Reliability Modeling of Field Data for Repairable Systems

David C. Trindade

David C. Trindade, Ph.D.
Bloomenergy
6005 Assisi Court
San Jose, CA 95138 USA
Internet (e-mail): dave@trindade.com

SUMMARY & PURPOSE

The purpose of the tutorial is to illustrate with several real-world examples the power of statistical modeling to provide insight into product reliability performance of repairable systems and thereby help identify factors contributing to reliability. Also, we show how incorrect modeling of repairable system field data using analysis methods for repairable units can mislead and result in incorrect actions to remedy developing reliability issues.

David C. Trindade, Ph.D.

Dr. David Trindade is the Chief Officer of Best Practices and Fellow at Bloomenergy. Previously he was a Distinguished Principal Engineer at Sun Microsystems. He has been a Senior Fellow (AMD) and a Senior Director of Quality, Reliability, and Applied Statistics (AMD, Phoenix Technologies, General Instruments, and IBM). His fields of expertise include reliability, statistical analysis, and modeling of components, systems, and software; and applied statistics, especially design of experiments (DOE) and statistical process control (SPC). He is co-author (with Dr. Paul Tobias) of the book Applied Reliability, 3rd ed., scheduled for publication in 2011. He has a BS in Physics, an MS in Statistics, an MS in Material Sciences and Semiconductor Physics, and a Ph.D. in Mechanical Engineering and Statistics. He has been an adjunct lecturer at the University of Vermont and Santa Clara University. In 2008, he was the recipient of the IEEE Reliability Society's Lifetime Achievement Award.

Table of Contents

1. Introduction 1
2. Reports and Investigation of Field Failures 1
3. Failure Mode Identified 1
4. Field Failure Data 1
5. Modeling Repairable Systems 2
6. Application of the HPP Model 3
7. Cause Implications 3
8. Physical Mechanisms/Remediation 3
9. Recurrence Analysis 3
10. Summary 5

1. INTRODUCTION

A repairable system, as the name implies, is a system which can be restored to an operating condition in the event of a failure. The restoration involves any manual or automated action other than replacing the entire system. Common examples of repairable systems include computer servers, network routers, printers, automobiles, locomotives, etc. Although repairable systems are common, the techniques for analyzing repairable systems are not as prevalent as those for non-repairable systems.

This tutorial provides several real-world examples of the power of statistical modeling of recurrence data to provide insight into product reliability performance and thereby help identify factors contributing to reliability issues. Also, we show how incorrect modeling of repairable system field data using analysis methods for non-repairable units can mislead and result in incorrect actions to remedy reliability issues.

1.1 Notation and Acronyms

ARR	annualized recurrence rate
NTF	no trouble found
MTBF	mean time between failures
MCF	mean cumulative function
MLE	maximum likelihood estimation
RR	recurrence rate
ROCOF	rate of occurrence of failures
HPP	homogeneous Poisson process
SER	soft error rates

2. REPORTS AND INVESTIGATION OF FIELD FAILURES

In 1999 a large manufacturer of servers began experiencing field failures in a new product line. The failures were sudden, unexpected, and could cause the system to shut down abruptly (referred to as a "panic"). Engineers spent considerable efforts to restore systems to operation and to prevent recurrence. Boards experiencing a failure were replaced and returned to the company for analysis. Extensive data logging of conditions at the time of the failure were recorded and reviewed.

It is helpful to understand the physical characteristics of a system board in a server. The boards are approximately 2'x2' in size and weight around 30 pounds. The typical cost of a board was about $100,000. There are thousands of pins involved in a board connection to a system chassis. So removal and replacement of a board is not a trivial matter and must be done carefully, especially to avoid bent pins. The boards that experienced a failure were removed and returned for analysis to the factory. Damage in transit was not uncommon. After analysis, over 95% of the returned boards were classified as "no trouble found" (NTF), that is, the failure was not reproducible and no physical evidence of the failure could be found. The boards were fully functional. Also, 5% of the boards were often damaged in transit, which added to repair costs.

There were wide-ranging actions to identify the cause of the failures. There was extensive stressing and testing of new and returned boards in systems. There was physical failure analysis of returned boards. In the field, there was replacement of failed boards with new boards. There were visits by engineers and management to customer sites for observations of the systems and environment. There were field environmental measurements (temperature, humidity, etc.). There was data logging activity using diagnostic software runs. There was consultation with suppliers and frequent reviews and update meetings of teams of engineers and management.

3. FAILURE MODE IDENTIFIED

After months of research, the failure mode was identified as parity errors in e-cache (external, L2) SRAMS as the problem location but determining the exact cause was elusive. See Figure 1.

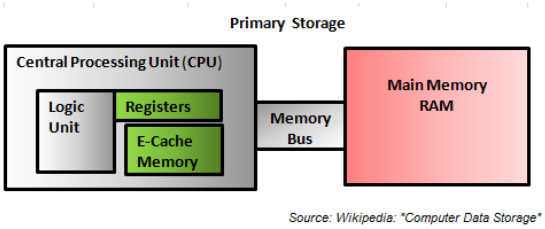

Figure 1. Primary Storage in Servers

There was slow progress in isolating the causes of the failures, which continued in the field. The service engineers worked diligently to diagnose failures and restore the systems to operation. The costs of field repairs escalated. Customers demanded prompt resolution.

4. FIELD FAILURE DATA

A data collection team was formed to collect data on field failures. Reliability data was collected from major customers' datacenters. The importance of acquiring time dependent data was emphasized. The data showed that some customers experienced no failures. Other customers saw high levels of failures for the same systems. One customer in a concrete vault below ground saw no failures. Other customers in high altitude environments (e.g., observation stations) had more frequent failures. Was altitude or barometric pressure a factor?

Additionally, the data showed application dependence on the rate of occurrence of failures (ROCOF). In the same datacenter, customers running different applications on identical systems experienced widely different ROCOFs. See Figure 2., which shows a the annualized repair rates (ARR) for four different applications running on 476 systems in a single datacenter.

Also, in the same datacenter, for identical systems running the same applications over the same time period, there could be systems with no failures, some with single failures, and some with multiple failures. See Figure 3 which shows the distribution of failures over 101 days on 48 systems running the same application.

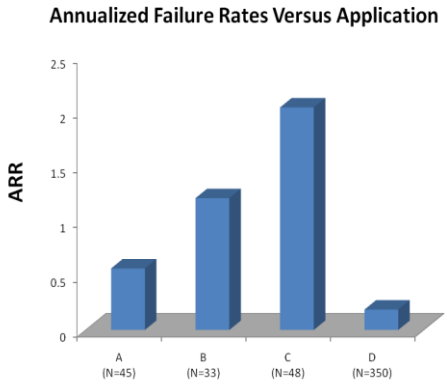

Figure 2. ARR versus Application in One Datacenter

Could statistical analysis and modeling of the data provide any insights into the cause? How could the application dependence be explained? Could the model agree with field behavior and predict future failures? Could we explain the distribution of failures across systems in a datacenter?

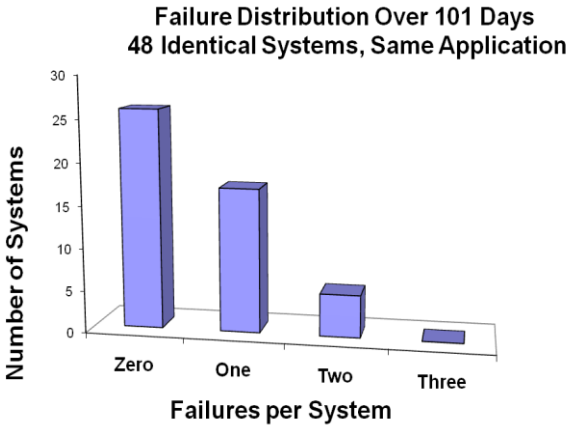

Figure 3. Failures Distribution Across Systems

5.MODELING REPAIRABLE SYSTEMS

There are two key reliability measures for repairable systems: times between repairs (interarrival times) and number of repairs over time. Reliability is a function of many factors, including basic system design, operating conditions, environment, applications, software robustness, types of repairs, quality of repairs, materials used, suppliers, and human behavior.

System age is the total running hours, that is the elapsed time on a system starting at installation turn-on. Age is often called the uptime. We must be careful to distinguish system age from times between failures (interarrival times) and device-hours or unit-hours.

An important property of repairable systems is that failures occur sequentially in time. If the times between successive failures are increasing, then the system reliability is improving, Conversely, if the times between failures are decreasing, then the reliability of the system is degrading. Thus, the sequence of system failures times can be very informative. If the times show no trend (relatively stable), the system is neither improving or degrading, a characteristic of what is called a renewal process.

In modeling system reliability, there are some critical questions. For a renewal process, the times between failures are independent and identically distributed (i.i.d.) observations from a single population. How can we verify such an assumption? In a renewal process, there is no trend. For a system, restoration to "like new," such as replacement of a failed component with one from same population, implies a renewal process (i.i.d.). There are statistical tests to check the assumption of a renewal process.[1]

Unfortunately, age related data is typically not available for systems. Field reliability data is often presented in terms of a mean time between failure, *MTBF*. It is much easier to count the numbers of failures in a given time period (e.g., one month) for a group of systems operating for that time period than it is to obtain the system installation dates to measure age and the time dependent history of the ages of failures. Are there other ways to model the field behavior?

For modeling, two variables are of key interest: $M(t)$ the mean number of repairs by time t, that is, the *MCF* and $T(k)$ the time to reach the k^{th} failure. For a renewal process, $M(t)$, the *MCF*, is also called the renewal function, which is the expected (or average) value of $N(t)$, the number of repairs by time t for a single system.

For a renewal process, the single distribution of failure times between repairs defines the expected pattern of repairs. Let Xi denote the interarrival time between the i^{th} and the (i-1) repair. The time to the k^{th} repair can be written as the sum of k interarrival times.

$$T(k) = \sum_{i=1}^{k} X_i$$

For example, if the first three interarrival times are 100, 150, and 75 hours, then the time to the third repair is 100+150+75 = 325 hours. Knowing the probability distribution (pdf) of Xi, we can theoretically find distributions for $N(t)$ and $T(k)$ along with $M(t)$ and the renewal or recurrence rate (*ROCOF*) $m(t) = dM(t)/dt$.

Suppose the interarrival times Xi are i.i.d. with exponential probability density function (pdf) having constant failure rate intensity λ, that is,

$$f(x) = \lambda e^{-\lambda x}$$

Then, we can show that $N(t)$ has a Poisson distribution with constant renewal rate intensity λ. The expected number of repairs in time t is λt. Note that λ is a rate (i.e., repairs/time) that is multiplied by time t to give the number of repairs by time t. Consequently, the probability of observing exactly $N(t) = k$ failures in the interval (0,t) is given by the Poisson distribution

$$P\left[N(t) = k\right] = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

We call this renewal process for which the interarrival times are exponentially distributed a *homogeneous Poisson process* (*HPP*).

For a *HPP*, the mean time between failures (*MTBF*) is constant and

$$MTBF = \theta = 1/\lambda$$

The expected number of repairs in time t is $M(t) = \lambda t = t/\theta$. The mean time to the k^{th} repair is $k/\lambda = k\theta$. We can rewrite the Poisson distribution for the *HPP* in terms of the *MTBF*, θ :

$$P\left[N(t)=k\right]=\frac{\left(t/\theta\right)^{k} e^{-t/\theta}}{k!}$$

By multiplying the calculated *HPP* Poisson distribution probabilities for a given failure rate or *MTBF* by the number of systems, we can estimate the expected distribution of failures across many similar *HPP* systems.

6. APPLICATION OF THE HPP MODEL

There were a total of 476 hosts in a large datacenter. By determining an overall failure rate or *MTBF* over the previous few months, we checked for the suitability of an *HPP* model that could predict over the next 101 days how many of the 476 systems would have no failures, one failure, two failures, and so on. This prediction was then compared against actual failure counts across all systems.

The model was in excellent agreement with observed results, confirming the *HPP* as shown in Figure 4. Comparison of *HPP* model for each application also showed excellent fit.

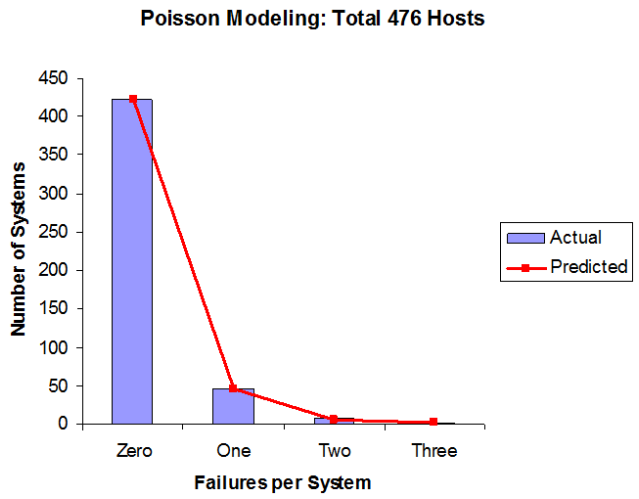

Figure 4. Comparison of HPP Model to Actual Data

7. CAUSE IMPLICATIONS OF MODEL

Since the results were consistent with a *HPP*, the implication was that the failure behavior for any system in the datacenter derived from a renewal process with a constant failure rate. Constant failure rates result from a constant source. There was no physical damage to the *SRAM* by the cause. The "good as new" assumption for a renewal process seemed valid. Failure rates were also determined to vary with altitude. Results confirmed that the only plausible source was radiation from cosmic rays causing single bit parity errors in the e-cache memory. Without error corrective actions, failures would occur and panic the systems.

8.PHYSICAL MECHANISM/REMEDIATION

The radiation environment includes alpha particles and high and low energy cosmic rays. Low energy cosmic rays can cause ^{10}B fission in boron-doped phosphosilicate glass (*BPSG*) dielectric layers of *ICs*, which generate electron-hole pairs in the silicon, disturbing memory bits, and resulting in soft errors. The factors that impact soft error rates (*SER*) are complexity, density, lower voltages, higher speeds, and lower cell capacitances. The susceptibility to soft error rates for *DRAM* and *SRAM* has increased with reduced dimensions (higher densities) and lowered operating voltages of advancing technology.

In Read-Write activity, the server writes to e-cache memory. Memory in e-cache can be saved to permanent memory. If a cosmic ray causes a parity error to occur in e-cache and an attempt is made to read data in e-cache or to write it to main memory, the parity error will be detected and the system will panic to prevent data corruption.

An effective solution was to incorporate mirroring, where every byte is duplicated and stored in two locations in *SRAM* along with a parity checker built into the *SRAM*. The confirmation of the solution is graphically shown in Figure 5, where a flat line indicates the introduction of the mirrored *SRAMS* and no subsequent e-cache failures over time.

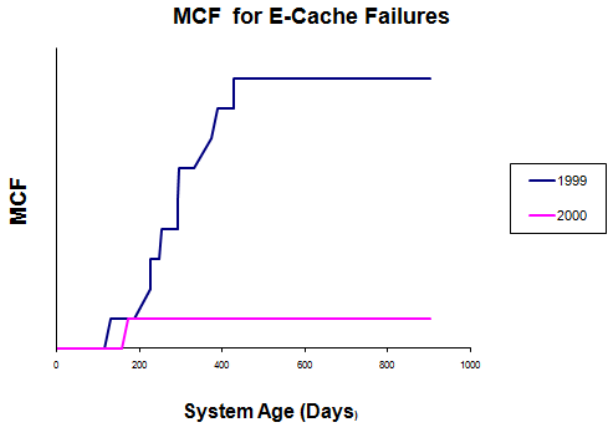

Figure 5. Confirmation of Solution

Application dependence can now be explained. If an application writes often to memory but does not read frequently, an e-cache error can be overwritten before a read cycle sees the error. Imagine an application updating minutes used by a cell phone user. Consequently, the failure rates will be low. If an application reads frequently, then e-cached errors will be detected quickly and cause failures. The failure rates will be high.

Until *SRAMS* were replaced with mirrored *SRAMs*, "Best Practices" were defined based on the modeling. Instead of removing a failed board, the most effective action was simply to reboot the system. No physical damage had occurred and the probability of a hit by a cosmic ray was purely random. In

addition, the costs of replacing boards and subsequent damage to the boards or systems (e.g., bent pins) could be avoided.

9. RECURRENCE ANALYSIS

The usual assumptions we make for non-repairable components is that the times to failure are a true random sample from a single population. Consequently, the observations are independent and identically distributed. The implication is that individual failure times can be combined for analysis, neglecting any order of occurrence in the original data. [1,2,3] Are these assumptions valid for repairable systems?

Example 1

New Production Equipment: A new system used in manufacturing contained a single, replaceable board. Upon failure, repairs were made by replacing the failed board with a new board from the same population in stockpile. Engineers wanted to model the reliability of the system based on failure data obtained during the first 1000 hours of operation.

Repairs were done at system ages (in hours) 108, 178, 273, 408, 548, 658, 838, and 988. A dot plot of repair times is shown as Figure 6..

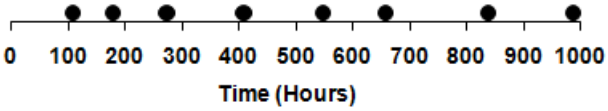

Figure 6. Dot Plot of Repair Times

The engineers analyzed the system data by taking the failure time for each replacement board, that is, the times between repairs, and treating those times as a group of independent and identically distributed (i.i.d.) observations arising from a single population of failure times. The actual order in which these times occurred (age of system at repair) was ignored. Analysis methods used were: Weibull probability plotting of data, parameter estimation, and model fitting. The times between repairs are called the interarrival times and are calculated in Table 1.

Repair Time (System Age)	Time Between Repairs (Interarrival Times)
108	108
178	178-108 = 70
273	273-178 = 95
408	408-273 = 135
548	548-408 = 140
658	658-548 = 110
838	838-658 = 180
988	988-838 = 150

Table 1. Interarrival Times for Weibull Analysis

For Weibull analysis, the order of interarrival times is not considered. The Weibull probability plot (Figure 7.) in *JMP* shows data points falling close to a straight line. The Weibull *MLE* parameters estimates were: characteristic life $\alpha \approx 136$ hours and a shape parameter $\beta \approx 4.3$. For the Weibull distribution, $\beta > 1.0$ indicates an increasing hazard rate.

Engineers concluded times between repairs followed a Weibull distribution. Of concern was that the estimated Weibull shape parameter, β, indicated an increasing "failure rate." The equipment engineers thus felt the machine needed to be brought down for additional repair and maintenance before "things got much worse".

Were these conclusions justified or misleading based on analyzing the boards as non-repairable components?

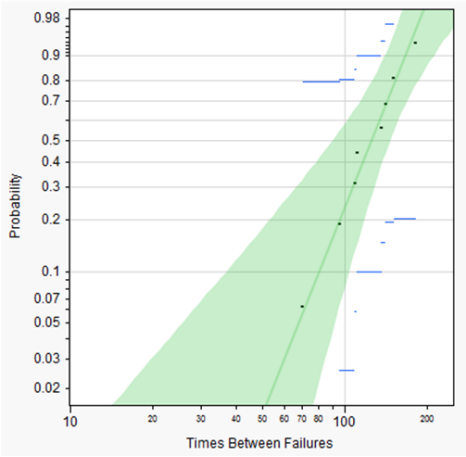

Figure 7. Weibull Probability Plot

Analyzing the data as a repairable system and plotting the times between failures versus the system age, we see (Figure 8.) that actual interarrival times are getting longer. Statistical analysis techniques confirm that the observed trend to longer times is significant. [2]

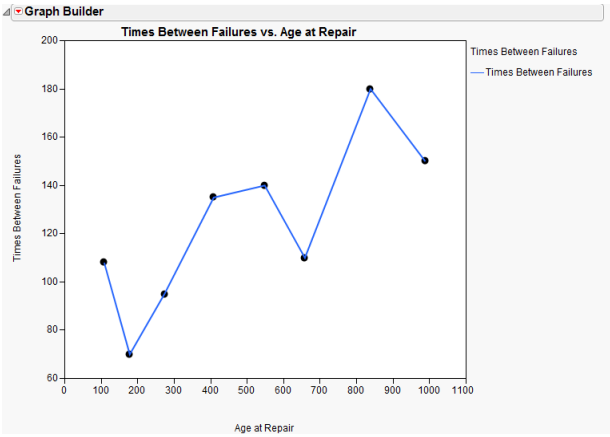

Figure 8. Plot of Times Between Failures Versus Age

Analysis of repairable system data using Weibull analysis methods for non-repairable systems produced misleading conclusions. Wrong interpretation was caused by the engineers' neglect of the occurrence order of failures in Weibull analysis. With correct analysis, engineers avoided expensive maintenance actions that were not necessary.

Example 2

Locomotive Valve Seat Replacements: Nelson (1995) [5] reports recorded valve seat replacements in locomotive engines. Treated as non-repairable components, a lognormal probability plot in *JMP* (Figure 9.) of the times between failures shows that the data fits a lognormal distribution very well.

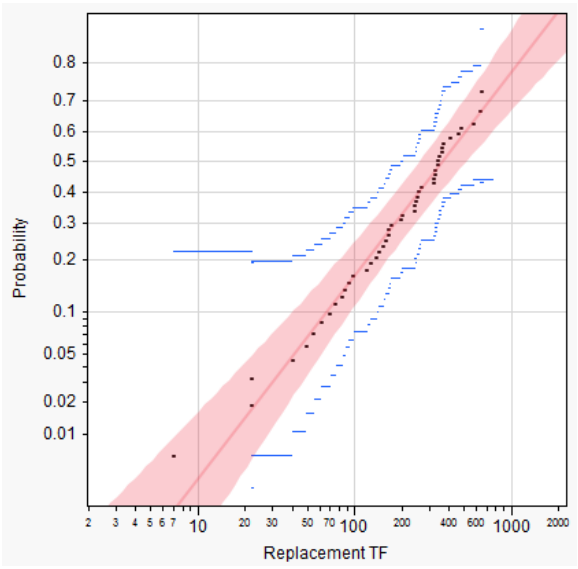

Figure 9. Lognormal Probability Plot of Locomotive Repair Times

Parameter estimates show hazard rate increasing early in life, peaking around 100 days, and then decreasing thereafter. See the hazard plot in Figure 10, created in *JMP*.

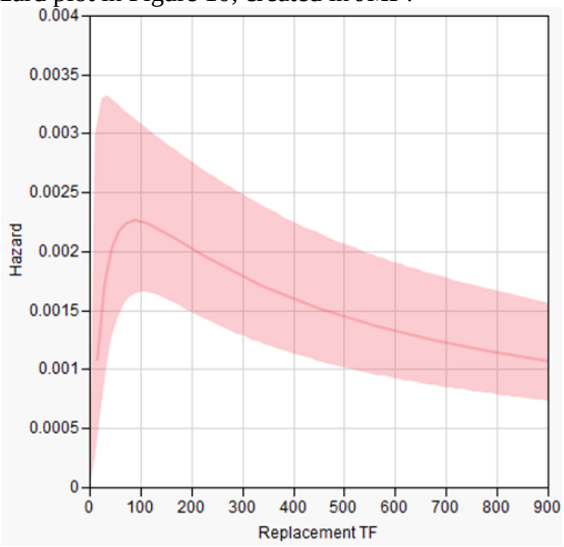

Figure 10. Hazard Rate for Lognormal Fit to Locomotive Repair Data

Analysis of valves on locomotive as repairable systems (See *MCF* [4] plot Figure 11.) shows repair rates sharply increasing at 500 days. Is wearout occurring? This result totally contradicts the original conclusion of a decreasing hazard rate. By using the times between failures irrespective of the age of the locomotives, we falsely cause the failures to appear as if they occur early when in fact the failures are occurring later in the life of the product.

10.SUMMARY

Field failures represent significant inconvenience to customers. Field failures remediation efforts are costly to system manufacturers. Complex systems make identification of causes difficult and challenging. Statistical analysis and modeling can provide valuable insights into causes.

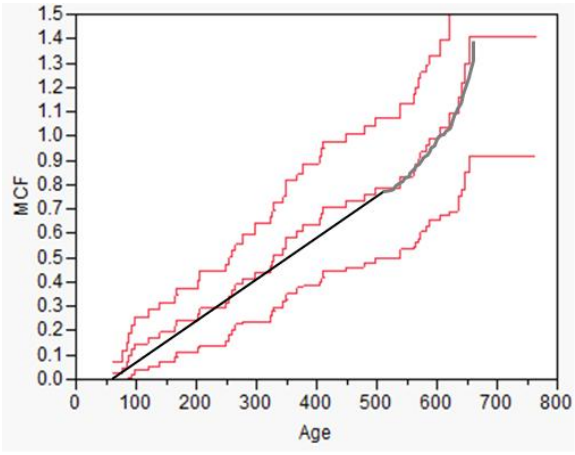

Figure 11. MCF Rate for Locomotive Repair Data

Analysis of repairable system data using non-repairable analysis methods can produce misleading results. For repairable systems, analysis of the distribution of failures across systems and the time order in which failures occur can provide valuable information. For individual systems, a cumulative plot shows the repair history graphically. For multiple systems, the *MCF* plot can reveal trends in the collective behavior of a group of systems.

11. REFERENCES

1. W. Nelson, *Recurrence Events Data Analysis for Product Repairs, Disease Recurrences and Other Applications*, ASA-SIAM Series in Statistics and Applied Probability, 2003.
2. P.A. Tobias, D. C. Trindade, *Applied Reliability*, 3rd ed., Chapman and Hall/CRC, 2011.
3. W.Q. Meeker, L.A. Escobar, *Statistical Methods for Reliability Data*, Wiley Interscience, 1998.
4. D. C. Trindade, Swami Nathan, Simple Plots for Monitoring the Field Reliability of Repairable Systems, *Proceedings of the Annual Reliability and Maintainability Symposium (RAMS)*, Alexandria, Virginia, 2005.
5. W. Nelson, "Confidence Limits for Recurrence Data-Applied to Cost or Numbers of Product Repairs", Technometrics, 37, 147-157, 1995.