, would violate this assumption. The models described in the next section are designed to correct this violation for specific types of analysis.

An empirical event history model

As mentioned above, two characteristics of the type of data needed for the kind of analysis we conduct here are event dependence and heterogeneity. These are common features of repeated event processes. However, a common problem with this is that it is usually impossible to determine whether one, the other, or both are the cause of the within-subject serial correlation.

Until recently, there were two types of models that could be used in the analysis of repeated event data: *variance-corrected* and *frailty models*. A recent strand of the literature has provided a third type of model, the *conditional frailty model*, which combines the ideas of the previous two and provides more flexibility to the assumptions needed for the estimates to be unbiased. In the next few sections, we describe these models.

Variance-corrected models

As the name implies, variance-corrected models account for subject-specific effects that are explicitly specified in the model by adjusting the variance-covariance matrix. While these models are particularly well suited for the case of event dependence, if the cause of the correlation between event or failure times is heterogeneity, the estimated effects of the explanatory variables in the context of the Cox model are biased but still consistent. This means that there would be systematic error in the estimates, but this error disappears as the sample size increases.

These types of models can take different forms depending on the type of analysis that is needed. The models differ in how the risk set is defined—whether events are in sequential order or whether they can occur simultaneously. They also differ in how time at risk of failure is counted—whether it starts from first observation and keeps counting to the last observation or whether it re-starts at zero after each event. Lastly, they also differ in whether the data are stratified—that is, whether the baseline hazard of the events differs by event number. It is important to note that stratification is particularly well suited for accounting for event dependence.

Each of these different versions of the model is appropriate for different types of dynamic processes and, therefore, different results will be produced.

Simulations conducted by Box-Steffensmeier, De Boef, and Joyce [12] show that these types of models perform best in the analysis of data with event dependence and no heterogeneity, and in cases with no heterogeneity and no event dependence. They also show that these models are not well suited for cases where heterogeneity is present. For that type of analysis, a frailty model may be more appropriate.

Frailty models

Frailty models are random effect models that capture the fact that some subjects, or parts in this case, may be more prone to failure than others. The assumption is that frailties are time-invariant, unobserved, individual effects.

Frailty models can also differ by risk set and by how time to failure is counted. These types of models incorporate heterogeneity in the model by treating the individual effects described above as random draws from a given parametric distribution. The parameters of the specified distribution are estimated in conjunction with the other parameters in the model. This ensures that the correlation between failure events is explicitly included in the model. So, conditional on the chosen parametric distribution, event times are assumed to be independent of each other. In other words, we assume that the distribution of the frailty across individuals captures any serial correlation in the data.

One problem with this type of model is that there is no particular guidance in the choice of distribution to capture the effects of frailty. Current research is inconclusive.

Another problem is that it is generally required that the frailty terms are independent of the covariants. Hausman [11] shows that violation of this assumption results in biased estimates.

Box-Steffensmeier, De Boef, and Joyce's [12] simulations show that this type of model is most suitable to problems without event dependence. The main reason for this is that the baseline hazard does not vary with the event number in conditional frailty models as it does in stratified variance-corrected models.

Conditional frailty—a more general modeling strategy

The models described above are well suited for dealing with either heterogeneity or event dependence, but not both at once. This creates a problem for research where both problems are present or where there is no way to discern which one is causing the within-subject serial correlation.

The conditional frailty model proposed by Box-Steffensmeier and De Boef [13] accounts for both heterogeneity and event dependence. It uses controls for heterogeneity through a frailty random effect. It also accounts for event dependence through event based stratification.

The conditional frailty model is cast in gap time. This means that the interpretation of the coefficients associated with the explanatory variables tells us the effect of the covariate on the hazard of failure since last occurrence.

Box-Steffensmeier, De Boef, and Joyce's [12] simulations show that the conditional frailty model performed at least as well as the variance-corrected and the frailty models whether or not event dependence and/or heterogeneity is present in the data.

The conditional frailty model is the most appropriate way to proceed given our specific analytic problem because both event dependence and heterogeneity are quite likely in our data. And, while we do have an extensive set of explanatory variables, we cannot be certain that this set of variables will capture all reliability nuances. Moreover, the conditional frailty model allows us to test whether heterogeneity is present in our sample.

Data and analysis

The analysis we present here is a proof of concept. In general and as shown in the literature review above, the focus of reliability analysis has been at the aggregate level. That type of analysis is mostly driven by the need to provide the Navy with reliable budget forecasts. The goal here is different. We examine reliability at the single part level and consider whether this type of analysis can benefit the maintenance community.

We show that there are important gains from examining single part maintenance data. In our proof of concept, we use data from the APG-65 radar assembly. Through our analysis, we are able to show that:

1. After a new part is inducted, the average number of FH between repairs drops quickly in the first few repairs, and then stabilizes. A new part is more reliable than an older part and the transition from new to old occurs quickly from the first to about the sixth repair. After the sixth repair, the average number of FH between repairs remains fairly constant.

2. Not all parts are created equal. Some parts have lower than average reliability—they require more repairs and provide less service hours than the average part.

3. It is possible to identify less reliable parts early in their life. This will allow the Navy to set maintenance policies and practices that minimize the effect of these less reliable parts on the maintenance system.

Our analysis also allowed us to test whether those factors thought to affect reliability at the aggregate level also have an effect when we conduct the analysis at the single part level.

In the next few sections, we describe the type of data necessary for this type of analysis and the types of results we were able to obtain.

Data for the APG-65 radar assembly

As a proof of concept, we use data from the APG-65 radar assembly for our analysis. We chose this radar because of its long maintenance history and the large number of observations. We point out that this type of analysis can be carried out on any type of part for which we have maintenance histories.

We use serial number data from the Aviation Standard Navy Maintenance and Material Management System (AV-3M). These data allowed us to construct maintenance histories for individual APG-65 radar assemblies during their service life. Whenever possible, we also used auxiliary data to identify new parts and to extract our explanatory variables.

Data from the AV-3M database, however, are not in a ready-to-use format for this type of analysis. As documented in Boning, Soto Arriagada, and Goodwyn [6], the part histories in the raw data contain large numbers of gaps. The main problem is that in many records, serial numbers are not recorded.

To correct these data problems, we use several different types of maintenance documents to construct maintenance histories that allow us to follow specific parts from the beginning of their service lives over a relatively large number of repair cycles. We describe the process below.

Constructing maintenance histories

The primary documents we used to construct maintenance histories for the APG-65 radar are material issue documents, material turn-in documents, and maintenance action forms (MAF). While the material issue and turn-in documents only identify parts using their National Item Identification Numbers (NIIN), MAFs also include manufacturer part numbers and serial numbers.

The serial numbers often contained errors that prevented us from seeing complete histories directly from the raw data. We used the documents above together with the careful matching of time lines to construct coherent maintenance histories.

However, because serial numbers were unreliably recorded, we constructed our own pseudo-serial numbers to keep track of the data. MAFs also made possible the use of Bureau Numbers (BUNO) to identify the aircraft in which a given part was installed. With this, we kept track of the TMS in which the part was installed and the number of FH of service before the next failure. This also allowed us to identify new aircraft, which, in turn helped us flag new parts inducted into the system with the new aircraft.

FH and reliability

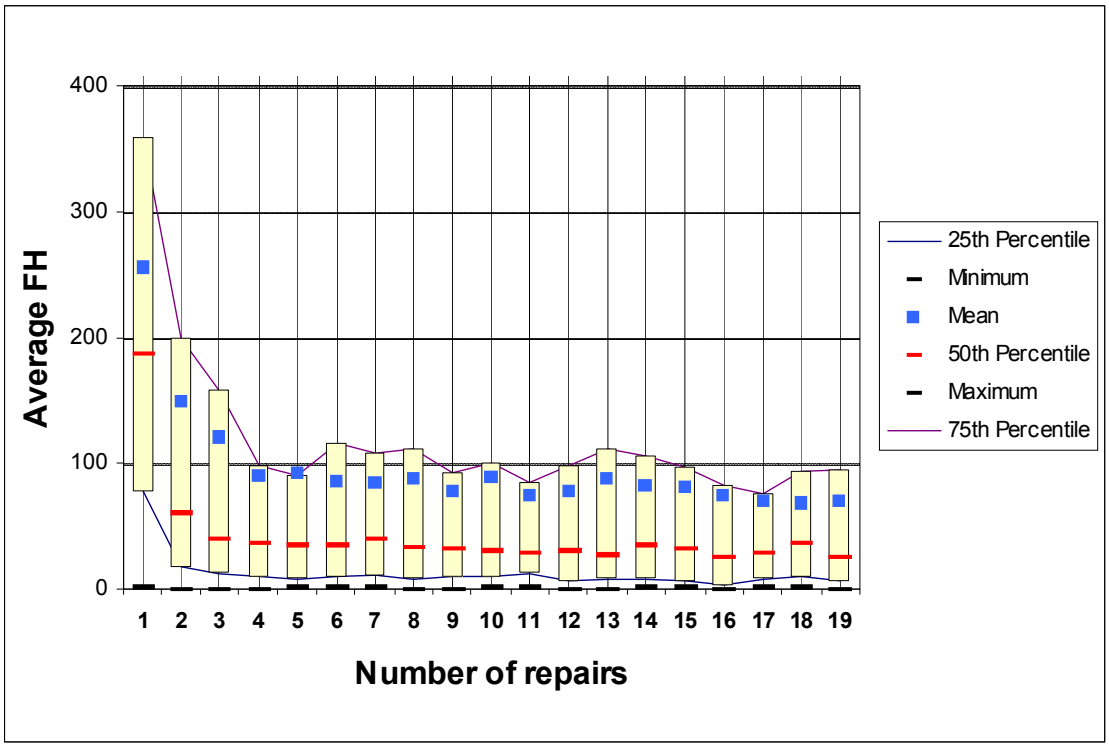
In this paper, we use FH between failures as a measure of reliability of a part. Throughout the paper, our goal is to identify those factors that determine how many FH a part can last after each repair.

A first look at the data gives us an interesting insight. The average number of FH a part can operate between repairs drops rapidly for a new part and stabilizes after a few cycles. This indicates that the average new part is more reliable than the average older part. The blue squares in figure 1 illustrates this.

These results could be used to define retirement policies. If, for example, an APG-65 is relatively cheap compared to its repair costs, it may be cost effective to retire the part early. However, if buying a new radar is relatively costly when compared to repair costs, it may be cost effective to continue to use the part until it can no longer be repaired.

The scope of this project does not allow us to pursue further investigation of this issue due to data constraints. However, with data on replacement and repair costs it would be possible to determine whether or not a retirement policy is appropriate for a specific part. Furthermore, with these data, it would also be possible to determine which retirement policies are cost effective. These policies could provide significant savings to the Navy because older parts that consume more maintenance resources from the system would be removed. Also, set retirement policies will allow for accurate procurement forecasts, which, in turn would facilitate programming and budgeting.

Figure 1. Average number of FH per repair[a]



a. Although we only report on a few repair cycles in this graph, for some parts we observe up to 40 cycles. We do not include all of them here because the sample sizes for higher numbers of repair cycles became small and, thus, produced spurious results.

Figure 1 shows us that the median (50th percentile—marked in red) is, in all cases, below the mean, while the mean is, in general, close to the 75th percentile (50 percent of all observations are within the 25th and 75th percentiles marked as the extremes of the yellow rectangles). This indicates that there are a few outliers that have much higher reliability than the rest of the sample. The low median, closer to the 25th percentile indicates that there may be a robust number of parts with low reliability.

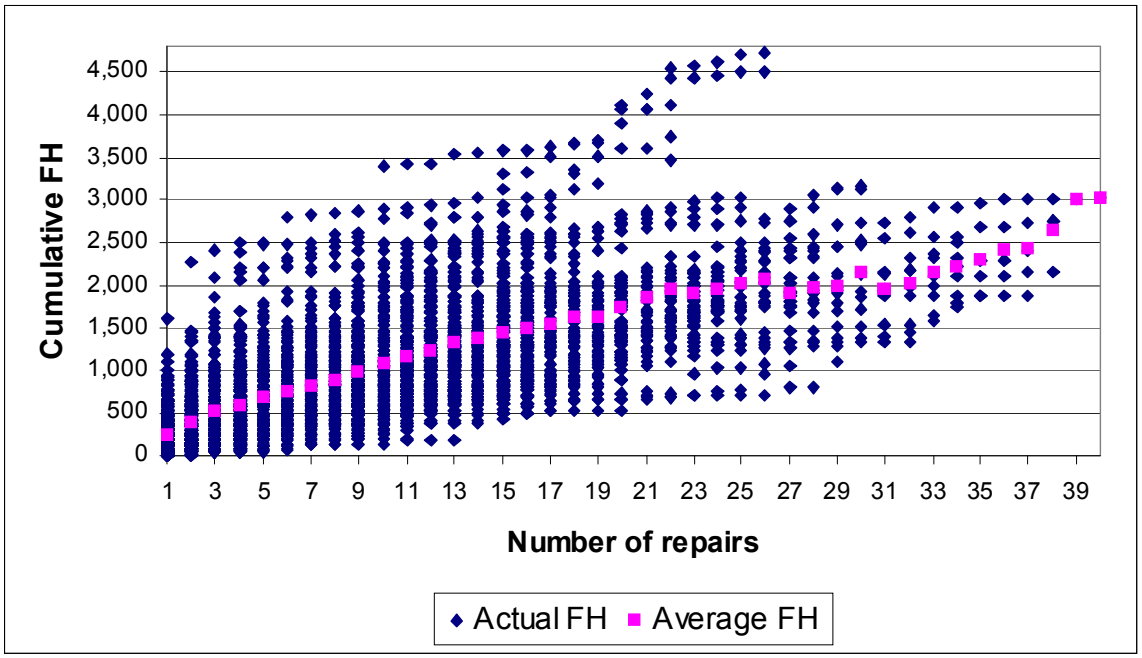
Fundamentally, shows us that MTBF does not tell us the whole story. To illustrate this point, we plot the cumulative number of FH against the number of repairs and compare it to the mean cumulative FH.

This gives us an illustration of the dynamics and distribution of reliability across our sample.

Figure 2 shows us the cumulative observed FH per repair in contrast to the average cumulative FH per repair. Every blue dot in the figure corresponds to an actual observation, while a purple square shows the average FH for each repair.

The story told by the average cumulative FH is consistent with the MTBF results. While average cumulative FH increase with each repair cycle, they do so at a decreasing rate. This means that the reliability of new parts degrades as the number of repairs increases. After that, it levels off.

Figure 2. Cumulative observed and average FH per repair



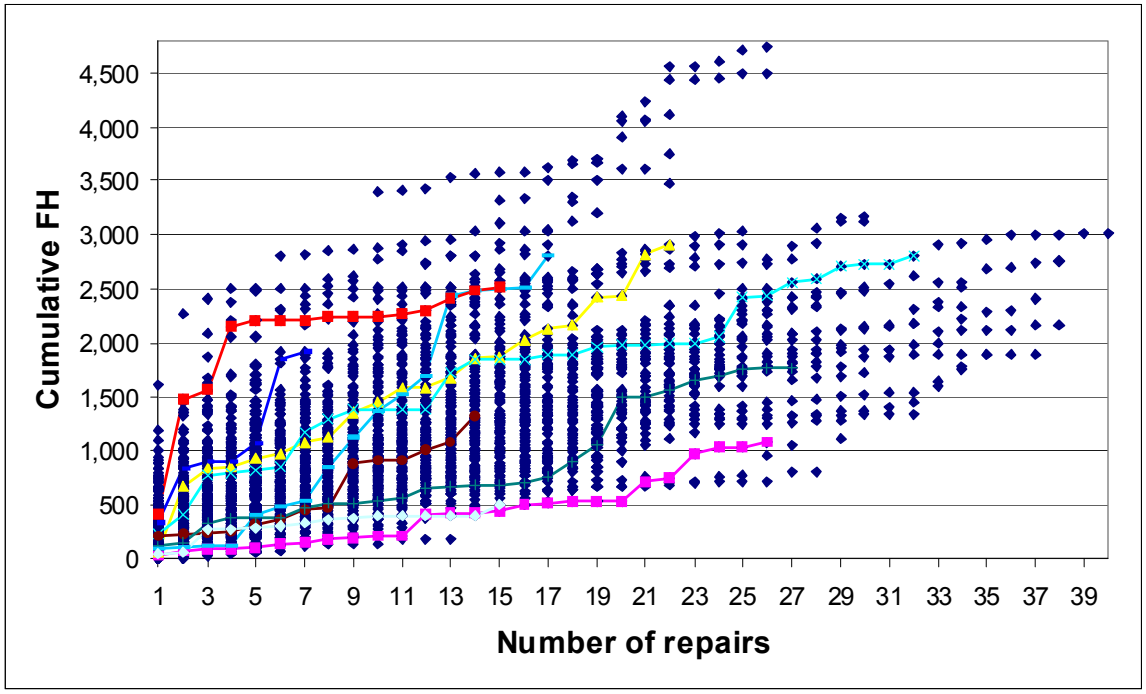
Note that beyond the 20th repair, the sample sizes become smaller (less blue dots) and the trends in the average become more erratic. Also note the wide distribution of FH around the mean.

Figure 3 helps us explain this point. Note that observations across repairs (blue dots in figure 3) correspond to specific parts, and

together they are the building blocks of a part history. Also note that trying to single out each part history in the graph would only make it more difficult to see the dynamics of the data. So, to illustrate this dynamic, we randomly selected a sub-sample of parts, which we join with colored lines for ease of identification.

As parts go through an increasing number of repairs, the number of cumulative FH the part is in service grows. The colored lines illustrate that this growth is not the same for all parts, and it is definitely different from the growth in the average represented in figure 2. In fact, it is easy to see that the average does not grow as rapidly as the rest of the parts due to the influence of less reliable parts.

Figure 3. Cumulative FH per repair by part[a]



a. For illustrative purposes, we highlight the life cycles of a randomly selected sub-sample of parts using colored lines.

To clarify this point, take the red line, for example, which shows a part that achieves a relatively large number of FH within a few repairs. This part, then, represents a very reliable part. In contrast, the purple line

only completes relatively few FH over a relatively large number of repairs. This is an example of an unreliable part. It is possible to see that the average is heavily influenced by both the new parts in the beginning of the series and by the less reliable parts at the end.

This points to two important conclusions; first, the average may not be telling us a complete reliability story, and second, there are great differences in the reliability of parts. However, without further analysis, we cannot make statements about the reasons for the differences in the reliability of these parts. It may be that the part that seems less reliable was operating in extreme environments. It could also be that there are differential levels of reliability inherent in the parts themselves.

In the next few sections, we use the conditional frailty model described above to disentangle the factors that affect the reliability of a part. We test whether heterogeneity of the parts themselves has an unobserved impact on reliability.

Event history analysis

Description of the data and analysis

We use the conditional frailty model to estimate the effects of different factors on the reliability of a part. This model is particularly suitable for this type of analysis because it allows us to test whether different explanatory variables have a significant effect on reliability. It also lets us test whether inherent heterogeneity in the reliability of parts plays a role in determining how often a part will fail. We follow the previous literature in deciding which variables to include in the analysis. Table 1 reports a full set of descriptive statistics for the variables included in the analysis.

Table 1. Descriptive statistics

Variable	Mean	Std. Dev.	Min	Max
Number of consecutive repairs	10.0730	7.5970	1	41.0
Cumulative FH	1024.2810	752.8859	0.8	4733.0
FH per repair	103.9993	160.0411	0.2	1613.3
Age in months	81.1041	56.0768	0	223.0

Table 1. Descriptive statistics

Variable	Mean	Std. Dev.	Min	Max
Days on wing	125.4830	218.3749	0	3137.0
Cumulative sorties	646.0826	481.9056	1	3528.0
Sorties per repair	65.0359	99.0766	1	1105.0
AV-8 sorties	3.1591	25.8557	0	614.0
FA-18 E-F sorties	11.3707	52.4316	0	1020.0
Other sorties	0.1341	4.8833	0	265.0
AV-8 FH	5.9555	48.0982	0	1012.1
FA-18 E-F FH	17.9656	81.5245	0	1499.5
Other FH	0.2293	8.6777	0	490.6
Removals	1.5733	0.9505	1	12.0
Cannibalizations	0.3362	0.6934	0	8.0
BCM	0.0316	0.1750	0	1.0
Navy Atlantic	0.1811	0.3851	0	1.0
Shipboard Atlantic	0.0975	0.2966	0	1.0
Shipboard Pacific	0.1293	0.3356	0	1.0
MC Atlantic	0.0797	0.2708	0	1.0
MC Pacific	0.1893	0.3918	0	1.0
Navy Pacific	0.2265	0.4186	0	1.0

Age

We measure age in months starting at zero, and we only follow those parts that we can identify as new at the beginning of their history. To determine whether a part was new, we selected parts that came with new aircraft.

The average age for the sample is 81 months, although the range of parts included in the analysis go from new to 223 months (approximately 18 years old).

Number of sorties and TMSs

As pointed out above, the literature reports a differential effect of FH and sorties on reliability. To capture these effects, we include total number of sorties up to the time of failure, as well as differential sorties and FH for different TMSs. The average number of sorties per repair is 65, although it may be different for different TMSs. We also include days on wing in the model. This variable should capture whatever time effects we cannot capture with age or time in operation.

Removals and cannibalizations

We also separate the effects of removals and cannibalizations in the analysis. Although we know that the reason for cannibalization is to use the part on a different aircraft, the reasons for removals are not so clear.

On average, an observation goes through 1.6 removals and 0.34 cannibalizations.

Repairs and environment

To complete our model, we include variables to show the location of where the repair was done or whether the part was BCMed. Location variables capture both the quality of the repairs and the effect of the environment on the part. Thus, we cannot draw strong conclusions from them. However, they need to be included in the model to avoid mis-specification biases.

Frailty and heterogeneity

To test for the presence of part heterogeneity, we assume that unobserved part reliability follows a gamma distribution. This assumption is in line with previous reliability research [9], and it is widely used in this type of model. Assuming that reliability heterogeneity across parts is distributed following a gamma distribution adds a single parameter to the estimation. If this parameter is statistically significant, we can conclude that there is heterogeneity in the reliability of parts beyond what we can explain with out set of explanatory variables.

Estimates

This model produces a frailty coefficient and coefficient estimates for each of the explanatory variables included in the model.[5] Note that most coefficients are statistically significant (p-values equal to 0.10 or smaller for the Chi square tests reported on table 2). We find that our results are generally consistent with the previous literature. We dis-

5. These variable include measures of age, number of sorties, removals and cannibalizations, and location.

cuss differences below. We also find statistically significant heterogeneity in the reliability of the parts under analysis.

To help with the interpretation of coefficients, table 2 includes a column labeled exp(coef).[6] The value for cannibalizations, for example, is 1.09. This means that cannibalization increases the likelihood of failure by 9 percent. Conversely, removals decrease the likelihood of failure by about 5 percent. Other estimates can be interpreted in the same manner.

Interpretation of the results

This model's results are consistent with the previous literature, but require some discussion. For example, the coefficient on age is marginally significant and shows a decrease in the risk of failure. However, this is not inconsistent with previous results. Because of our definition of reliability as FH per repair cycle, age is better captured by employment, in terms of sorties and FH rather than by calendar time. In fact, higher age in these data may point to more reliable parts that have lasted for a long time and are still in service.

The coefficient on days on wing provides further support for this argument. Days on wing measures the amount of time a part spends installed on an aircraft before failure. The effect of this variable on reliability is close to zero and, in fact, it is not statistically different from zero. We can only conclude that FH capture the effects of employment on reliability, while the time that a part spends on wing does not matter. Again, this points to the fact that calendar time does not have a direct effect on reliability, but using the part does.

Table 2. Coefficient estimates

	coef	exp(coef)	se(coef)	Chi-sq	p-value
Age in months	-0.001450	0.998551	0.000713	4.13	0.0420
Days on wing	-0.000022	0.999979	0.000170	0.02	0.9000
FA-18 A-D sorties	-0.074500	0.928207	0.001401	2830.60	0.0000
AV-8 sorties	-0.004810	0.995202	0.004128	1.36	0.2400

6. Exponential function of the coefficients.

Table 2. Coefficient estimates

	coef	exp(coef)	se(coef)	Chi-sq	p-value
FA-18 E-F sorties	0.000572	1.000572	0.003300	0.03	0.8600
Other sorties	-0.004170	0.995839	0.026348	0.03	0.8700
BCM	-0.169000	0.844509	0.117255	2.07	0.1500
AV-8 FH	-0.040900	0.959925	0.002334	306.21	0.0000
FA-18 E-F FH	-0.043600	0.957337	0.002258	373.05	0.0000
Other FH	-0.040700	0.960117	0.014980	7.39	0.0066
Removals	-0.046600	0.954469	0.028366	2.70	0.1000
Cannibalizations	0.088800	1.092862	0.039731	4.99	0.0250
Navy Atlantic	-0.141000	0.868489	0.168509	0.70	0.4000
Shipboard Atlantic	-0.401000	0.669650	0.173164	5.35	0.0210
Shipboard Pacific	-0.446000	0.640184	0.170673	6.82	0.0090
MC Atlantic	-0.463000	0.629393	0.178218	6.75	0.0094
MC Pacific	-0.349000	0.705393	0.168699	4.28	0.0390
Navy Pacific	-0.143000	0.866754	0.165967	0.75	0.3900
frailty (dist=gamma)		0.120000		339.81	0.0000

R-square=0.916 (max possible=1)
Likelihood ratio test=9308 on 166 df, p-value=0.0000
Wald test=3112 on 166 df, p-value=0.0000
Number of observations=3766

As found in previous analysis, in our model cannibalization has same effect on reliability. Removing a part that is functioning correctly to replace it on another aircraft has detrimental effects on the part without any particular gain.

On the other hand, removals require some discussion. While only marginally significant, the direction of the effect and its magnitude are important. Our estimates show that removing a part has positive effect on reliability.

A story consistent with this finding is that a part would be removed only if thought to be faulty to begin with. However, a removal without a repair, as it is captured in this variable, may mean that the part itself was not faulty, but rather that it was incorrectly set and, therefore, appeared to be faulty. Re-installation of the part would then ensure

that it is properly set and would allow it to work properly for a longer period of time.

Another set of coefficients that requires discussion are FH and sorties. We include sorties for all the TMSs in the sample, but for FH we leave one category out. The reason for this is that the variable that we are trying to explain is FH, so the excluded category acts as a baseline.

Interpretation of the sortie coefficients should be made with respect to each other, while interpretation of the FH should be made with respect to the omitted category. For example, we can say that FH on an FA-18 A-D are worse for reliability than FH in other TMSs. This result is also weakly true for FA-18 E-F sorties.

The last set of explanatory variables are the location variables. As mentioned above, these variables confound the effects of the location of the last repair and the operating environment of the part before failure. So, while serving as an important control in the model, they are difficult to interpret.

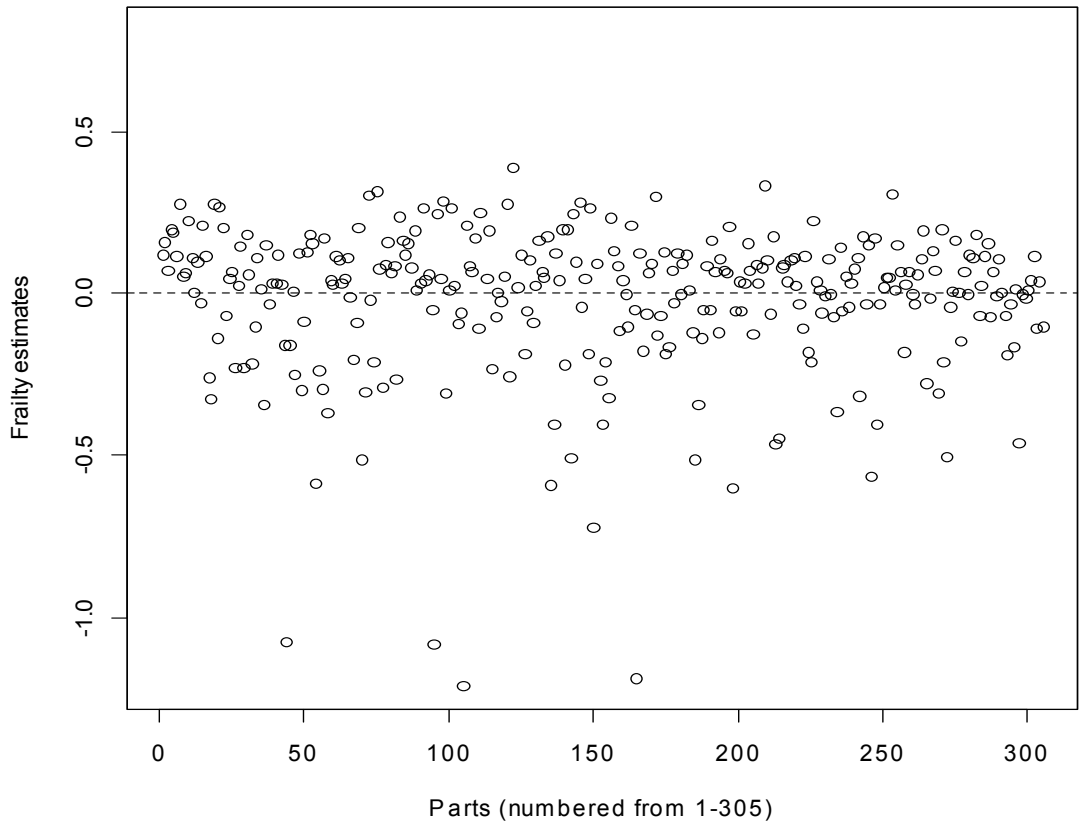
Like the TMS FH variables, the location variables also have an omitted category that serves as a baseline, and part of the interpretation depends on that. The reason for the exclusion of one category here is somewhat different than above. The location variables are indicator variables, so our concern here is co-linearity.

The omitted category includes parts without a location of previous repair (because they were new or by omission) and parts repaired in reserve and non-fleet facilities. Our results indicate better reliability for the included categories.

Heterogeneity

Lastly and most importantly, our estimates show that the frailty coefficient is highly statistically significant. This means that there is unexplained unobserved or unobservable heterogeneity in the data. Given the completeness of our explanatory variables, the main conclusion we can draw is that there is inherent variation that does not change over time and that makes each individual part have different levels of reliability. This points to the fact that some parts are more likely to fail no matter what.

Figure 4. Estimated frailty of each part in the sample[a] [b]



a. Two outliers with frailties below -1.5 do not appear in figure 4.
b. Positive frailties indicate unreliable parts. Negative frailties indicate reliable parts.

Figure 4 depicts the estimated frailty for each part included in our sample. Note that frailty illustrates the likelihood that a part would fail more often than its counterparts for none of the reasons observed in the data. So, a negative level of frailty indicates a part that is more robust and less likely to fail, and a positive estimate indicates a more frail part than average.

What is important about this result is that it indicates that we may be able to identify these parts, maybe early enough, in order to prevent them from using up excessive amounts of maintenance resources. In the next section, we investigate this possibility.

Finding *"lemons"*

A reliable part would give us a large number of FH with a relatively low number of repairs during its lifetime. This ideal part would give us many service hours and at the same time consume a relatively low level of maintenance resources.

However, our results in the previous section indicate that there is heterogeneity in the reliability of parts and, therefore, some parts are more reliable than others. In this section, we use cluster analysis to classify parts into different levels of reliability as early as possible during their life cycles. This analysis allows us to differentiate between the reliable parts and the *lemons*.

Cluster analysis

Cluster analysis is a statistical method used to assign sets of observations into a predetermined number of categories. This is an exploratory data tool that does not require a priori knowledge of the categories themselves or even the number of categories that should be classified.

K-means cluster analysis classifies data into k clusters by assigning data points to the cluster whose center, or centroid, is the nearest. The centroid is the average of all points in the cluster for all the data dimensions that describe the data. In our case, because of our focus on reliability, we use the FH metrics from the model above to classify our data into clusters.

The algorithm to perform the clustering is simple. After the analyst decides on the number of clusters, k, to classify the data into, the algorithm proceeds to randomly select k seed points and assigns each data point in the data to the nearest seed. Because the seeds are randomly selected, they seldomly produce the best classification. So, the algorithm proceeds to computes the centers for each of the seeded clus-

ters and reassigns the data. Then it recomputes new centroids and repeats the procedure again. The algorithm is repeated until convergence of the results.[7]

Note that the algorithm described above requires the analyst to specify the number of clusters into which the data should be classified. Without prior knowledge of what the number of clusters should be, one way to make this decision is to use the number of clusters that provides the most distinct clustering—where the data groupings are as different to each other as possible.

To conduct this analysis, we used the same FH variables included in the data described in the previous section.

Lemons and *peaches*[8]

We proceeded with our analysis by first running the clustering algorithm using all of the observations in our dataset. This allowed us to create a baseline of reliability clusters with all the information about the reliability of the part. We used these results as a benchmark to see whether we could produce reliable results with less information.

In our analysis, we use the Calinski & Harabasz pseudo-F index to find the optimal number of clusters. Larger values of this index indicate more distinct clusters. Our results using the whole dataset indicate that four clusters produce the most distinct classification of the observations.[9]

Because we are classifying observations into groups, the most transparent way to show these results is graphically. We use figure 5 to illustrate our results. Also note that, for clarity, figure 5 only shows the

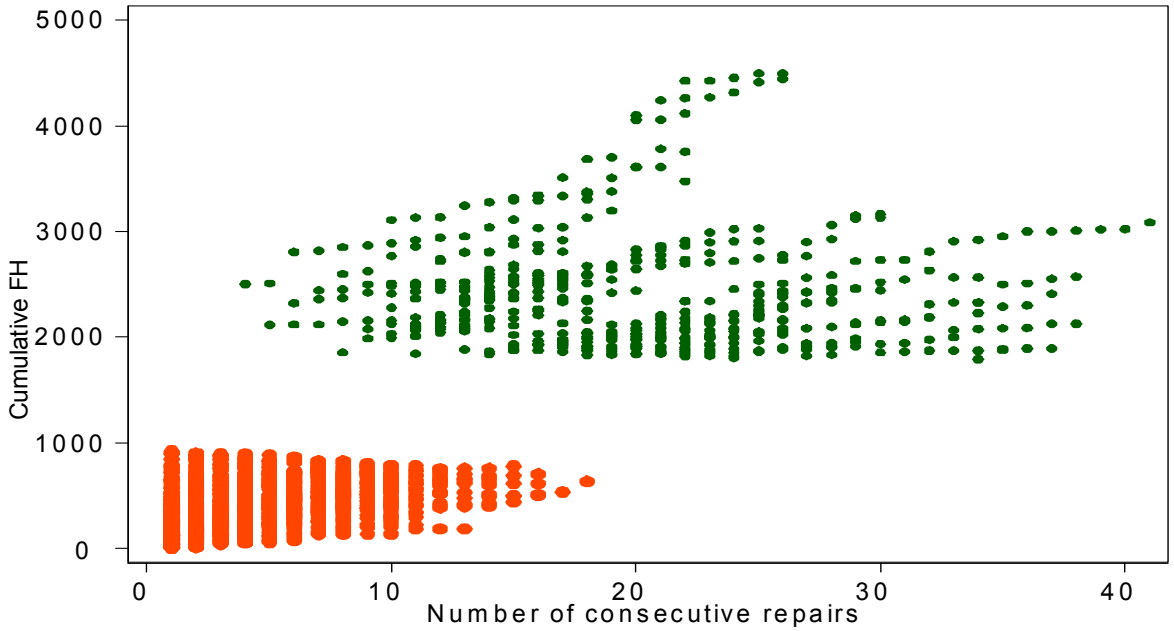
7. Convergence is achieved after consecutive iterations do not change the classification results by more than a pre-decided threshold.

8. Peach and lemon are common ways to refer to cars that are abnormally reliable or abnormally unreliable. These two terms are commonly used in the economics literature.

9. The value for the Calinski & Harabasz pseudo-F index was 7144.93 for four clusters, 6862.50 for three, and 7144.93 for five.

results for the most and least reliable of the clusters. This gives us a clear comparison of the reliability of two groups of parts.

Figure 5. Cluster analysis over the whole life history of the parts



The analysis shows a clear distinction between the parts that provide a relatively large number of FH with a relatively low number of repairs. These would be the *peaches*, which are shown in green. The *lemons* have few cumulative FH and a large number of repairs. They are shown in red.

Early detection of lemons

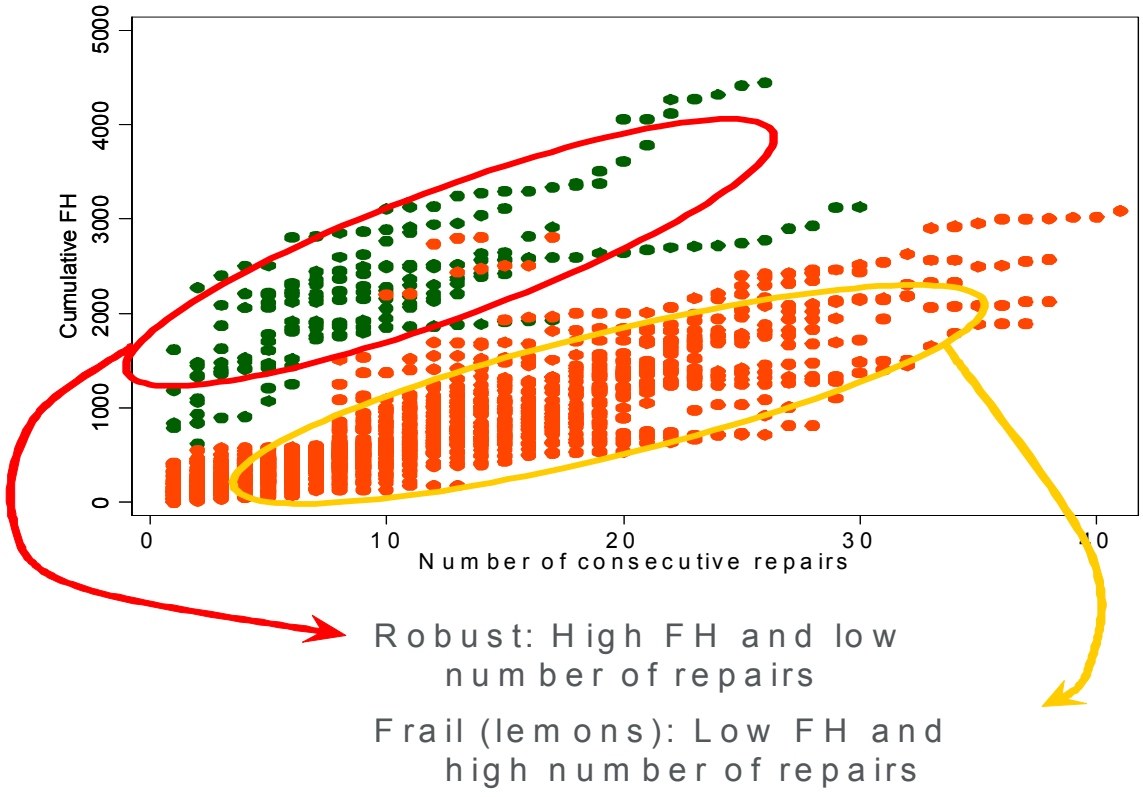
The next step in our analysis is to see if we can approximate this classification early in the life of the parts. Note that our analysis is exploratory and that, without direct cost measures to construct an optimal rule to differentiate lemons and peaches from the rest of the parts,

the results presented here should be considered notional and for illustrative purposes.

We follow the same procedure described above but with two important differences. First, we take the same number of clusters as given from the analysis of the complete dataset. So, for the rest of the analysis, we use four clusters as the natural way to classify these data. Second, we tried to use as little of the life history of the parts to produce reasonable results.

We iterated our results starting with the first repair cycle and continued on until we found the earliest repair cycle that would give us a reasonable approximation to the clustering obtained from the full dataset. We report our results in figure 6.

Figure 6. Reliability clustering at the seventh repair cycle



Implications of the analysis

There are several potential policy implications of this finding. First it may be cost efficient to retire lemons as soon as they can be identified. This would allow the Navy to free up maintenance resources as lemons require more than their fair share. If lemons are replaced with random parts, the replacements will increase the average reliability of the overall stock of parts. Moreover, the replacement part will be new and therefore more reliable than the parts they are replacing. All of which will increase average reliability and provide cost savings.

However, replacement of lemons is not the only alternative. A simple practice such as that of tracking the reliability of parts may provide the Navy with cost savings—a Best in, First Out (BIFO) practice could be implemented. This would mean that if we can identify the most reliable parts, these should be the first ones issued from supply when needed. This will ensure that it is the most reliable parts that are in the air—the most likely to last for a long time before needing their next repair. Other hybrid policies are also possible.

This page intentionally left blank.

Conclusions

Our analysis of the reliability of single parts over their life cycles shows that:

1. Analysis of MTBF by repair cycle can help us determine optimal retirement rules.
2. Event history analysis of the maintenance histories of parts can help us identify the factors that affect the reliability of a part. It also shows that it is possible to test whether parts of the same type inherently have different levels of reliability.
3. Analysis of single part maintenance histories allows for the early identification of lemons, which could help the Navy set cost efficient maintenance policies.

The analysis presented in this paper is a proof of concept designed to test whether analysis at the single part level has some advantages over more aggregated analysis. We have found that several results can lead to maintenance policies that can help the Navy lower its maintenance costs.

First, we find that the FH service time for a new APG-65 radar assembly declines quickly to level off after the first few cycles of repair. This result confirms that newer parts consume less maintenance resources and provide more up time. Further more, this suggest that, depending on the difference between the cost of repair and the cost of replacement of the part, there may be optimal part retirement rules.

Using an event history model, we also found several factors that affect the reliability of parts. One important finding is that the calendar age of a part is not an important determinant of the reliability of parts. Employment in terms of hours of service, but not sorties, are a better metric for the *age* of a part.

We also find that cannibalizations decrease the reliability of a part; however, we do find a positive effect of removals on reliability. This indicates that even operational parts may benefit from a visual inspection if thought to be faulty.

We found differential effects of sorties and FH on different TMSs for the reliability of parts. We also found that there is some evidence that BCMed parts become more reliable after depot repairs—we cannot claim statistical significance, however.

Lastly, our model allowed us to test our hypothesis that parts of the same type may have different levels of reliability for unexplained, unobserved or unobservable reasons. We found statistically significant evidence that parts can be more or less reliable than the average part consistently over their life times.

We used cluster analysis to identify low reliability parts, or lemons, as early as the seventh repair cycle. We note that many parts last beyond 20 repairs and some last as long as 40.

Based on this analysis, we propose different policy alternatives, such as early retirement of lemons. Another policy may be Best In, First Out (BIFO), in which the parts with best forecasted reliability after a repair are always the first in line for further use.

We note some limitations to this study. First, because it is a proof of concept of limited scope, we do not have reliable replacement and repair cost data to be able to set optimal policy rules for the retirement of lemons. Second, we focused solely on one particular part, the APG-65 radar assembly, but these methods could be applied to other types of spare parts.

Access to cost data for this and other parts would allow us to investigate when specific retirement rules would be optimal and to identify where such rules are never optimal. These data would also allow us to find the optimal time to identify lemons, balancing the probability of misidentification at an early point in the life of a part with the cost of keeping a lemon in stock for longer.

References

[1] Robert Levy. *The Effect of Cannibalization and Other Maintenance Actions on F-14A Failure Rates* (U), Sep 1988 (CNA Research Memorandum 2788016200/Final)

[2] Robert Levy. *The Effect on Failure Rates of the Removal and Reinstallation of Avionics Components* (U), Mar 1989 (CNA Research Memorandum 2788020100/Final)

[3] Sam Kleinman. *The Effects of Aircraft Conversions on Reliability* (FOUO), Mar 1984(CME 0583193700/Final)

[4] James Jondrow et al. *The Effect of Aging Equipment on Depot-Level Repair of Aircraft Components* (U), Mar 2002 (CNA Research Memorandum CRM D0004643.A2/Final)

[5] Brent Boning and Keith Brown. *Modeling AVDLR BCMs Using Airframe Age and Flying Hours* (U), Nov 2006 (CNA Research Memorandum D0014782.A2/Final)

[6] Brent Boning, Leopoldo Soto Arriagada, and Craig Goodwyn. *Causes of the Drop in AVDLR Consumption* (U), Jan 2008 (CNA Research Memorandum D0016735.A4/1Rev)

[7] Robert Levy. *Distinguishing the Effect on Failures of Changes in Sortie Rate and Sortie Length* (U), Mar 1986 (CNA Research Memorandum 2786004100/Final)

[8] David Boodman. *Reliability of Air Intercept Radars* (U), May 1952 (CNA OEG Study 10 1000048000/Final)

[9] Dean Follmann and Matthew Goldberg. *Distinguishing Heterogeneity From Decreasing Hazard Rates* (U), Dec 1986 (CNA Research Memorandum 2786026000/Final)

[10] David Cox. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society. Series B (Methodological)* Volume 34, issue 2, 1972: 187-220

[11] Jerry Hausman. "Specification Tests in Economics." *Econometrica* Volume 10, issue 46, 1972:1251-1271

[12] Janet Box-Steffensmeier, Suzanna De Boef, and Kyle Joyce. "Event Dependence and Heterogeneity in Duration Models: The Conditional Frailty Model." *Political Analysis* Volume 15, 2007: 237-256

[13] Janet Box-Steffensmeier, and Suzanna De Boef. "Repeated Events Survival Models: The Conditional Frailty Model." *Statistics in Medicine* Volume 24, 2006: 3518-3533

List of figures

This page intentionally left blank.

List of tables

This page intentionally left blank.