Report No. 182 / UILU ENG-98-4017

Statistical Modeling of Rainflow Histograms

by

Jason Scott Roth

B.S., The Pennsylvania State University, 1996

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Mechanical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1998

Urbana, Illinois

ABSTRACT

The ability to quantify usage in durability is an important idea, because despite the effectiveness of any fatigue model, if the loads are not representative of the real in-service loadings, that model will inevitably produce results that are uncharacteristic of the actual expected life. The modeling of variable amplitude loading histories is a difficult proposition because the magnitudes of the loadings are random in nature, but when those load histories are placed into a rainflow counted histogram, the data in those histograms forms a pattern which can be statistically modeled.

The models presented involve the use of non parametric density estimation, and in particular the kernel method, to quantify the underlying density in the histograms. Then Monte Carlo simulations are conducted to predict the future results of two different problems. The first problem is given a loading history, model the loadings that could be anticipated if that durability test were to be conducted over longer periods of time. The other problem is given a set of loading histories, from that set, model the loadings that could be expected from the more extreme usages in a similar but much larger set of loading histories. The theory of these models, as well as procedural instructions for implementing the models, and some of the results of these models are presented.

TABLE OF CONTENTS

Page

1. INTRODUCTION

A major factor in the design of vehicular components is the anticipated severity in the usage of the component. In many situations, the reliability of the component reflects directly upon the safety of the vehicle, as well as the overall quality of the vehicle in the eyes of the consumer, so designing for the anticipated service is of extreme importance. Quite often, in order to determine the loads that a component must be designed to withstand, extensive testing is conducted using a sampling of the population as a test base, and hundreds of sets of load data are measured.

Typically, a component will be designed with the premise that it will have a life of at least x usages for some percentage of the population of users, let's say for instance 99% of the population, so one of the keys of design is determining what the loading associated with that 1% most extreme usage is.

Another factor in the design process is in verifying that the design life of the component is actually achievable for a large majority of the population. However, it is inconvenient and expensive to test them to failure when they are in service because a typical component is designed to have a relatively long life. Instead, it would be convenient to conduct in-service testing on the part for a period of time much shorter than the expected life, and then use the loads that were found to occur in that period of time to predict the life of the component.

These two ideas for determining both the extended in-service loads and the single most extreme loading scenario that a component will be subjected to, are the basis of this project.

Due to the varying service conditions that a vehicle can be placed under, loading histories are random processes, and in time-series format, as in Figure 1.1a, loading histories have the same general characteristics for an individual driving route, but are not predictable. For example, the loads shown in Figure 1.1a represent 10 passes over a series of bumps. The magnitude of load for any one pass cannot be predicted in advance. When the data from a loading history is rainflow-counted and placed into a histogram format, such as the from-to format depicted in Figure 1.1b, the distribution of the data takes on a definite shape and size, and forms an inherent pattern.

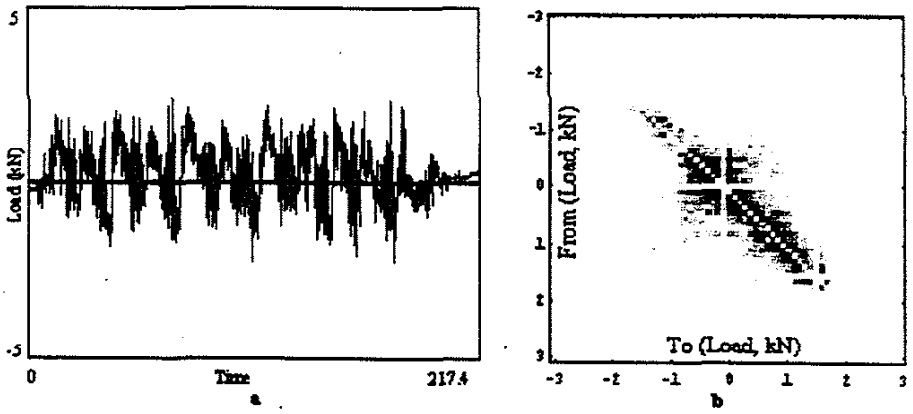


Figure 1.1 a) Loading history in time-series form, **b)** the same loading history in a rainflow counted from-to histogram.

It is the aim of this research to use statistical methods to model the behavior of rainflow histograms in order to attempt to solve the following two problems: a) given a single rainflow histogram, model the statistical properties of that histogram to extrapolate it to a much longer time period; and b) given a small set of durability data, where each individual piece of the data set is from a distinct user, model the underlying statistical

properties of that data set to predict the most severe histogram for a much larger set. That is, from a number of individual histograms, predict a single histogram that represents an extreme user. Statistical models are constructed to attempt to solve these two problems, and preliminary results of those models are given.

2. BACKGROUND

As was mentioned in the introduction, this model seeks to use the inherent statistical properties of rainflow-counted histograms in order to extrapolate data obtained from in-service testing. In order to accomplish this goal, statistical non-parametric density estimation is utilized, but before the theory of the model can be presented, it is essential to understand the mathematical foundations on which the model is built. The following sections give enough background information on the statistical techniques used in the modeling procedures so that an average engineer can comprehend the theory of the model, and replicate the results obtained herein.

2.1 STATISTICAL PRELIMINARIES AND DEFINITIONS

The entire basis of this model lies in the fact that the magnitudes involved in variable-amplitude loading histories are random; that is, the actual values are not precisely known. At the present time, to make sure that loading histories used for design purposes are realistic, those loading histories must consist of data taken from actual tests. But in order to assure that the data being used is representative of the real population, that test data must be collected over long periods of time and from many different sources, requiring extensive testing schedules. In the model outlined in this thesis, the uncertainties of the loadings are estimated by probability distributions, and those probabilities are used to statistically model the loadings. Therefore, it is important to have a good understanding of the theory of probability.

2.1.1 Definition of Probability

Probability is defined as the likelihood of the occurrence of a particular event, relative to the likelihood of occurrence of any other possible event. The probability of a discrete event, x, is symbolized $P(x)$.

In order to quantify the probability of all possible values of a random event, a probability distribution is used. If X is a random variable, its probability distribution can be described by its cumulative distribution function (CDF):

$$F_X(x) \equiv P(X \le x) \text{ for all } x \tag{2.1.1}$$

Note: A standard notation is to denote a random variable with a capital letter, and its value with the corresponding lowercase letter

The random variable, X, is called a *continuous* random variable when x can assume any value in a range $a \le x \le b$, while on the other hand, X is called a *discrete* random variable if x can only assume certain discrete values in a range $a \le x \le b$. In the model presented here, the random loads are assumed to be continuous random variables, because load values are not distinct, but can potentially take on one of any infinite number of magnitudes.

For a continuous set of random variables, X, the CDF is the following:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(\xi) d\xi \,, \tag{2.1.2}$$

where $f_X(x)$ is called the probability density function (PDF). The PDF has the property that $\int_{-\infty}^{\infty} f_X(\xi) d\xi = 1$, or in other words, $F_X(x = \infty) = 1$.

The PDF describes the distribution by quantifying the probability that values of X will be contained in the interval $(x, x+dx]$ as follows:

$$P(x < \mathrm{X} \leq x+dx) = f_X(x)dx\,. \tag{2.1.3}$$

By knowing the PDF of a distribution, it is a simple procedure to determine a few characteristic properties of that distribution.

2.1.2 Statistical Characteristics of Probability Functions

The most simple descriptor of a distribution is its average, or mean value, and for a continuous variable, X, and a corresponding PDF, $f_X(x)$, the mean value, designated as $E(\mathrm{X})$, is defined as is:

$$E(\mathrm{X}) = \int_{-\infty}^{\infty} x f_x(x)dx \tag{2.1.4}$$

The second most important descriptor of a distribution is its variance, $Var(\mathrm{X})$, which is defined as:

$$Var(\mathrm{X}) = \int_{-\infty}^{\infty} (x - E(\mathrm{X}))^2 f_X(x)dx \tag{2.1.5}$$

When given a set of data as we are with loading histories, rather than a theoretical distribution, as has been described in this section, the procedures to calculate the characteristic values of the data set are slightly different than those described here. Those techniques, called descriptive statistics, are described in an upcoming section.

Since a basic foundation of probability theory has been presented, we now take a look at some well-known probability density functions, and their corresponding characteristics.

2.2 USEFUL THEORETICAL PROBABILITY DISTRIBUTIONS

In statistics, there are numerous continuous probability distributions to choose from, and in this model, two distributions play an integral role. Those two distributions are the normal (or Gaussian) distribution and the Weibull distribution, and are described below.

2.2.1 The Normal Distribution

The normal distribution is by far the most widely used distribution in statistics and probability, and with good reason. The normal distribution simply represents a distribution of data where the density tails off exponentially in both directions about the mean value. The PDF for the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} Exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty < x < \infty \tag{2.2.1}$$

and the CDF is

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} Exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \tag{2.2.2}$$

The parameters μ and σ are the mean and standard deviation, respectively, of the distribution, and quite often the normal distribution *f(x)* is designated as $N(\mu,\sigma)$. An oft used normal distribution is that with a mean of zero and a standard deviation equal to one, N(0,1), and this distribution, shown below in Figure 2.1, is called the *standard* normal distribution.

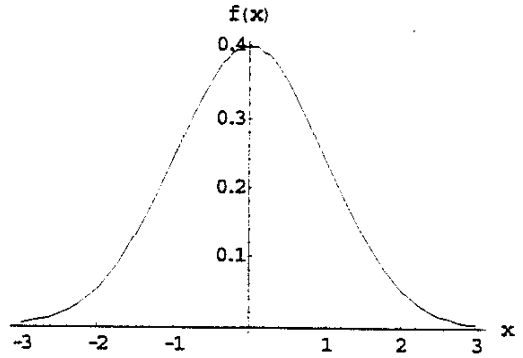


Figure 2.1 Standard Normal Distribution, N(0,1)

The log-normal distribution is a special case of the normal distribution where x in the distribution is replaced by $ln(x)$.

2.2.2 The Weibull Distribution

The Weibull distribution is used very frequently to characterize reliability, and is an extremely flexible distribution, making the Weibull distribution an attractive tool in statistical modeling [2]. The Weibull distribution is described mathematically as

$$f(x) = \frac{\beta}{\eta^{\beta}} x^{\beta-1} Exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right], \; x \geq 0. \qquad (2.2.3)$$

In the normal distribution, the shape parameters are the mean and standard deviations of the data, but in the Weibull distribution, the parameters β and η are the shape parameter and the scale parameter, respectively.

The flexibility of the Weibull distribution is exemplified in Figure 2.2 below, where three such distributions are plotted. In the distributions in this figure, η=1, and β was allowed to vary.

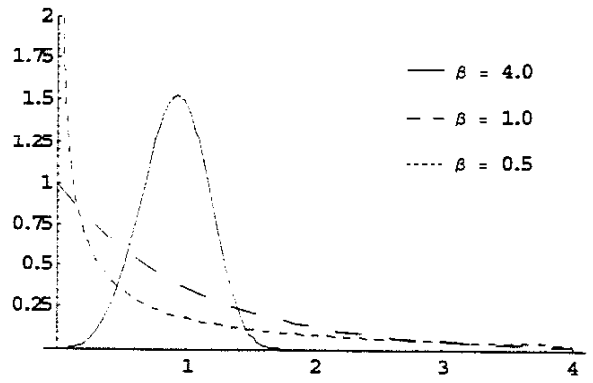


Figure 2.2 Three Weibull Distributions with η=1, and β as Shown

In the model presented in the next chapter, the Weibull distribution is used as a tool to model and extrapolate the cumulative exceedance diagram, and is also used to estimate the damage done in extreme loading cases.

Thus far, the distributions discussed have been theoretical in nature, but the problem of determining the distribution of an actual sample of data still exists, and this problem is solved by using *density estimation* techniques.

2.3 Density Estimation

For a given set of data taken from a continuous population, X, there are countless ways to construct a probability distribution of the data. All of these density estimation techniques, however, can be categorized as being one of two general classes of density estimates: parametric or non-parametric.

2.3.1 Parametric Density Estimation

In parametric density estimation, an assumption is made that the given data set will fit a pre-determined theoretical probability distribution, such as one of the previously

described distributions. In general, a majority of data sets will approximate some known theoretical distribution, and using parametric density estimation on those simply involves finding the shape parameters of the distribution. In order to use a theoretical density function to characterize a data set, it is often necessary to understand some characteristics of that data set, and to do that we use *descriptive statistics*.

2.3.1.1 Descriptive Statistics

The two most common properties used to characterize a data set $\left[x=(x_1,x_2,\ldots x_n)\right]$ are the sample mean, $\bar{x}$, and the sample variance, s^2, defined as follows:

$$\bar{x}=\sum_{i=1}^{n}\frac{x_i}{n}, \text{ and} \tag{2.3.1}$$

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 . \tag{2.3.2}$$

The sample mean can be thought of as the centroid of the distribution of data, and its value, $\bar{x}$, has been shown to be the best estimator of the mean, μ, of the normal distribution when the data is normally distributed.

The sample variance is a value that describes how the data is distributed about the mean, and again, if the population is normally distributed, its value, s^2 is the best estimate of the variance parameter, σ^2, used in the theoretical normal distribution [1,2].

From the sample mean and the sample variance, the sample standard deviation and the sample coefficient of variation (COV) can be calculated:

$$S_X=\sqrt{s_X^2} \text{ and } COV=\frac{S_X}{\bar{x}} . \tag{2.3.3}$$

Both of these are good quantifiers of the variability of a data set, and both are used in the statistical modeling procedures described in the next chapter.

The problem with parametric density estimation comes in the cases where the data does not approximate any known theoretical distribution. It is in these cases where *non-parametric density estimation* techniques are necessary.

2.3.2 Non-Parametric Density Estimation

Contrary to parametric density estimation, the non-parametric class of density estimators make no assumptions about the distribution of the data set. Because of its generality, the non-parametric class would appear to be superior, since it determines the density based on the data itself, rather than basing the density on a theoretical distribution, as the parametric estimators do. But as we will see, with increased flexibility comes the problem of increased complexity.

One of the first and most widely used non-parametric density estimators is the histogram. To create a histogram, all that is necessary are an origin, x_0 and a bin width, h, and the bins of the histogram are defined to be the intervals $[x_0 + mh, x_0 + (m+1)h)$ for integers m. The histogram is then constructed by counting the number of points in the data set that fit within the limits of each bin. The major pitfall of histograms is that the resulting frequency distribution is strongly dependent on the bin width [3,4]. As the bin width is increased, the resulting frequency distribution is smoothed out, and as the bin width is decreased, the distribution is discretized.

Every non-parametric density estimation technique involves the choice of a parameter similar to the bin width, and the general term for this parameter is the "smoothing parameter". The problem of choosing a valid smoothing parameter is an inherent problem in non-parametric density estimation, and will make a major difference between a good density estimate and a poor density estimate. Because of this smoothing parameter problem, when a theoretical distribution can be inferred from a set of data, it is most often preferable to use parametric density estimation, but there are many cases where parametric density estimation is just not possible. The problem of describing the density of a rainflow histogram, as is done in this model, is one of those cases where non-parametric estimation is necessary.

There are several more descriptive methods of constructing a non-parametric density estimation than the histogram, and the method used in this model to calculate the density is the *kernel method.*

2.4 THE KERNEL METHOD

The method which plays a vital role in the procedures associated with the rainflow histogram density estimation is a multi-dimensional kernel method, but before this multi-dimensional method is explained, it is instructive to introduce the kernel method in one-dimension.

2.4.1 The 1-D Kernel Method

A kernel in one dimension is simply a function, K(x), which satisfies the following equality:

$$\int_{-\infty}^{\infty} K(x)dx = 1, \qquad (2.4.1)$$

so a kernel is in essence a PDF. The kernel can take on any shape, as long as it satisfies the above equality. Choosing a kernel to use for density estimation is a subjective decision, but in most cases, statisticians will use a symmetric probability density function, and the standard normal kernel is the most frequently used kernel. Listed in Table 2.1, and plotted comparatively in Figure 2.3 are three popular kernel functions.

Table 2.1 Three popular 1-D kernel functions.

Normal kernel	$K(x) = \frac{1}{\sqrt{2\pi}} Exp\left[-\frac{1}{2}x^2\right]$	for all x
Epanechnikov kernel	$K(x) = \frac{3}{4}\left(1-\frac{x^2}{5}\right)$	for $\lvert x\rvert < \sqrt{5}$
	0	all other x
Biweight kernel	$K(x) = \frac{15}{16}\left(1-x^2\right)^2$	for $\lvert x\rvert < 1$
	0	all other x

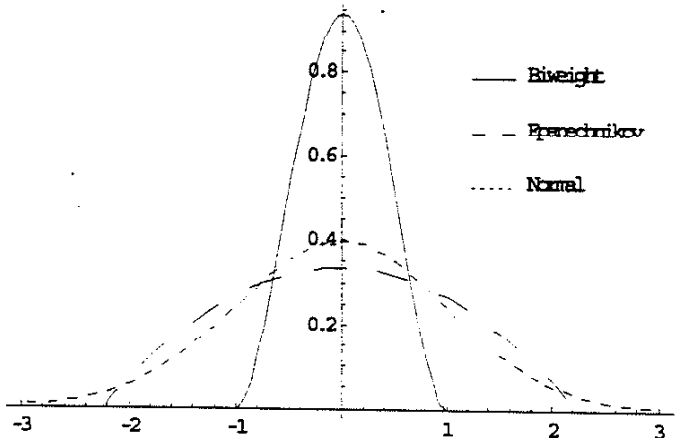


Figure 2.3 Comparison of three different kernel functions.

Just as in the earlier discussion of histograms, there is a smoothing parameter associated with kernel density estimation. In the case of the histogram, the problem was in determining how large to make the bin width, and similarly, with kernel estimation the problem lies in determining what the width of the kernel should be. The width of the kernel is called the *kernel bandwidth*, and this bandwidth must be taken into account when summing the kernels. Actually, it is widely agreed that it is the choice of bandwidth, not the choice of the kernel function, that is the major factor in determining whether or not a given density estimation procedure is valid [4]. That is, the width of the kernel is generally more significant than is the shape of the kernel. However, determining the optimal bandwidth is often very tricky, and this will be discussed further in the following section on the 2-D kernel method.

Once the kernel has been chosen, a density estimate is constructed by placing a kernel centered on each observation in the 1-D data set, and then summing the value of the individual kernels to obtain a density estimate. The following is the mathematical definition of the density function, or probability distribution function (PDF), calculated using the 1-D kernel method

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathrm{K}\left(\frac{x - \mathrm{X}_i}{h}\right). \qquad (2.4.2)$$

To calculate f at each x, the magnitudes of the kernels corresponding to points X_i, are summed at each point, x

For illustrative purposes, a simple example of the kernel method is shown below in Figure 2.5. In this example, a sampling of data points corresponding to a normal distribution was constructed, a kernel function was placed at each data point, and the

density function was obtained by summing those kernels. While this is a very simple example, it illustrates the power of kernel density estimation, because while no assumptions had to be made about the data set, the result is clearly a normal distribution.

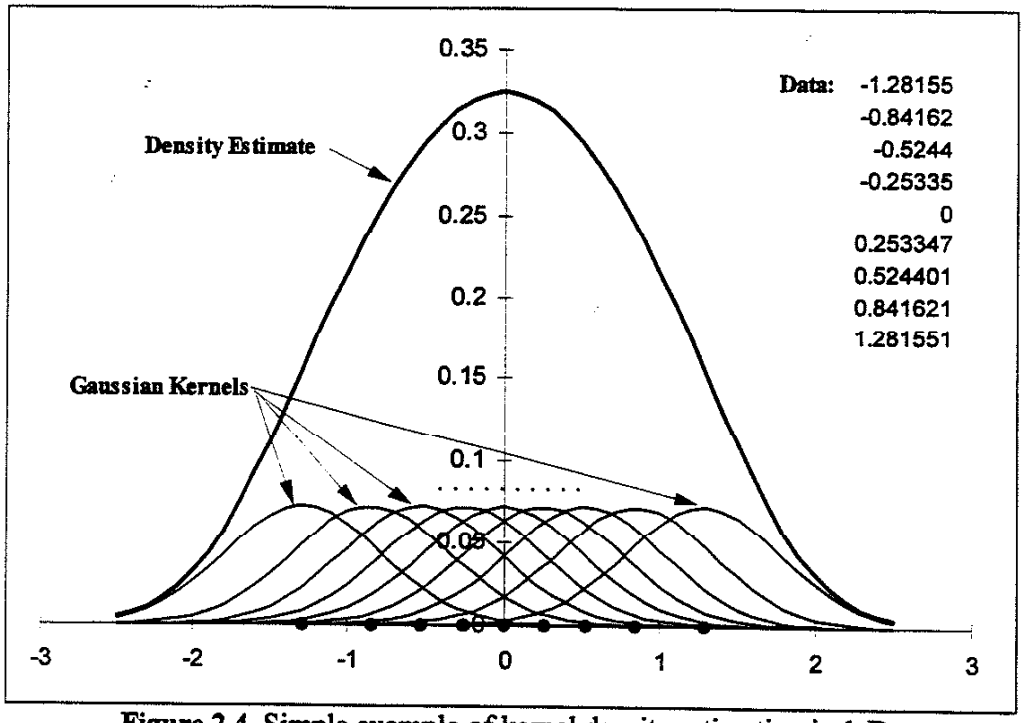


Figure 2.4 Simple example of kernel density estimation in 1-D.

2.4.2 The 2-D Kernel Method

The ideas introduced in the previous section on the kernel method in 1-D space can easily be extended to applications in 2-D space. In a similar manner to the 1-D case, where a simple kernel $K(x)$ is placed over each data point, in the 2-D case, a kernel $K(x,y)$ is placed over each data point in the x-y plane, and those kernels are then summed over the plane.

In a similar fashion to the 1-D case, where the area under the kernel must integrate to 1, in the 2-D case, the volume under the kernel must integrate to 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(x,y)dxdy = 1 \qquad (2.4.3)$$

Again, similarly to the 1-D case, statisticians will typically choose as a 2-D kernel a radially symmetric probability density function [3,4]. Two often-used kernels are the standard bivariate normal density function and the 2-D Epanechnikov kernel, presented in Figure 2.5 below.

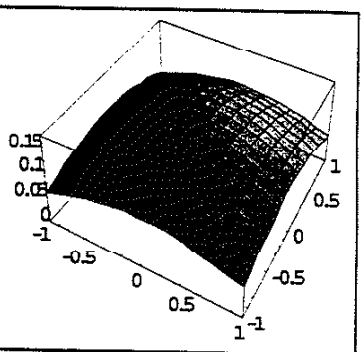
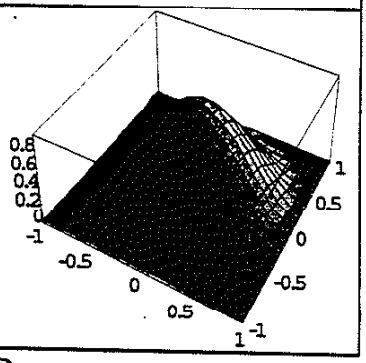
Standard bivariate normal kernel	$K(x,y) = \frac{\mathrm{Exp}\left[-\frac{1}{2}\left(x^2+y^2\right)\right]}{2\pi}$ for all x,y	
Epanechnikov Kernel	$K_e(x,y) = \frac{2}{\pi}\left(1-(x^2+y^2)\right)$ if $x^2+y^2<1$ $K_e(x,y) = 0$ otherwise	

Figure 2.5 Kernel functions in 2-D.

The density function is determined at each point in the plane *(x,y)* by summing the kernels and determining the magnitude of the kernels that lie over that point with the following formula:

$$f(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}, \frac{y-Y_i}{h}\right). \qquad (2.4.4)$$

Again, the density summation is dependent on the kernel bandwidth, h. As was mentioned previously, determining the magnitude of the bandwidth to be used in a

density estimate is not a trivial procedure. For details of the derivation of a formula for the optimal bandwidth, see Cacoullos [5] and Epanechnikov [6]. These researchers showed that for normally distributed data, with a standard deviation of σ,

$$h_{opt} = \sigma \cdot A(K) n^{-\frac{1}{6}} \tag{2.4.5}$$

for normal kernel, A(K) = 1.0

for Epanechnikov kernel, A(K) =2.40

The magnitude of the bandwidth is a physical representation of the variability that a point in the plane is allowed to exhibit. In Figure 2.5, each point represents a bin in the histogram, and for illustrative purposes, a circle is drawn to represent the boundary of the kernel that is centered at the most extreme point in the loading history, which is at roughly (2,-2). The point at which the kernel boundary is centered has a variability that is dependent on the bandwidth of its kernel.

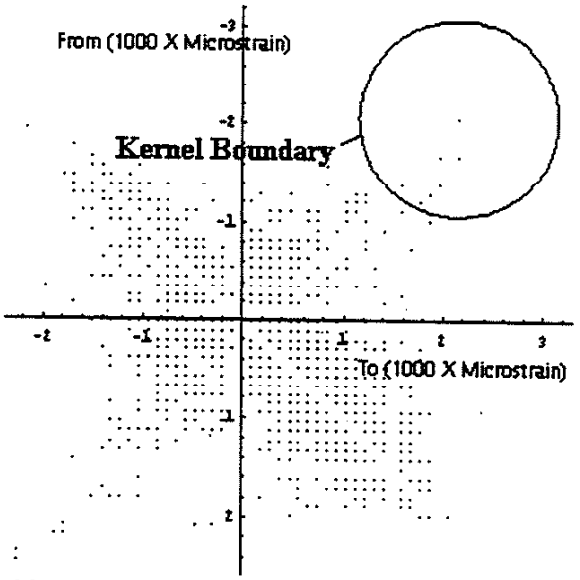


Figure 2.6 The kernel boundary represents the variability assigned to the point which the kernel is centered at. Note that this diagram is just for an example, and is not representative of any real data.

The kernel boundary represents the location where that data point can possibly occur in a future test, or in other words, it represents the variability that the point can potentially exhibit. Of all of the possible positions in that boundary, its most likely position in a future test is directly in the center of the kernel boundary because that is where the cycle occurred in a previous test, however, it can potentially exist anywhere in the boundary.

The problem with the kernel method as presented so far is that it mandates that the bandwidth, h, be the same for all data points, (X_i,Y_i), in the set. This means that the every data point will have the same variability, and this lack of flexibility is a major drawback of the kernel method.

One of the principles of the model is to allow the magnitudes of the extreme cycles to vary much more than the magnitudes of the low-load cycles, and in order to facilitate the idea of allowing the variance to change for different regions of data, it is necessary to use the *adaptive kernel estimation* technique.

2.4.3 Adaptive Kernel Estimation

The basic idea of the adaptive kernel technique is that it allows the bandwidth, and hence the variance, to change from point to point. This is done by using a broader kernel (greater variability) in sparse regions of data, and using a narrower kernel in the more dense regions of data.

The problem now is how to objectively determine if a data point is in a sparse region. Silverman explains the solution to this problem [4]:

> The adaptive kernel approach copes with this problem by means of a two-stage procedure. An initial estimate is used to get a rough idea of the density; this estimate yields a pattern of bandwidths corresponding to the various observations and these bandwidths are used to construct the adaptive estimator itself.

The actual procedure involved in calculating the density with the adaptive kernel technique has three steps, the first of which is to calculate that initial estimate. The initial estimate should be calculated by way of some non-parametric estimation technique, and Silverman, among others, has determined that the method chosen to get this initial estimate, (called the *pilot estimate*) is of little consequence. He suggests using the fixed-bandwidth kernel estimation technique described earlier to arrive at the pilot estimate, and that advice is adhered to in the model [4].

The second step in the procedure is to use that pilot estimate to determine a factor, λ_i, for each data point, by which to scale the bandwidth of the kernel that is centered at that data point. The value of each bandwidth factor is calculated in the following manner:

$$\lambda_i = \left\{ \frac{f(X_i, Y_i)}{g} \right\}^{\alpha}, \tag{2.4.6}$$

where g is the geometric mean of $f(X_i, Y_i)$. The geometric mean is calculated by

$$g = \left(\prod_{i=1}^{n} f(X_i, Y_i) \right)^{-n}$$

or for numerical overflow purposes, (2.4.7)

$$Ln(g) = \frac{1}{n} \sum_{i=1}^{n} Ln\left(f(X_i, Y_i) \right).$$

The value of α is called the *sensitivity parameter*, and it has been shown by Abramson [7] and others that α=1/2 generally yields good results. Finally, the density function resulting from this adaptive kernel estimation technique can be calculated for any points (x,y) with the following summation:

$$f(x,y)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{(h\lambda_i)^2}K\left(\frac{x-X_i}{h\lambda_i},\frac{y-Y_i}{h\lambda_i}\right)\right]. \qquad (2.4.8)$$

The adaptive kernel estimation technique is especially effective at quantifying the density in the tails, or extreme values, of a distribution. The tails of the distribution are of particular interest to us, since the most damaging cycles are located in the tails of the rainflow histogram. The specifics of the adaptive kernel technique as they relate to the model are discussed in Chapter 3.

This concludes the theoretical background information that should be understood before the modeling procedures are introduced. The next chapter goes thru a step-by-step process of explaining the construction and application of the model.

3. PROCEDURES

The models presented in this chapter seek to use the inherent statistical properties of a rainflow histogram, or set of rainflow histograms, to solve two distinct problems.

Problem I: Given one time history, use the rainflow histogram extracted from that time history to model the rainflow histogram that would be expected if that test had been allowed to run for a much longer time period.

example: given the results of a driver making 10 laps around a test track, predict the histogram for that same driver making 100 laps around that same track

Problem II: Given a set of several time histories, each corresponding to a different usage, construct a model to predict the histogram of the single-most damaging usage that would exist in a much larger, but similar set of usages.

example: given the results of 10 drivers making 10 laps around a track, predict the histogram corresponding to the most damaging driver if that group of 10 drivers were instead a group of 100 drivers

The models developed to approximate the solutions to these problems use the procedures and concepts introduced in Chapter 2 as their theoretical basis. The models themselves are now presented in the following sections.

3.1 EXTRAPOLATION OF A SINGLE HISTOGRAM

The principle governing this problem is that each time an event is repeated, for instance hitting a particular pothole, the manner in which that event takes place is not perfectly repeatable. A human driver will not hit that pothole at the same speed, same steering angle, etc., therefore it is likely that the loads associated with that event will also not be repeated. It follows that the more extreme an event is, the less repeatable it will

be. Quantifying this repeatability, or in other words the variability associated with each event, is the main obstacle of Problem I.

In a statistical sense, this variability is set by the kernel bandwidths. The larger the bandwidth, the more variability in the loadings, and vice-versa. The first step of this model involves the extrapolation of the cumulative exceedance diagram, and then using this extrapolation to set the variance in the model. The steps used to extrapolate the exceedance are described in the following section.

3.1.1 Cumulative Exceedance Extrapolation

A rainflow counted history can be presented in a cumulative exceedance diagram, where the range is plotted on the y-axis, and the number of cycles that have a range that is equal to or exceeds that range value is plotted on a log-scale, on the x-axis. A typical exceedance diagram can be seen below in Figure 3.1.

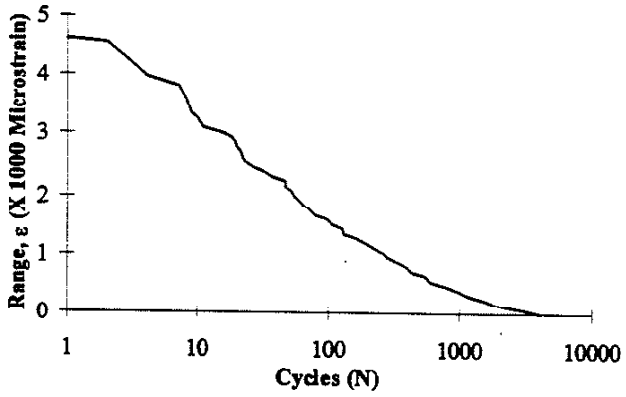


Figure 3.1 Example of an Exceedance Diagram

Socie [8] has suggested that these exceedance diagrams be curve-fit using two Weibull probability distributions, one for the higher loads, and a second distribution to describe the lower loads. The distribution that he constructed is the following:

$$N(\varepsilon_i) = N_{\max} Exp\left[-\left(\frac{\varepsilon_i}{\varepsilon_{\max}}\right)^k Ln\left(N_{\max}\right)\right], \tag{3.1.1}$$

where ε_{max} is a known value from the exceedance diagram, and k and N_{max} are the scaling parameters of the Weibull Distribution. Using this distribution and the method of linear least squares, the exceedance diagram is fit to obtain the unknown scaling parameters. The procedure is described below.

3.1.1.1 Curve Fitting Procedures

Simple linear regression involves the following as the proposed functional relationship:

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \tag{3.1.2}$$

where β_0, and β_1 are the intercept and slope, respectively, and e_i is the error associated with the approximation. The first step is to linearize the distribution proposed above. This is done by taking the natural log of both sides of the distribution, and simplifying to obtain

$$Ln\left(N(\varepsilon_i)\right) = Ln\left(N_{\max}\right) - \left[\left(\frac{\varepsilon_i}{\varepsilon_{\max}}\right)^k Ln\left(N_{\max}\right)\right] = Ln\left(N_{\max}\right)\left[1 - \left(\frac{\varepsilon_i}{\varepsilon_{\max}}\right)^k\right],$$

which is now of the form $y=mx$, where

$$\begin{aligned} y &= Ln\left(N(\varepsilon_i)\right), \\ m &= Ln\left(N_{\max}\right), \\ \text{and } x &= \left[1 - \left(\frac{\varepsilon_i}{\varepsilon_{\max}}\right)^k\right]. \end{aligned} \tag{3.1.3}$$

In the particular distribution that we used, $\beta_0 = 0$, so the distribution can be written in terms of the error using the above defined x, y and m:

$$e_i = mx_i - y_i \tag{3.1.4}$$

The premise of the method of least squares is that minimizing the sum of the squares of the error will yield the parameters corresponding to the best fit. So to find the value of m that will minimize the sum of the squared error, we take the first partial derivative of the squared error with respect to m, set it equal to 0, and solve for m. This is done in the following steps:

$$\sum_i e_i^2 = \sum_i (mx_i - y_i)^2 = \sum_i m^2 x_i^2 - \sum_i 2mx_i y_i + \sum_i y_i^2$$

$$\frac{\partial \sum_i e_i^2}{\partial m} = \frac{\partial}{\partial m}\left(\sum_i m^2 x_i^2 - \sum_i 2mx_i y_i + \sum_i y_i^2\right) = \sum_i 2mx_i^2 - \sum_i 2x_i y_i = 0$$

$$\sum_i 2mx_i^2 = \sum_i 2x_i y_i \rightarrow m = \frac{\sum_i x_i y_i}{\sum_i x_i^2}, \tag{3.1.5}$$

but remember that $m = Ln(N_{max})$. So, when the optimum least squares fit is found, the parameter $N_{max} = Exp[m]$.

The other parameter that must be solved for is the value of k that resides in the equation for x_i. The solution for the value of k that minimizes the sum of the squared errors is easiest found via an iterative technique. Since an equation for the optimum value of N_{max} is known for a fixed k, it's a simple procedure to iterate thru values of k, and find the k that minimizes the value of $\sum_i e_i^2$. Then N_{max} is found by calculating m using Equation 3.1.5, and then using Equation 3.1.3 to find N_{max}.

This summarizes the least squares method, but there are some special techniques used in the model to ensure a good fit in the high-range region, and these are described below.

Determination of the high-range fit parameters

Since the intention is to solve for two separate sets of parameters, one for the high-range values and the other for the low-range values, it is helpful to apply a weighting factor that will place greater emphasis on the error of the higher-range values, as opposed to the error of the lower-range values, so that more emphasis is placed on fitting the higher-range values.

Fatigue damage is proportional to ΔS^n, where n typically varies between 4 and 10. This means that a few high-amplitude cycles are much more damaging than a large number of low-amplitude cycles. An appropriate weighting compensation for the error is to multiply the error by $(\Delta\varepsilon_i)^4$. Doing so causes the fitting parameters to be strongly dependent on the goodness of fit in the high-range area of the exceedance diagram, while the low-range region of data has little effect in determining the parameters. Thus, the fit obtained is weighted to have the best fit in the high-range portion of the exceedance diagram, as seen in a typical exceedance curve fit in Figure 3.2.

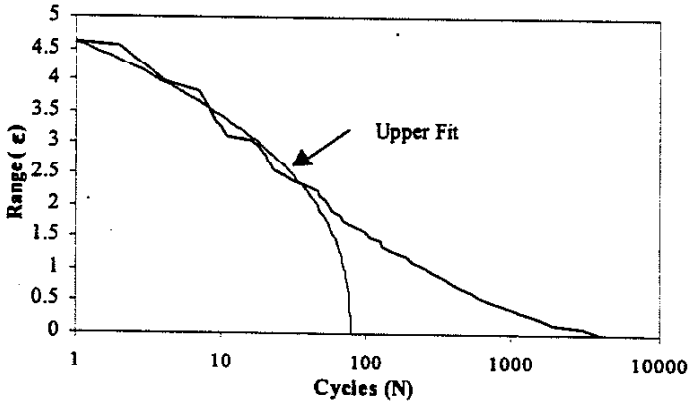


Figure 3.2 Example of upper fitting result.

Smoothing of high-range data

Quite often, a time history may contain some extreme load values that are inconsistent with the rest of the data set, as in Figure 3.3. If parameters for that data set are determined with the procedure outlined above, problems will arise. Namely, the extreme data causes the parameters to become overly liberal, thus throwing off the impending extrapolation.

While this data is uncharacteristically high, it cannot be simply ignored as erroneous, because it most likely holds some validity, so the question of how to handle this data arises. The method chosen to handle the extreme data in this model is called *kernel smoothing*, and uses the same principles as 1-D kernel estimation, which was introduced in Section 2.5.1. The difference is that instead of trying to predict the density, as is discussed in Section 2.5.1, we are now trying to intentionally oversmooth the given data. To accomplish this, we intentionally choose a bandwidth that is larger than the optimal bandwidth, and this large bandwidth has the effect of smoothing the extreme data down to a reasonable magnitude.

An example of the results of this kernel smoothing procedure can be seen below in Figure 3.3. While the actual portion of the curve that is smoothed is relatively small, the difference that the smoothing makes in the curve fitting procedure is significant, because of the weighting placed on the extreme data.

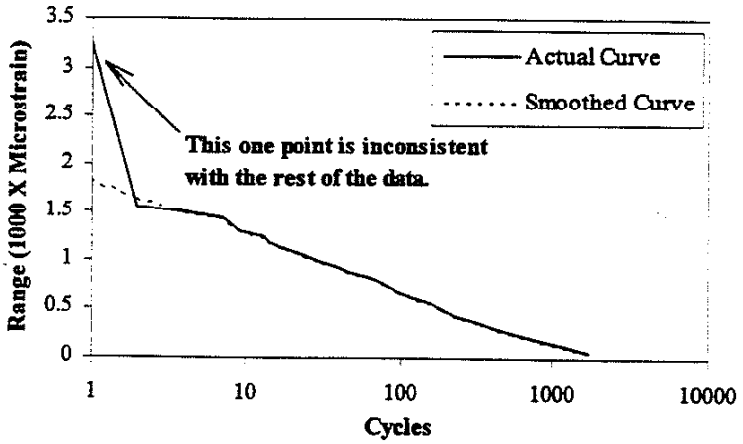


Figure 3.3 Example of kernel smoothing of extreme data.

Determination of the low-range fit parameters

Once the upper portion of the exceedance curve has been fit, a very similar procedure is used to fit the lower portion of the curve. We define the lower portion of the curve as the region of the curve between the end of the curve (N=N_{max}) and the point where the curve and the upper fit intersect. See Figure 3.4.

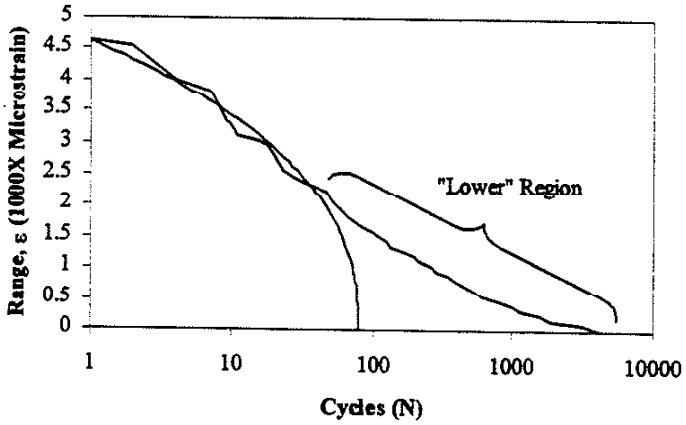


Figure 3.4 Definition of the *Lower* region.

The procedure is to then find a set of parameters to describe a curve that, when added to the upper fit, approximates the actual curve in the lower region. This is easily

done using the same procedure as described earlier in this section, with two differences. The first difference is that with the lower fit, it's not necessary to apply a weighting function to the error, because this fitting procedure is strictly confined to the data in the lower portion of the curve, so there is absolutely no influence from the upper portion of the curve. The other difference is that with the lower fit it's not necessary to use the kernel smoothing, as was discussed with the upper fit, because while it would smooth the data, it's simply not worth the effort.

Once the curve fitting procedures are completed, we have a curve that approximates the actual data, as in Figure 3.5 below, and does so using only 6 parameters: k^{upper}, $N_{\max}^{upper}$, $\varepsilon_{\max}^{upper}$, k_{lower}, $N_{\max}^{lower}$, and $\varepsilon_{\max}^{lower}$. This parameterized curve is then used to extrapolate the exceedance diagram.

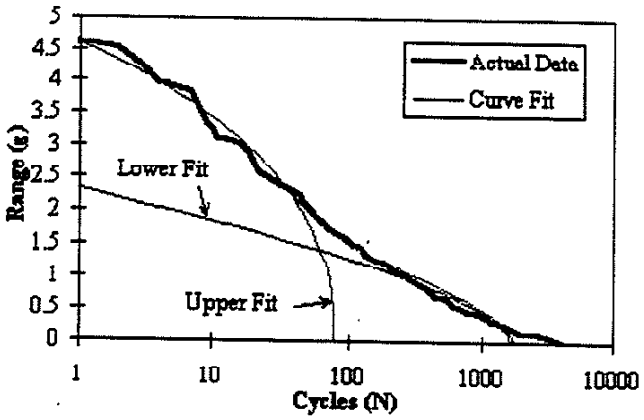


Figure 3.5 Example of completed exceedance curve fit.

3.1.1.2 Extrapolation Procedure

The goal in this step is to determine the projected exceedance diagram at a total number of cycles, N_{max} = *Extrapolation Factor* * N^0_{max}, where N^0_{max} is the number of cycles in the given durability test. The extrapolation is estimated by using the curve fit parameters to calculate N(ε_i) for incrementing ε_i until $N(\varepsilon_i) = \frac{1}{ExtrapolationFactor}$. Then the number of cycles at each ε_i is multiplied by the extrapolation factor, to effectively shift the exceedance to the right, and yield our anticipated exceedance. See Figure 3.6

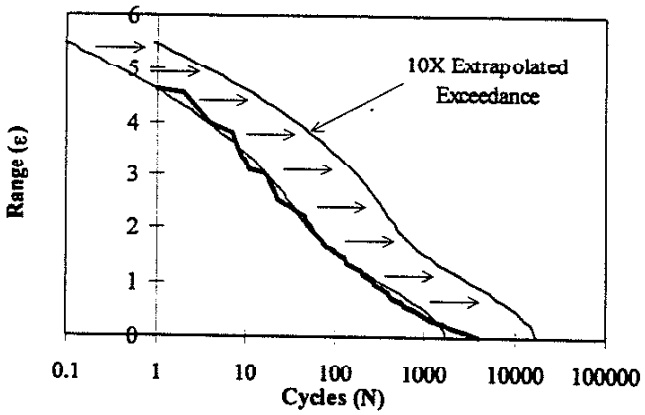


Figure 3.6 Example of a 10X extrapolated exceedance diagram.

This extrapolated exceedance is used at a later step as the gauge for setting the variance in the model. The next step in the modeling procedure is to calculate the density function.

3.1.2 Density Calculation

The calculation of the density is based on the 2-D kernel method, whose theory is presented in Section 2.5.2, but there are a few set-up steps that are used to prepare the histogram for density calculation, and these set-up procedures are described below.

3.1.2.1 Set-Up Procedures

One of the widely recognized problems associated with the kernel method are difficulties that arise when the data to be analyzed is in a bounded domain. This is indeed the case with rainflow histograms, because in each rainflow histogram, there is a diagonal set of bins that are set up to hold cycles with a range of zero, but in rainflow counting it is recognized that a zero-range cycle cannot exist, therefore, these diagonal bins are empty. If these empty bins are included in the density estimation, they cause the density estimate to be artificially low in the regions surrounding the diagonal. This region is highlighted in Figure 3.7 below.

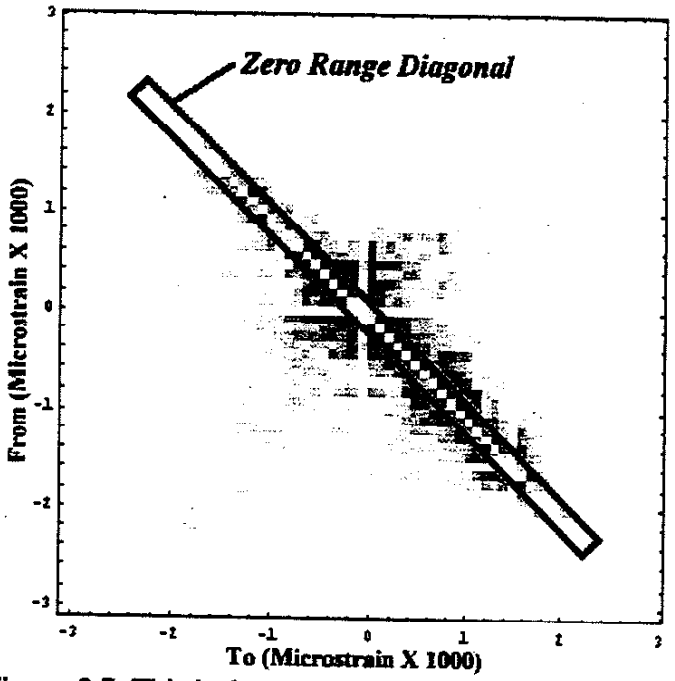


Figure 3.7 This is the zero-range diagonal of bins which causes a problem in the kernel method.

There are a few techniques that can be used to correct this problem of bounded domains, and the one that was used in this model is a reflective technique [4]. Since the

2-D kernels typically used are symmetric, augmenting the data by adding reflections on that zero-range diagonal, or in essence mirroring the data, will yield a valid density estimate at the bins near the boundary.

In order to create a mirror, it is necessary to split the data into two separate data sets: one corresponding to data "above" the diagonal, and the other being data "below" the diagonal. Once they are separated, the mirror is created along the entire diagonal, in the manner described in Figure 3.8.

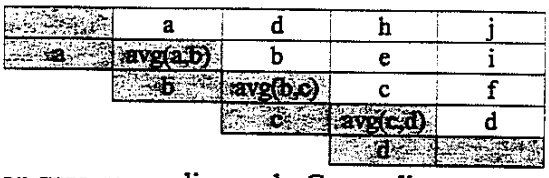


Figure 3.8 Mirroring on zero-range diagonal. Gray cells represent the mirror, white cells represent the original histogram. Note that this is just a small portion of the histogram, and that the remainder of the diagonal is mirrored in the same manner.

For the remainder of the modeling, these two data sets are handled separately. This procedure of splitting the histogram into two data sets, and then using the mirroring procedure to account for the bounded domain weakness of the kernel method are the set-up steps that are necessary before the density estimation procedure can begin. It is now appropriate to begin a discussion of the procedural tasks involved in the calculation of the density estimation.

3.1.2.2 Implementation of the Adaptive Kernel Method

As was discussed in the background section, the quality of a density estimate is widely recognized to be primarily determined by the choice of smoothing parameter, and is only in a minor way affected by the choice of kernel. The kernel that was chosen for use in this model is the radially symmetric Epanechnikov kernel, for two reasons. The

first reason is that unlike the normal kernel, the Epanechnikov kernel has a definite radial boundary, and the probability of any data point occurring outside of that radius is zero. This boundary was useful in being able to mandate a specific variability in the model. The other reason that the Epanechnikov kernel was used is that calculations involving this kernel are simpler and less time consuming than some of the other kernels available, and one of the virtues of the kernel method is it's inherent simplicity, so using a simple kernel is sensible.

To reiterate what was mentioned in the background section, instead of using the fixed-bandwidth kernel method to obtain a density estimate, this model makes use of the adaptive kernel method, which allows for better density estimation in the regions of extreme cycles.

The first step in the adaptive kernel estimation technique is to use the fixed-bandwidth kernel method to obtain a pilot density estimate, which is then used to determine the magnitudes of the adaptive bandwidths. To use the fixed-bandwidth kernel method, it's first necessary to choose a bandwidth, h. The precision of h is not of extreme importance in determining the pilot estimate, so a few liberties are taken in determining this bandwidth. Since a closed-form solution of the optimal bandwidth is known for a normal distribution of data, our model makes the assumption that the data in the histogram takes on a normal distribution, and uses the following as a value of h [3,4]:

$$h = Min(\sigma_x, \sigma_y) * 2.42 * n^{-1/6} \qquad (3.1.6)$$

where σ_x is the standard deviation of the data in the x-direction and σ_y is the standard deviation of the data in the y-direction. The minimum standard deviation is chosen since it is better to undersmooth the data in the pilot estimate than it is to oversmooth, and a

smaller h results in less smoothing. Then, using the Epanechnikov kernel function, the bandwidth, h, and the fixed-bandwidth kernel method described in Section 2.5.2, the value of the pilot estimate is found at each data point (X_i,Y_i).

The next step in the procedure is to find the adaptive bandwidth factors λ_i. This is done by calculating the geometric mean, g, of the pilot density values at all data points (X_i,Y_i), and then using equation 2.4.6 to calculate λ_i. These multiplication factors, λ_i, describe how the bandwidth varies from kernel-to-kernel, and equation 2.4.8 is then used to calculate the density function. In this calculation, another bandwidth, h, is required, and it is this value of h that really sets the allowed variance in a density estimate.

As stated previously, in this model the variance is dictated by the magnitude of the extrapolated exceedance diagram, which was found using the procedures in Section 3.1.1. So, the bandwidth to be used in equation 2.4.8 is determined by iterating for h to find a value such that $[Range(X_i,Y_i)+h^* \lambda_i^{max}]_{max}$ is approximately equal to the maximum range of the extrapolated exceedance, where $Range(X_i,Y_i)$ is the range of the cycle(s) at point (X_i,Y_i). Using this bandwidth, and equation 2.4.8, the magnitude of the density function is calculated at each bin in the histogram.

Finally, after the density estimate has been determined for each bin in the histogram via the procedures outlined above, the final step in the model is the simulation of the random loadings.

3.1.3 Simulation of Random Loadings

As was briefly mentioned in the background section, there are two classes of random variables, discrete and continuous random variables. Thus far, all that has been

dealt with are continuous random variables, because the space in the x-y plane is continuous, but in this simulation we know that there are a discrete number of bins in the histogram which the cycles can be placed into. For instance, in a 64 X 64 histogram, there are only 4096 potential bins that a cycle can be placed in, rather than an infinite number of possibilities as would be the case with continuous random values, so it is sensible to simulate the histogram as a set of discrete random variables.

The probability of a random cycle occurring in each bin in the histogram is known from the calculated density function, so with these probabilities, it is simple to conduct a Monte Carlo simulation of the cycles. In the set-up of the problem, the data was split up into two separate sets, one consisting of data above the diagonal, and the other set consisting of the data below the diagonal. The model keeps those sets separate for the random cycle generation as well. The simulation generates a number of cycles, N, where $N = N_1 + N_2$ and

N_1 = *Extrapolation Factor * Number of Original Cycles in Set Above Diagonal*
N_2 = *Extrapolation Factor * Number of Original Cycles in Set Below Diagonal.*

An example of a 10X extrapolated histogram is shown in Figure 3.9 below. Because of the variability that this model allows for, there is not one particular 10X extrapolation that is generated every time that the simulation is run. That is, for each simulation, a similar but slightly different histogram will be generated. The results of five simulations run with the same original data and same extrapolation factor as that in Figure 3.9 are shown in exceedance form in Figure 3.10.

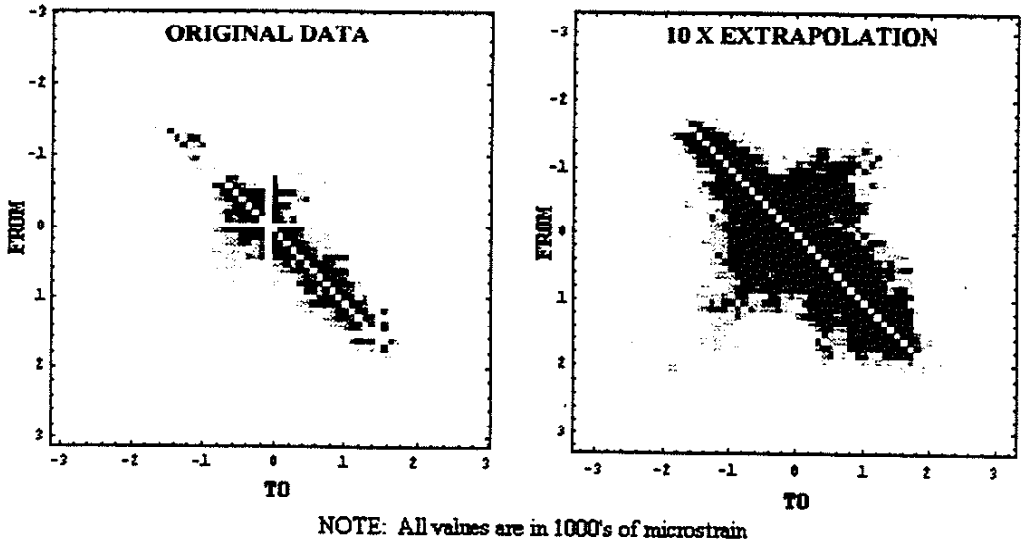


Figure 3.9 10X Extrapolation of a loading history from the ATV project.

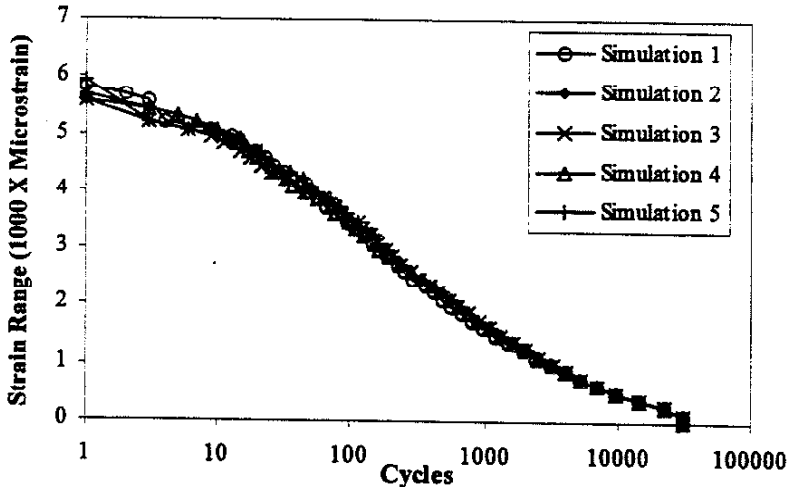


Figure 3.10 The results of five 10X extrapolation simulations. Note that the results differ slightly.

Further examples of results obtained with this model, and a discussion of those results, follows in Chapter 4.

3.2 Extrapolation of Multiple Histograms

In this scenario we are given a set of histograms corresponding to a variety of different usages, and we would like the model to predict the one most damaging histogram in a larger, but similar set of operations. The main obstacle that we had in the first problem, which was determining the variance, is more easily solvable in this problem, because we already have a *set* of histograms, and the variability within the set itself is then used to determine the variance in the model. The procedure to quantify this variance will be described in more detail in an upcoming section.

In this problem, the main dilemma is instead in determining how many cycles should be in the most extreme histogram. Then there is a problem of determining a method of generating these cycles in a manner that is random in nature, but still exhibits a definite extremeness.

In setting up the model, it was necessary to construct two fundamental guidelines upon which the model could be built. The first assumption is that there is not necessarily a correlation between the damage generated by a loading history and the total number of cycles in that history. In other words, there is no reason to believe that the most damaging history in a set of histories will contain the most cycles. The other assumption is that the most damaging history is not necessarily going to contain the most extreme loadings, but it is most likely a history that contains a large number of mid-to-high range cycles.

These assumptions tell us that we must use some technique to ensure that the cycles in our most extreme histogram are generated in a controlled, yet random manner. The term "controlled" means that there must be a way to generate a histogram where a

higher proportion of the cycles are in the mid-to-high range than is the case for an average histogram. The method chosen to control the cycle generation was to discretize the histogram into regions of equivalent damage content.

3.2.1 Discretization of the Histogram

In order to utilize the inherent density of the set of histograms, and still be able to dictate that there be a large proportion of cycles in the high range regions (relative to the proportion of cycles in the same regions in an average loading history), it was recommended by Dreßler, et. al. [9], to break the data set into a series of clusters, and then base the analysis on these clusters.

The first step is the process of determining how these regions should be distinguished from each other. Several methods were investigated, and the one which was chosen for this model involves discretizing the histogram into a series of triangular regions, as in Figure 3.11 below.

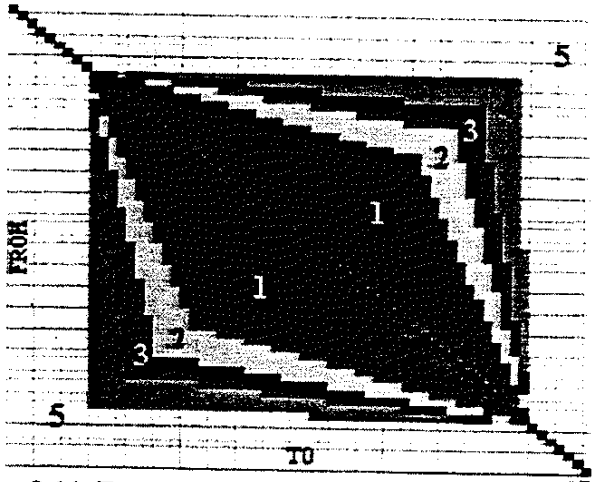


Figure 3.11 Example of discretization of damage regimes.

The regions in this model are chosen so that the sum of the damage, calculated with the stress-life equation,

$$D_i = \sum_{\text{all bins in regime i}} N(\Delta S)^4 \tag{3.2.1}$$

where i = *(1,...,total num of regions)*, is approximately equal in each region. The total number of regions used in the analysis is not of major consequence, but implementing the model when using five distinct regions gave reasonable results.

Once the regions have been defined, the model loops thru each histogram in the given set, and counts the number of cycles that are located in each regime, and saves these values for use in a correlation analysis later in the model. The next step in the model involves the estimation of the damage in the most extreme loading case.

3.2.2 Estimation of Damage in Extreme Usage

Using the strain-life equation [10],

$$\frac{\Delta\varepsilon}{2} = \frac{\sigma_f'}{E}\left(2N_f\right)^b + \varepsilon_f'\left(2N_f\right)^c \tag{3.2.2}$$

we can calculate the total damage done by each loading history, and for use in a later step, we also calculate the damage done by the cycles in each previously defined regime, for each loading history. Then, using those damage totals, it is easy to use parametric density estimation, as described in Section 2.3.1, to obtain a probability distribution of damage values for all of the measured data. After examining some of the more frequently used probability distributions, it was determined that the Weibull distribution is especially adept at describing data such as damage totals, which tail off exponentially. To verify that the Weibull is a viable distribution, the damage results of 19 ATV tests have been

calculated, and the non-parametric kernel density estimate of those damage values (calculated with 1-D kernel method, Section 2.5.1) is plotted on the same graph as the Weibull distribution, in Figure 3.12.

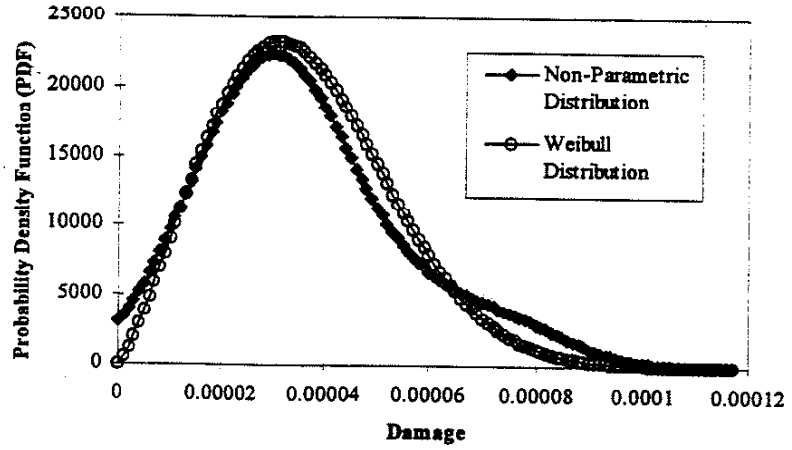


Figure 3.12 Distributions of damage calculations for 19 ATV tests.

The two distributions in Figure 3.12 are very similar, especially in the tail of the distribution, which is of the most interest in determining the extreme usages. Because of this similarity, the assumption is made that the Weibull distribution is a valid representation of the distribution of damage totals.

In the model, we chose to generate the distribution using a parametric estimation method rather than the kernel method, because the problem of selecting an appropriate bandwidth in the kernel method can create errors in the distribution, and as mentioned earlier, because the Weibull distribution is a good approximation.

Then, using this distribution f_X, an estimation of the damage total is calculated for that one most damaging usage in N in the following manner. Find the value of x such that

$$\int_{-\infty}^{x} f_X(\xi)d\xi = 1 - \frac{1}{N},$$

and that value of x is then an approximation of the damage done in the one most damaging usage in N. This x is now used to predict the number of cycles that will be in each regime for that most extreme load history.

3.2.3 Correlation Analysis

This correlation analysis is used to determine if a linear relationship exists between the number of cycles in each regime, and the total damage done for each of the known usages. The linear statistical correlation between two variables X and Y is quantified using the normalized covariance, also known as the correlation coefficient, ρ:

$$\rho = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{N}}{\sqrt{\left(\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{N}\right)\left(\sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{N}\right)}}, \qquad (3.2.3)$$

where N is the number of data points that are being correlated. The value of ρ can range from -1 to +1. A value of +1 means that the variables X and Y have a perfect linear correlation; that is, for each i, $x_i = \alpha * y_i$, where α is constant. On the other hand, a $\rho = -1$ is a perfect linear correlation, where $x_i = -\alpha * y_i$. Finally, if ρ=0, there is absolutely no linear relationship between X and Y. Any other value of ρ means that there is some degree of linear relation between X and Y, where the closer ρ is to ±1, the stronger the linear relationship is between X and Y, and the closer ρ is to 0, the less linear relationship there is between X and Y. If ρ=0, it does not necessarily mean that there is no quantifiable relationship between X and Y, but just that there is no *linear* relationship.

The model generates one correlation coefficient for each triangular regime in the histogram, and those coefficients are used to estimate how many cycles should be in each regime for the case of that one most damaging usage.

3.2.4 Making Use of the Correlation Coefficient, ρ

There are an infinite number of ways that a loading history can accumulate a damage which is representative of the one most damaging history in N, so one of the problems encountered is in determining which of these infinite possibilities is the *most likely* to occur in that most damaging history. This is approximated by using the correlation coefficients determined previously. If we are anticipating a total damage of D, we can use the correlation coefficients to predict the number of cycles that can be expected in each region when the loading that creates a damage of D is applied in an *average* fashion. This is done with the following equation:

$$E(Y|X=x) = \mu_Y + \rho\left(\frac{\sigma_Y}{\sigma_X}\right)\left(x-\mu_X\right), \tag{3.2.4}$$

where $E(Y|\ X{=}x)$ is the average of the variable Y, given that the variable $X{=}x$. This calculation is carried out to determine the number of cycles in each of the regimes, or Y, given the estimate of the total damage, $X{=}x$, determined with the Weibull distribution. Once the number of cycles expected in each regime is calculated, the final steps in the model are to calculate the density, and then generate the random loadings.

3.2.5 Density Calculation

The calculation of the density in this problem is carried out in a very similar fashion to the density estimation procedure in the first problem, with a few significant differences. The procedure is explained below.

3.2.5.1 Set-Up Procedures

The major difference between the set-up of Problem II and the set-up of Problem I is that in this problem, we are dealing with a set of histograms, whereas prior to this, the density estimation was conducted using one histogram. So before any manipulation of the data begins, the data from all of the histograms is superimposed onto one *summation* histogram, and this summation histogram is then used in the density estimation procedures.

Once the summation histogram is constructed, the data is split into two data sets, and the data is mirrored along the diagonal, in an identical fashion to the procedures described in Section 3.1.2.1.

3.2.5.2 Implementation of the Adaptive Kernel Method

For Problem II, the adaptive kernel method is invoked in the same manner as was described in Section 3.1.2.2, with one major exception. In the previous problem, the variance was determined by the desired extrapolation factor, or more accurately, by the exceedance expected at the given extrapolation factor. In this case, because we start off with a set of histograms, the variance can be determined from the data itself. The variable that is used to set the variance is again the bandwidth, h, in equation 2.4.8. It is

common to use the same value of h in equation 2.4.8 as is used in the pilot estimate, and this h is determined by equation 3.1.1:

$$h = Min(\sigma_x, \sigma_y) * 2.42 * n^{-1/6}. \qquad \text{(same as 3.1.6)}$$

The remainder of the density estimation procedure is carried out in an identical manner to the procedures in Section 3.1.2.2. Once the density estimate has been generated, the final step is to actually generate the random cycles.

3.2.6 Simulation of Random Loadings

The process of generating the random loadings for this problem is again very similar to the procedure for Problem I, which is described in Section 3.1.3. The difference here is that rather than knowing the total number of cycles to be generated for the entire histogram, we instead know how many cycles should be generated in each regime. To implement the simulation, the model loops thru each of the regimes, and generates random cycles that are limited to that region.

Once the cycles have been randomly generated, the damage done by those cycles is summed, and that total is checked against the estimated damage for the most extreme usage in a set of N usages. The process of looping thru the regimes and generating random cycles is repeated until the comparison between the generated damage, and the predicted damage is within predefined limits.

Several data sets were modeled using the previously described procedures, and the results are presented in the following chapter, along with a discussion of those results.

4. RESULTS AND DISCUSSION

Thus far, the theory behind the models has been introduced, and the procedural steps involved in implementing the models have been delineated, and this chapter presents examples of the models. One of the difficulties associated with designing a model such as those presented here is that in order to verify the success of the model, or instead to discover the reasons for failure of the model, the researcher requires a reasonably large amount of data. At the time of writing this thesis, the maximum amount of data available to test Problem I with was a time history of 10 laps around a test track. This is not enough data to make a reliable verification of the model, but because there is no alternative, this small amount of data has been analyzed, and that analysis is presented in Section 4.1. In testing Problem II, three different sets of data were available for analysis. The first is a set of 19 tests, the second a set of 54 tests, and the final set contains 334 tests. These three sets have been used to test the model, and the results of these simulations are presented in Section 4.2.

4.1 VERIFICATION OF THE MODEL FOR EXTRAPOLATION OF A SINGLE HISTOGRAM

In a previous project at the University of Illinois, durability tests were conducted on an ATV (all terrain vehicle), and a set of 19 tests were accumulated during that project. Each of the tests consists of one driver making ten passes over a test track. More details of the tests are found in Park [11].

To check the validity of the first model, it is instructive to extract one lap from a test, and see how the model's predicted histogram for a 10X extrapolation of that single lap compares with the actual rainflow histogram of that driver's entire test.

The following, Figure 4.1, is the loading history of one of those drivers traversing the track ten times, with the portion of the test that was selected for modeling designated as such.

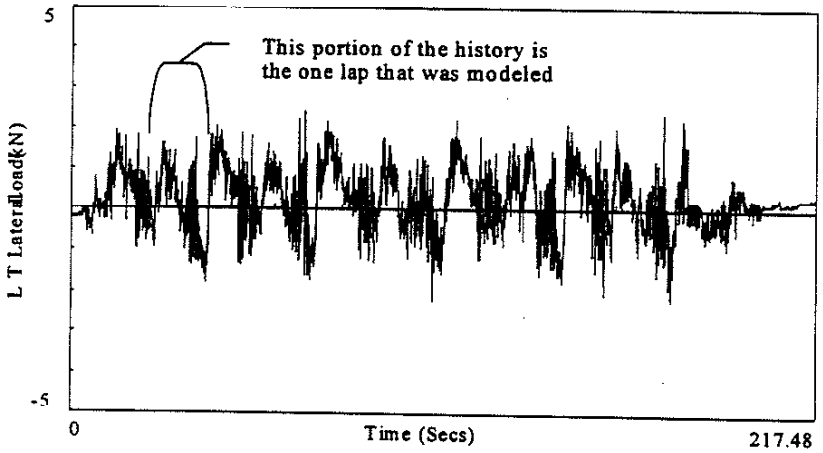


Figure 4.1 Loading history of an ATV test drive.

The lap chosen for analysis was selected at random, and is shown in more detail in Figure 4.2.

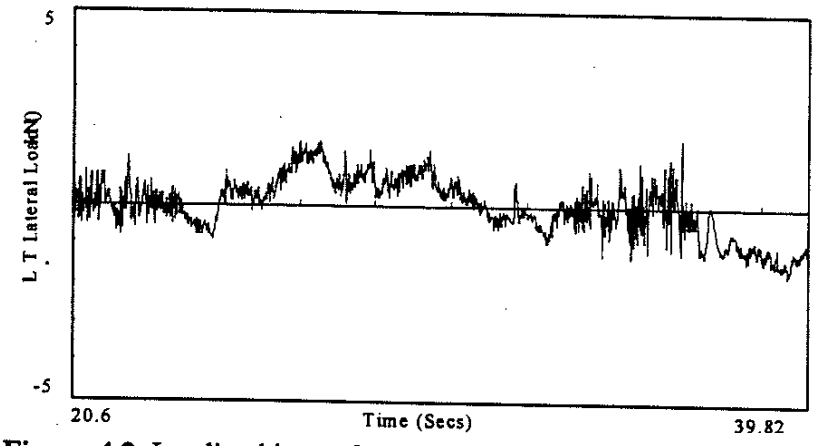


Figure 4.2 Loading history for a single pass on the ATV test track.

When this portion of the history was rainflow counted using SoMat Ease, it yielded the following rainflow histogram, Figure 4.3.

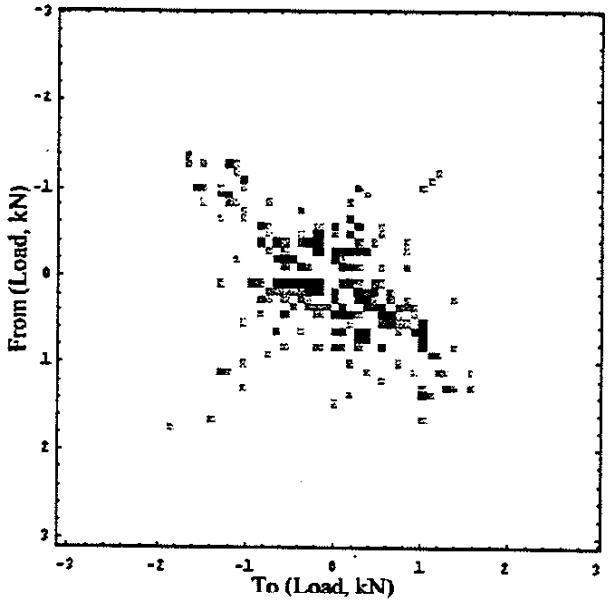


Figure 4.3 Rainflow counted histogram of the loading history in Figure 4.2.

The only two input variables that the model requires are this rainflow histogram, and the extrapolation factor, which in this case is 10. These are input to the model, and the modeling procedure begins.

The first step of the model is to extrapolate the exceedance diagram, and for the above histogram, the exceedance parameters for the upper part of the curve along with the plot are given in the figure below.

k^{upper} =1.244

N_{max}^{upper} =56.17

$\varepsilon_{max}^{upper}$ =3.56

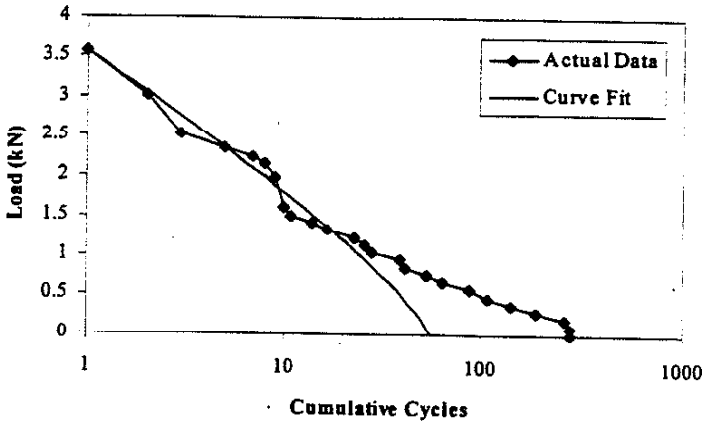


Figure 4.4 Exceedance diagram and associated exceedance parameters, of the upper half of the rainflow histogram in Figure 4.3.

Only the upper half of the exceedance curve is fit at this point, since the upper part describes most of the cycles that cause damage, and will dictate the variance in the model.

This curve fit is then extrapolated 10×, to give the estimated exceedance diagram in Figure 4.5.

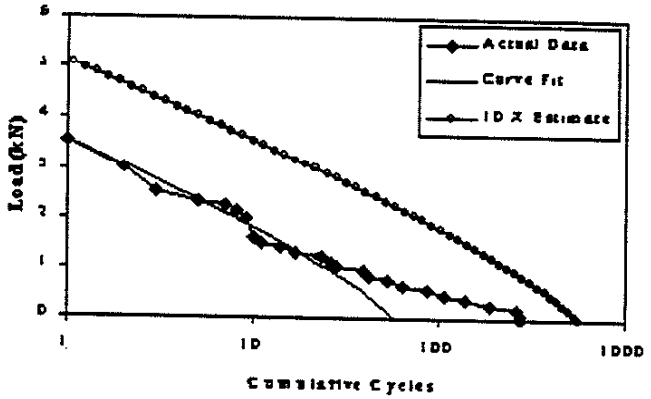


Figure 4.5 Estimated 10X exceedance extrapolation.

The next step in the procedure is to calculate the pilot bandwidth, and using the given histogram and equation 3.1.6, the pilot bandwidth calculated was 0.608. This

bandwidth is used to create the pilot estimate, and then calculate λ_i using formula 2.4.6. The adaptive bandwidths are shown in comparative fashion in Figure 4.6

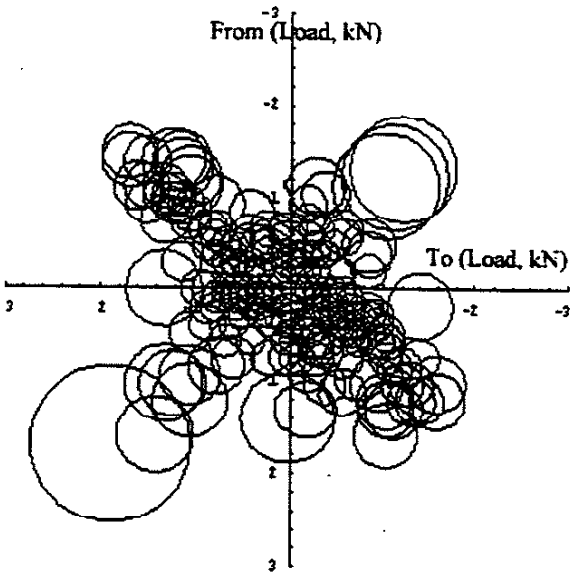


Figure 4.6 Kernel bandwidths for histogram in Figure 4.3

Notice in Figure 4.6, that the more extreme cycles (further from the diagonal) have much larger kernel bandwidths, thus allowing the location of that cycle to vary much more than cycles in more populated portions of the histogram.

Next, the density estimate is constructed using the adaptive kernel estimate. By making use of the previously determined extrapolated exceedance diagram as discussed in Section 3.1.2.2, the model sets the bandwidth in Equation 2.4.8 to a value of 0.409, and then uses Equation 2.4.8 to calculate the density function.

Finally, a Monte Carlo simulation is run, and a resulting histogram is constructed. The following is the result of one such simulation of this data, compared with the actual result of 10 laps around the track for this given driver.

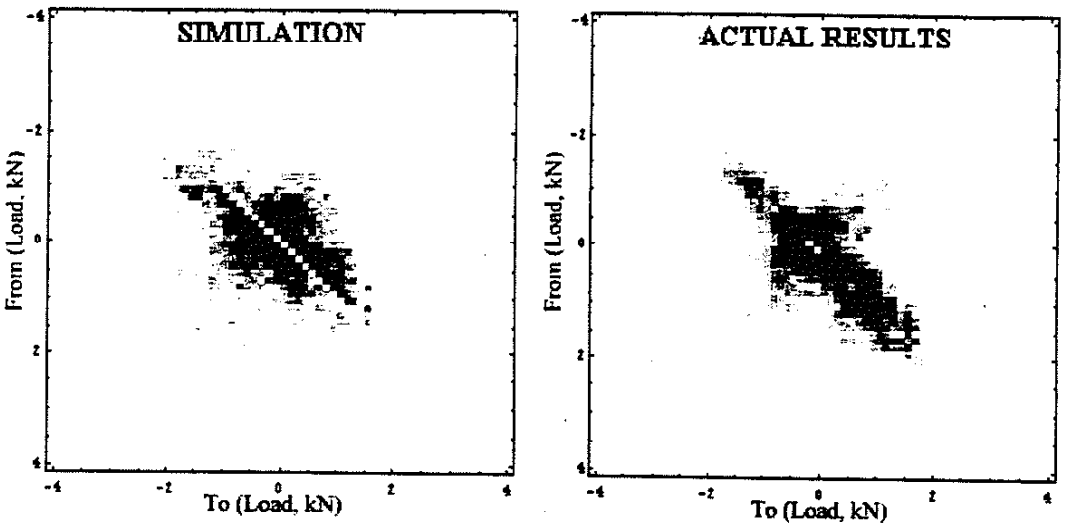


Figure 4.7 Results of the histogram generated in a simulation of the 10X extrapolation, along with the results of the actual test after 10 laps.

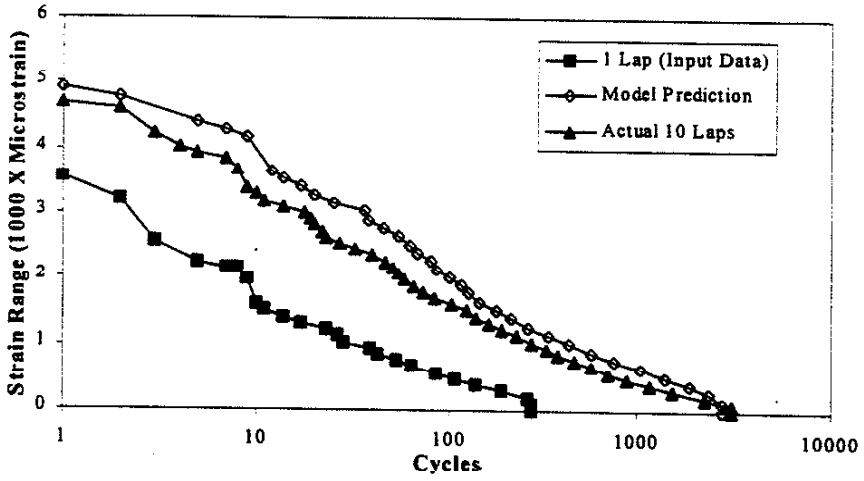


Figure 4.8 Exceedance of histograms in Figure 4.7, and of the histogram from one lap.

The simulation in Figure 4.7 is not a perfect duplicate of the actual result, but keep in mind that the modeled result is meant to be an approximation. The actual loads that this model is attempting to simulate are random, and any model in which the goal is to predict random events is certain to have uncertainties involved.

A major reason for error in this example problem is that the data given to the model (the histogram in Figure 4.3) is extremely sparse. It would be much more reasonable to give the model 10 laps, and predict the result for 100 laps, but because 10 laps was the longest available time history at the present time, the room for experimenting with the model was limited. Given the sparseness of the input data, the resulting prediction is encouraging.

Another simulation was run using the same set of data, but this time, the first 5 laps of the time history were used as the input data, and a 2× extrapolation was conducted. As would be expected, because the input histogram contains a larger portion of the history than in the previous example, the results of this extrapolation are more representative of the actual data. The results of this simulation are in Figures 4.9 and 4.10.

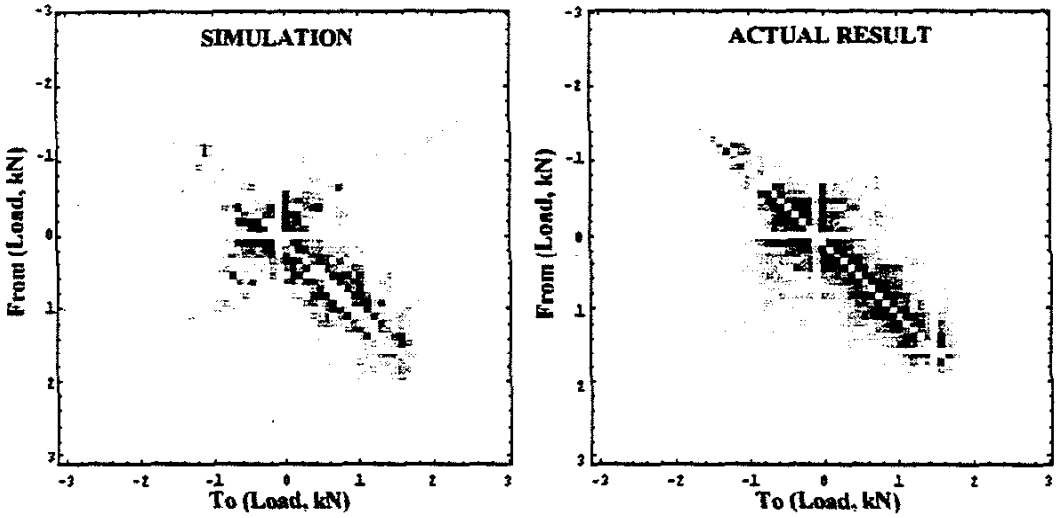


Figure 4.9 Results of the histogram generated in a simulation of the 2X extrapolation, along with the results of the actual test after 10 laps.

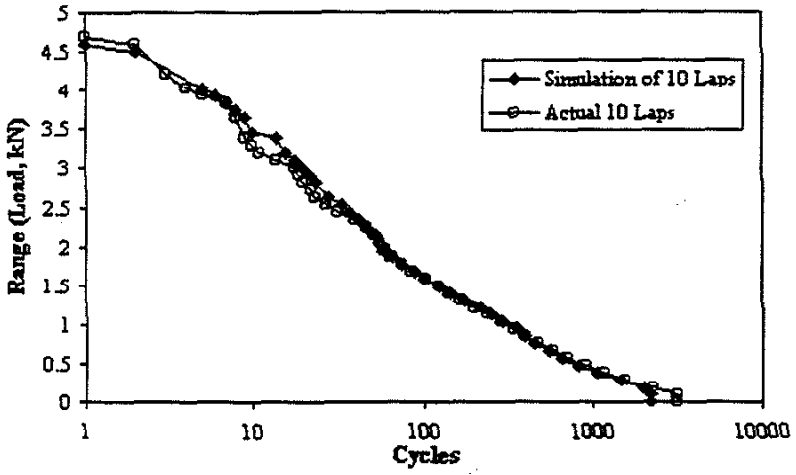


Figure 4.10 The exceedance diagrams of the histograms in Figure 4.9.

One of the shortcomings of this model is in the way that it estimates the total number of cycles in an extrapolated histogram. It does so by multiplying the number of cycles in the given data by the extrapolation factor, and this can cause problems, since it is very unlikely that the total number of cycles in a history will be split up equally over a long period of time.

An example of this error is evident in the example problems presented above. In the first example, the predicted 10× extrapolation contained 2780 cycles, and the second example, the 2× extrapolation, contained only 2188 cycles, while the actual data that these were supposed to simulate (the entire history in Figure 4.1) actually contains 3133 rainflow counted cycles. This explains the discrepancy in the density of the histograms, especially along the diagonal, where cycles accumulate very quickly. But from a durability viewpoint, it is not necessary to accurately model all of the cycles in a history, only the most damaging cycles.

Despite the fact that this model has been subjected to a minimal amount of testing, the results presented here are very encouraging, and further testing of the model would certainly be worthwhile.

The final section of this chapter gives the results of some analyses using the model designed to solve Problem II.

4.2 Verification of the Model for Extrapolation of Multiple Histograms

The first set of data analyzed in this section is from the same ATV tests analyzed in Section 4.1. That set consists of 19 different drivers, each making 10 passes over the same test track on the same ATV. The second set of data consists of 334 different flights of a single airplane. The final set of data consists of 54 in-service tests of a lawn and garden tractor. The loading histories come from 54 different owners from throughout the US, and each test was conducted using one of two tractors specially configured for data acquisition. The difference in relative variability of life of these three sets of tests is shown in Figure 4.11. This figure was generated by fitting the data to a log normal distribution, and then plotting the data on log normal probability paper. The portion of the data that is of importance in this model are the data points with short life, or that data which is below the x-axis.

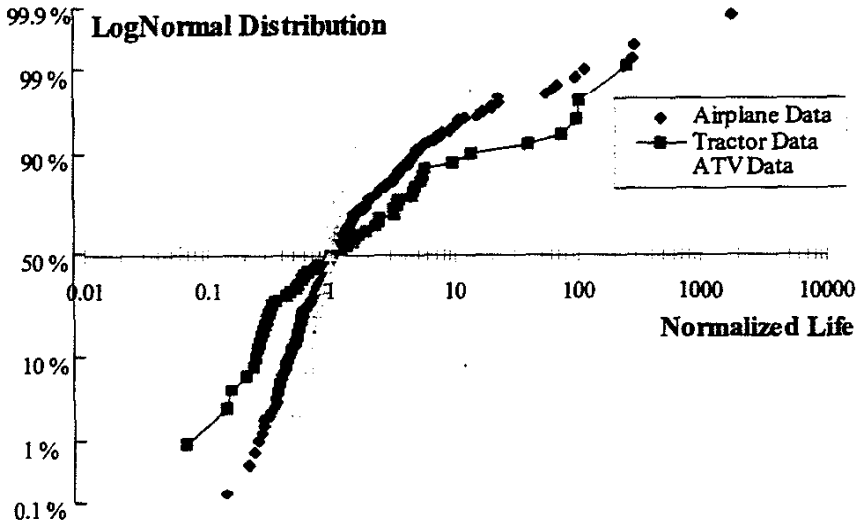


Figure 4.11 Log-normal distribution of life, normalized with respect to the median life.

The tractor data has the greatest variability, which is expected, because out of the three data sets, the usages involved in the tractor testing have the most widespread test base. Since the terrain of the work environment differed substantially for each user, and because there were two different tractors involved in the testing pool, we could expect widely varying loads, and widely varying life.

The airplane data is the next most variable data, because while the same plane was used in each test, the nature of the flights differed from flight-to-flight, and therefore the loadings varied similarly.

Finally, as could be expected, the ATV data is the least variant, because all of the testing was done on a single ATV, and all of the tests were run over the same track, so the only cause of variance in the data is the varying the character of the different drivers.

These three data sets were used to test the model created to solve Problem II. Given a set of histograms, the objective of Problem II is to be able to predict the results of

the most extreme histogram in a larger, but similar set of histograms. For the remainder of this thesis, the term *99%* is used to classify the one most extreme histogram in a group of 100, the term *99.9%* is used to describe the one most extreme histogram in 1000, and so on. The details and results of example simulations of the three previously mentioned data sets are given in the following sections.

4.2.1 Analysis of ATV Durability Data

The ATV data set consists of 19 tests, and an assumption was made that to get a good approximation of the inherent distribution of the damage in these 19 data sets, we need only take a random sampling of 5 sets. The goal of this example problem is to verify the model on a very small scale by predicting the histogram that will be created by the most extreme loading history in a set of 19, when taking a random sampling of 5 of the histograms.

To start the problem, we randomly select 5 of the histograms without replacement, meaning that we cannot select the same histogram more than once. The histograms that were selected for this simulation are shown in Figure 4.12, and the exceedance diagrams for those 5 histories are shown in Figure 4.13.

In the model for Problem I, the exceedance diagram is used to determine the variance, but the exceedance is not involved in the model for Problem II. However, the exceedance diagram is used throughout the rest of this section as a visualization tool because with an exceedance diagram, it is easier to distinguish differences in loading histories than if the same data were to be plotted in a set of histograms.

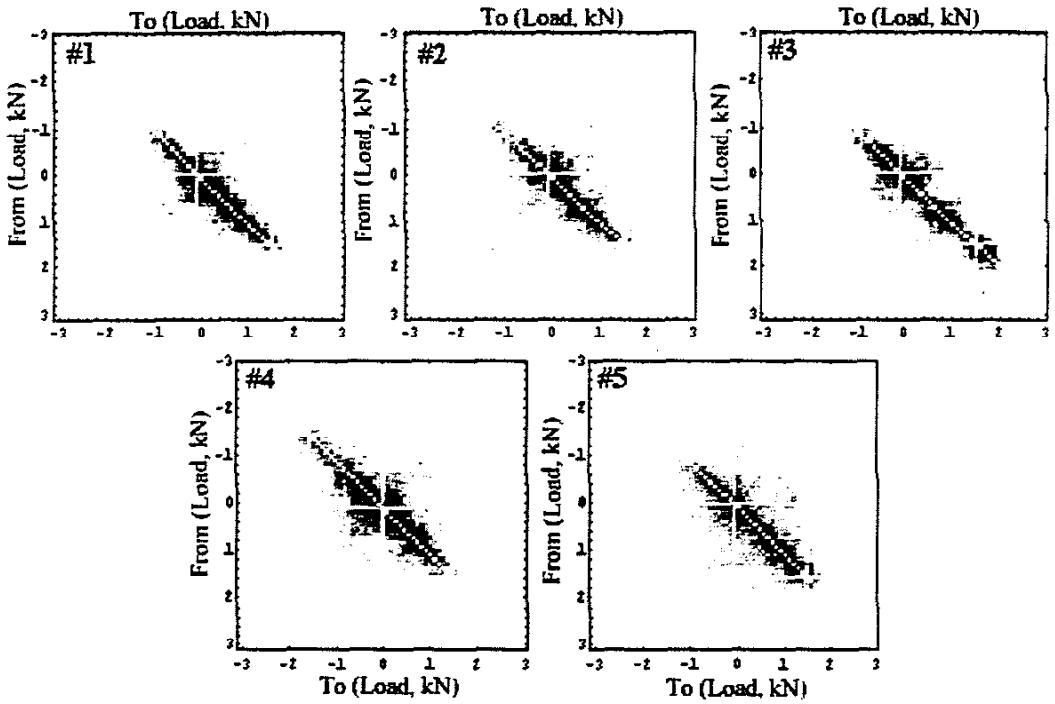


Figure 4.12 Five randomly selected histograms from ATV tests.

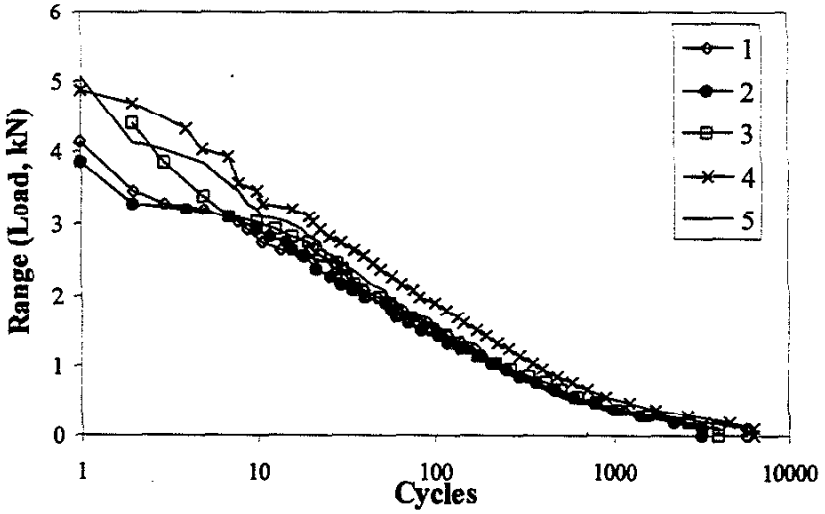


Figure 4.13 Cumulative exceedance diagram of the five randomly selected ATV tests.

Once the histograms have been randomly selected, the first major step in the procedure of this model is to descretize the histogram into triangular regions of equal

damage, as described in Section 3.2.1. The best way to discretize the histogram is to first apply a smoothing effect on the data, which can be done by calculating the density function.

To calculate the density function, a summation histogram is first constructed by simply superpositioning all of the random histograms on top of one another and adding the total number of cycles in each bin. The 2-D adaptive kernel method is then applied to the summation histogram in a similar manner to that described in Section 4.1, with the exception being that the bandwidth used in Equation 2.4.8 in this model is the same as the bandwidth used in determining the pilot bandwidth. In this example problem, the bandwidth value used in the procedures was $h=0.29$, which was calculated using Equation 3.1.6. The density function $f(x,y)$ is calculated for each bin in the histogram.

After calculating the density function, the discretization process begins by calculating the damage done at each bin with the following equation:

$$D(x_i, y_i) = f(x_i, y_i) \cdot (\Delta range)^4$$

but becasuse $f(x_i,y_i)$ is not the same as N_i (see Equation 3.2.1), $D(x_i,y_i)$ is not really the damage, but is instead directly proportional to the damage, and its value is used solely for the purpose of discretizing the histogram.

Then this damage matirx, $D(x_i,y_i)$, is iterated thru to find triangular regimes that constitute equal sums of damage. The regimes determined in this problem are shown in Figure 4.14.

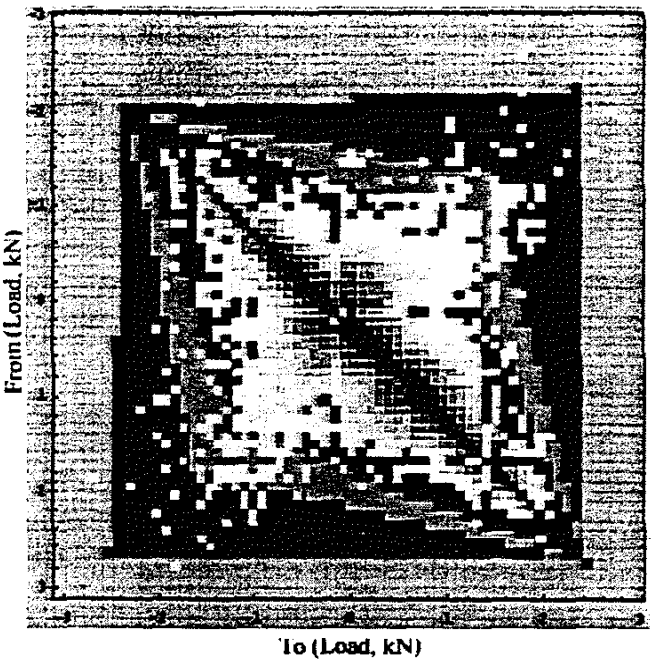


Figure 4.14 Damage regimes of the five random ATV tests. The histogram is superimposed over the damage regions on the diagram.

Next, the total damage for each of the five histograms and the damage done by the cycles in each of the five regimes were calculated using the strain-life equation. With these total damage values that are calculated, we then determine the Weibull distribution that describes those damage totals.

To determine the Weibull parameters in the first model, we used a linear least squares approximation because we wanted to apply a weighting factor to the data, but in this model, an easier parameter approximation method called Maximum Likelihood Estimation (MLE) is used. MLE says that the parameters β and η can be determined using the following equations [12]:

$$\left[\frac{\sum_{i=1}^{n}\left(x_i^{\beta}\ln(x_i)\right)}{\sum_{i=1}^{n}x_i^{\beta}}-\frac{1}{\beta}\right]-\frac{1}{n}\sum_{i=1}^{n}\ln(x_i)=0,\ \text{and} \qquad (4.2.1)$$

$$\eta^{\beta} = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^{\beta}\right] \tag{4.2.2}$$

where we use Equation 4.2.1 to iteratively solve for β and then use 4.2.2 to solve for η. Then this Weibull distribution is used to estimate the damage that will be done by the X% histogram, by using the inverse Weibull distribution:

$$D = \eta\left\{Log\left(\frac{1}{1-\alpha}\right)\right\}^{\frac{1}{\beta}} \tag{4.2.3}$$

where α is the probability. So, for instance, in estimation of the damage in the 99.9% histogram, α=.999.

A correlation analysis was then conducted to determine the linear correlation between the number of cycles in each damage regime and the total damage done, for each test. Then, knowing an estimate of the damage, and these correlation coefficients, an estimate is made on how many cycles will be in each regime for that worst case loading, using the theory presented in Section 3.2.4. The results of a correlation analysis for the 5 random histograms

The most important piece of information from Table 4.1 is the predicted number of cycles, and the details of how these values are actually calculated is presented in the Appendix.in this example are presented in Table 4.1.

Table 4.1 Results of correlation analysis for ATV data, for 99.99% histogram.

	Damage					
Test #	**Reg 1**	**Reg 2**	**Reg 3**	**Reg 4**	**Reg 5**	**TOTAL**
1	1.69E-07	1.31E-06	4.90E-06	7.54E-06	0	**1.39E-05**
2	9.61E-08	6.48E-07	4.15E-06	2.55E-06	3.47E-06	**1.09E-05**
3	8.61E-08	7.74E-07	2.59E-06	5.34E-06	2.15E-05	**3.03E-05**
4	1.48E-07	1.84E-06	7.46E-06	1.23E-05	5.49E-05	**7.66E-05**
5	6.02E-08	4.74E-07	5.30E-06	1.92E-05	2.68E-05	**5.18E-05**
Avg	1.12E-07	1.01E-06	4.88E-06	9.38E-06	2.13E-05	3.67E-05
StDev	4.52E-08	5.59E-07	1.78E-06	6.53E-06	2.20E-05	2.76E-05

	Cycles					
Test #	**Reg 1**	**Reg 2**	**Reg 3**	**Reg 4**	**Reg 5**	**TOTAL**
1	5605	32	14	3	0	**5654**
2	3114	30	14	3	1	**3162**
3	3911	38	15	8	4	**3976**
4	6238	71	28	13	7	**6357**
5	3352	40	24	11	2	**3429**
Avg	4444	42.2	19	7.6	2.8	4515.6
StDev	1397.48	16.62	6.56	4.56	2.77	1413.41

Correlation Coefficient:	0.43066	0.9215	0.9715	0.9686	0.8426
	Reg 1	**Reg 2**	**Reg 3**	**Reg 4**	**Reg 5**
Prediction of Cycles:	7430	118	51	30	14

Estimated Total Damage at 99.99%:	**0.000174**

Finally, a separate Monte Carlo simulation is run for each regime, where the number of cycles generated in that regime is equal to the number of cycles predicted in the correlation analysis.

The first simulation that was run with these five ATV tests was to find the prediction of the histogram of the most extreme loading in as set of 19 histograms, or the 94.7% histogram. The histogram generated in this simulation is presented in Figure 4.15, along with the actual most damaging histogram in the set of 19 ATV tests. A better

comparison of the damage resulting from these histograms can be seen in the exceedance diagram, Figure 4.16

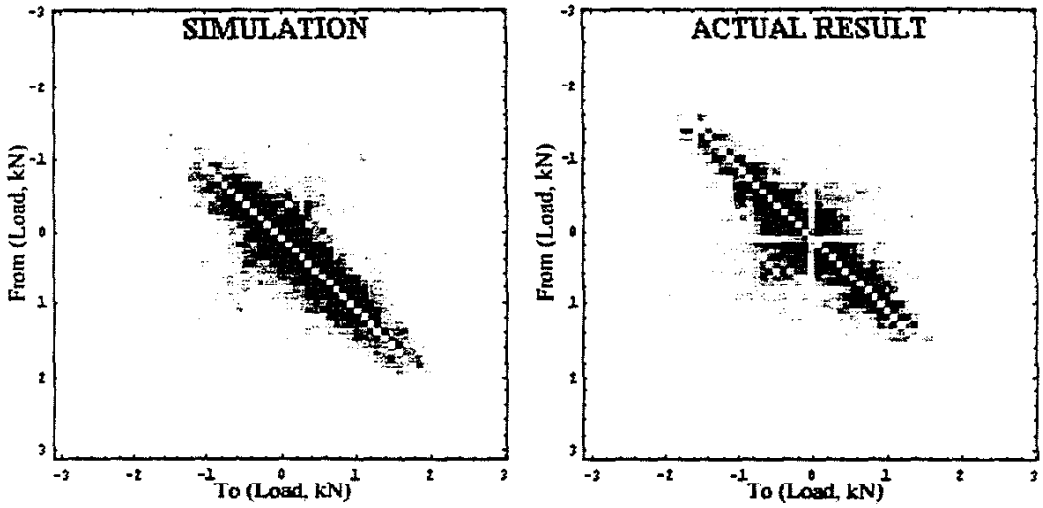


Figure 4.15 Comparison of the model's prediction of the 94.7% loading history, and the most damaging histogram in the set of 19 ATV durability tests.

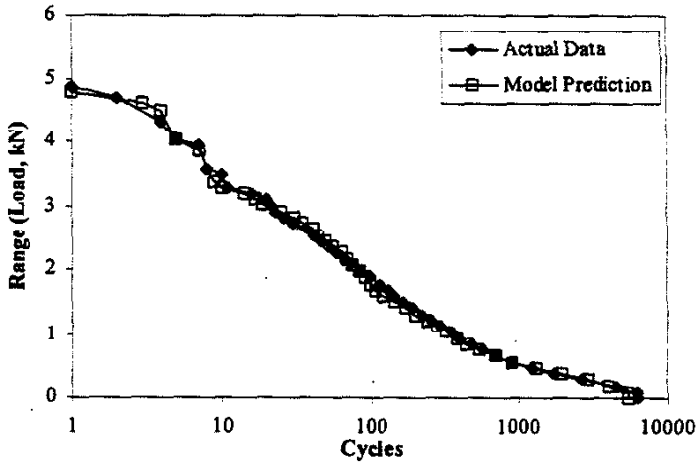


Figure 4.16 Comparison of the exceedance diagrams of the histograms presented in Figure 4.15

The results of this simulation are good, but this ATV problem was the easiest of the three examples to model, because of the three examples, the ATV tests had the least variability. The results of the other two examples are presented later in this section, but

before moving on to those problems, it was interesting to see what this model predicts for the histogram corresponding to the 99%, 99.9% and 99.99% loadings when given all 19 data sets. The resulting histograms are shown in Figure 4.17, and the exceedance diagrams are shown in Figure 4.18.

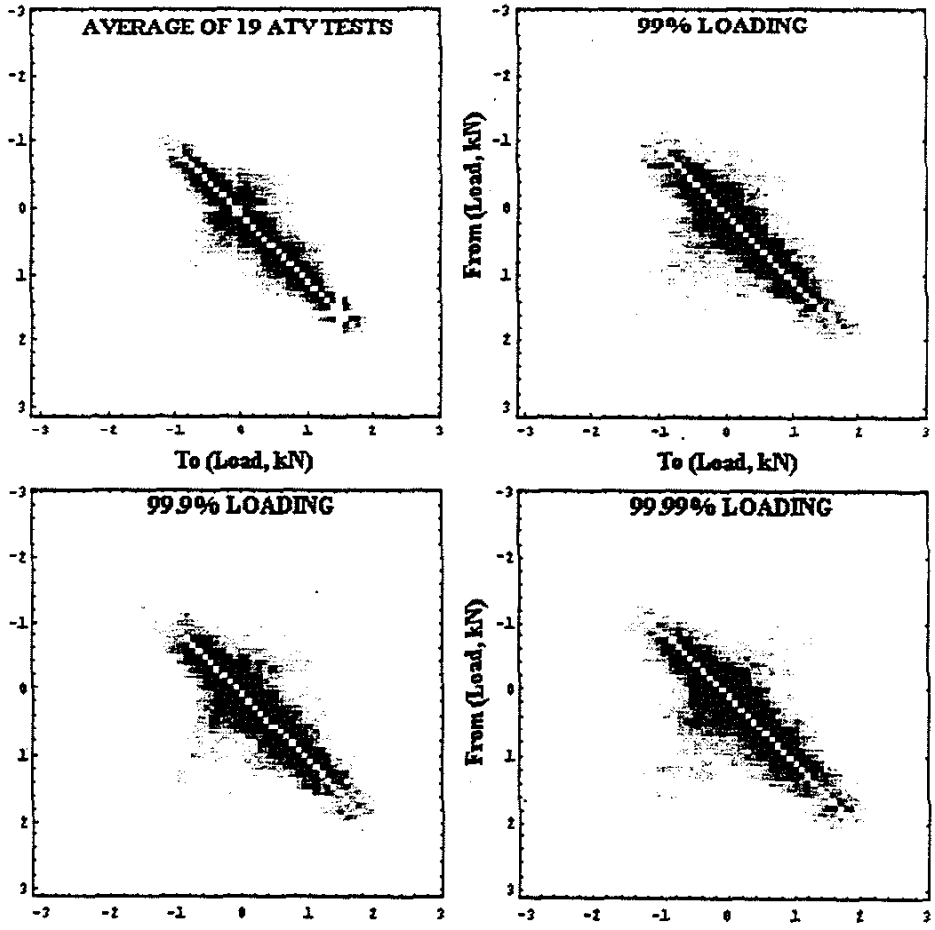


Figure 4.17 Simulations of the 99%, 99.9% and 99.99% loadings, when given all 19 ATV tests.

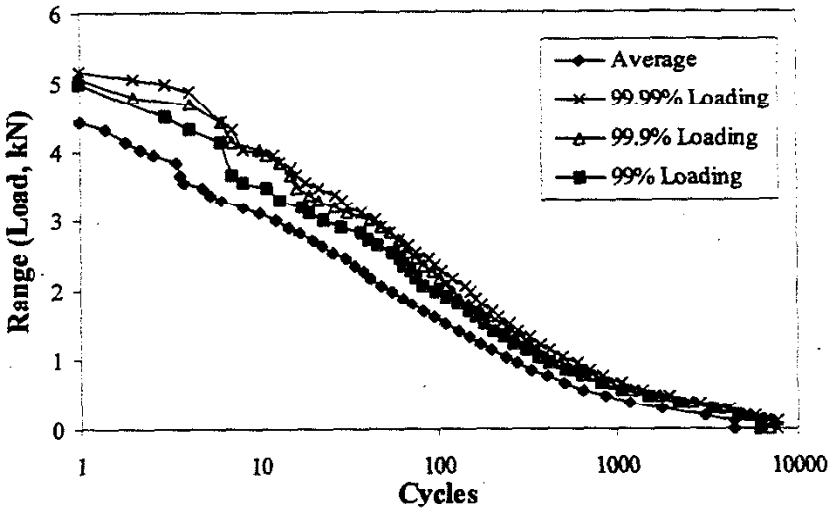


Figure 4.18 Exceedance diagrams for the four histograms in Figure 4.17.

The difference between the three loadings is not visible by looking at the histograms, but the exceedance diagram shows that there is indeed a difference between them.

4.2.2 Analysis of Airplane Durability Data

The second example problem examined in this section involves the 334 airplane loading histories described earlier. This set has the second largest variability of the three example problems.

This set is slightly different in that the data is stored in peak-valley format rather than the from-to format that all of the previous data has been presented in, thus only half of the histogram is populated. The model used in this problem is the same as the model presented in the previous problem, with the only difference being that the statistical analysis was only performed on half of the matrix.

Statistical theory suggests that to approximate statistical charactersitics of a very large set of normally distributed data, it is necessary to randomly sample only 30 data points in the set. In this example problem, even though the data is not normally distributed, it's assumed that a sampling size of 30 can describe the data set of 334 tests. The premise of this problem is to randomly sample 30 of the 334 data sets, and from those 30 sets, predict the histogram that is generated by the most damaging of the 334 data sets.

The procedures used to conduct this simulation are the same as those used in the previous example problem, so we will skip over the procedures and just present the results obtained by the model. The result of attempting to model the loadings that would be generated by the most damaging history in a group of 334 when given a set of 30 histories is presented in Figure 4.19 along with the actual histogram of the most damaging test in our set of 334 airplane histories. The exceedance diagrams of both histograms are shown in Figure 4.20.

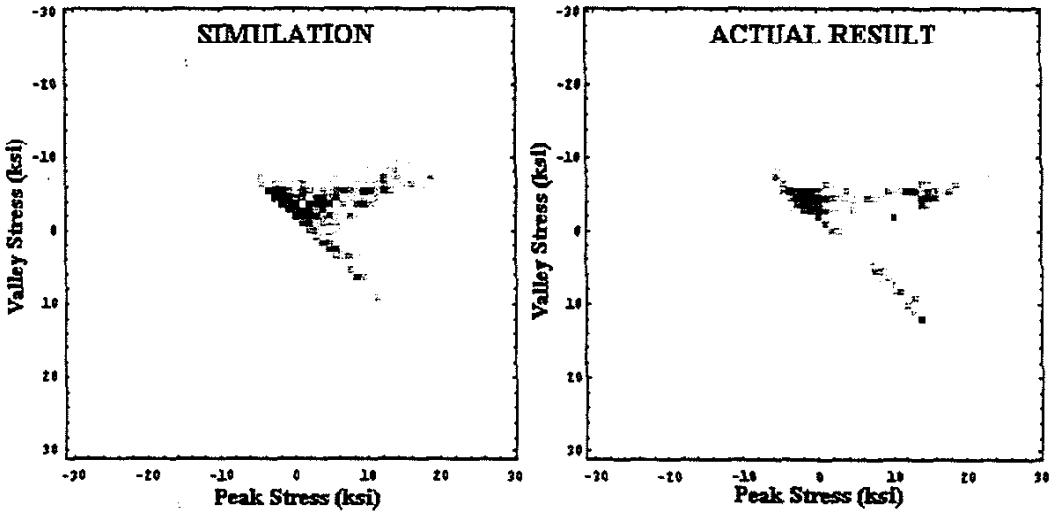


Figure 4.19 Result of simulation of the most extreme history expected in a set of 334 histories, when given a set of 30 histories.

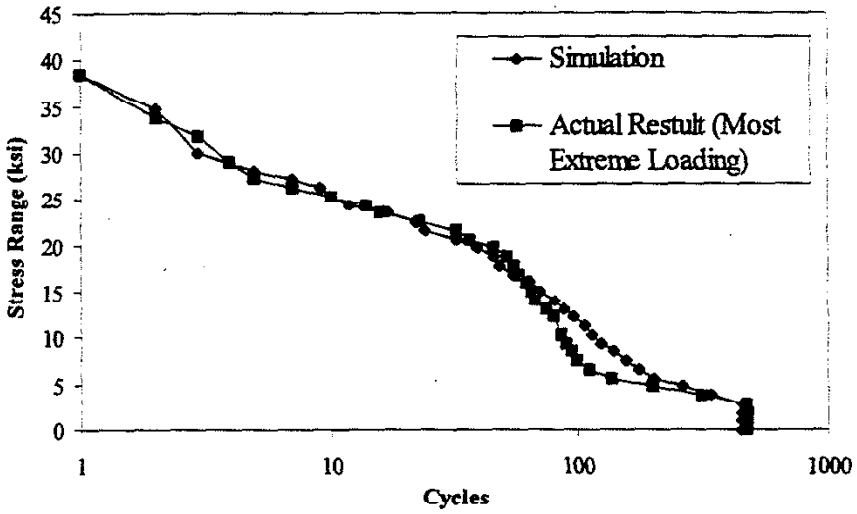


Figure 4.20 Exceedance diagrams of results presented in Figure 4.20

The results obtained here are again very good, with the major difference being in the number of low-load cycles predicted, and because those small cycles do little or no damage, they are of minimal consequence in the modeling process.

Finally, the last example problem is that involving the tractor histories, and because this data set has the most variance, we expect that it will be the most difficult to model.

4.2.3 Analysis of Tractor Durability Data

This tractor data consists of 54 histograms, and the idea of this problem is that given a subset of these histograms, we'd like the model to again predict the most extreme loading in this set of 54 histories. In this example, a random sampling of 15 of the data sets was assumed to give a good representation of the data set as a whole.

Next, 15 histograms were randomly selected, and the model was run on those data sets to predict the most damaging histogram in a set of 54 histograms. Again, the procedures of running the model are the same as those used in the first example problem, so only the results will be presented for this problem. The histogram generated by the model prediction and the actual histogram are shown in Figure 4.21.

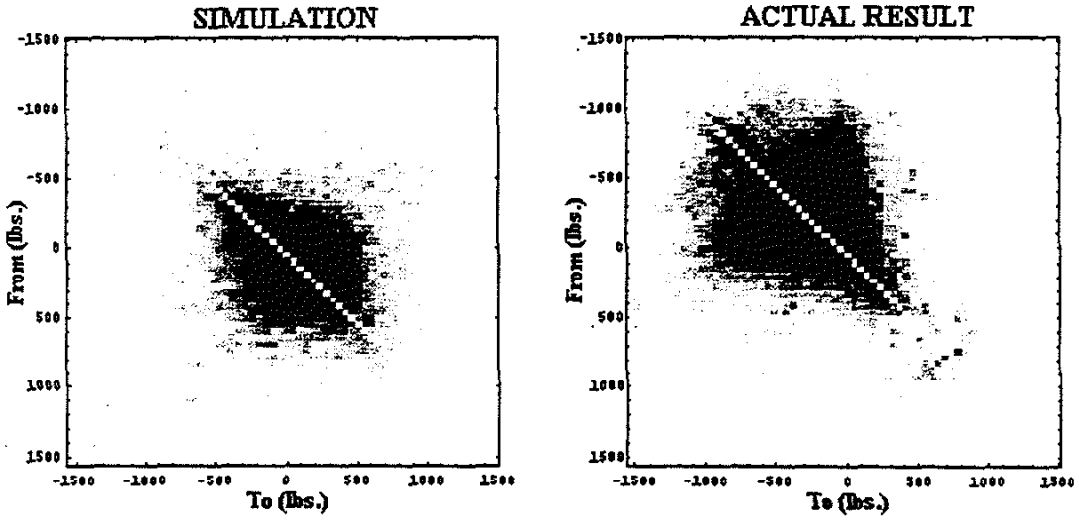


Figure 4.21 Model prediction of the most damaging histogram in a set of 54, given 15 histograms, versus the actual most damaging histogram.

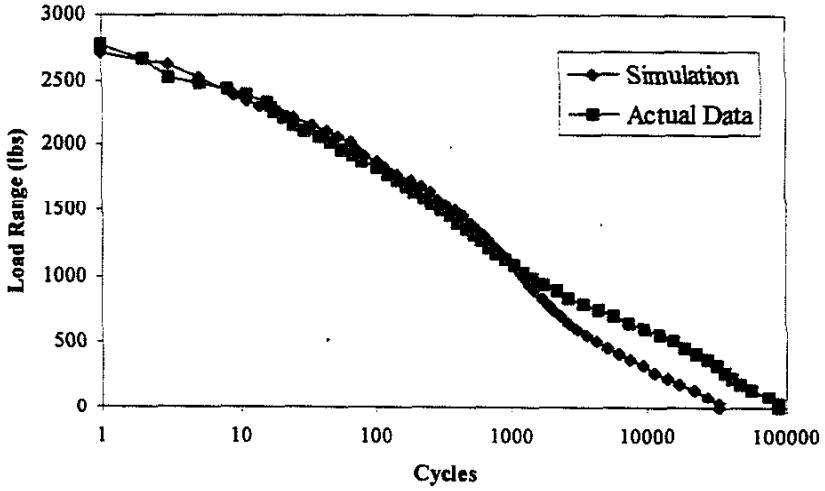


Figure 4.22 Exceedance diagram of the histograms presented in Figure 4.21

Although the damage total is similar between the actual histogram and the predicted histogram, as can be seen in the Weibull probability plot in Figure 4.23 below, it is clear that they go about accumulating that damage in somewhat different ways.

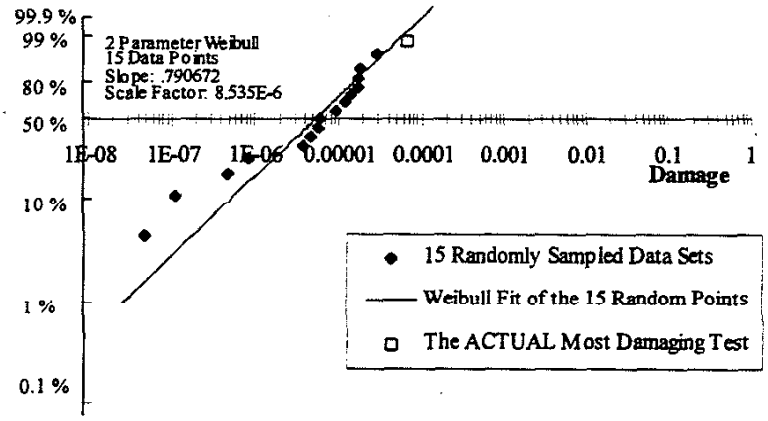


Figure 4.23 Weibull Probability Plot of damage, based on the 15 randomly selected tests.

Based on looking at all 54 data sets, it appears that it would be impossible to predict that most damaging histogram, because it seems to be an anomaly. That is to say that the histogram of that 54th data set is inconsistent with the other 53 histograms in the set. Because of this anomaly, an attempt was made to instead predict the 53rd most damaging histogram in the set. The results of this simulation are shown in Figures 4.24 and 4.25.

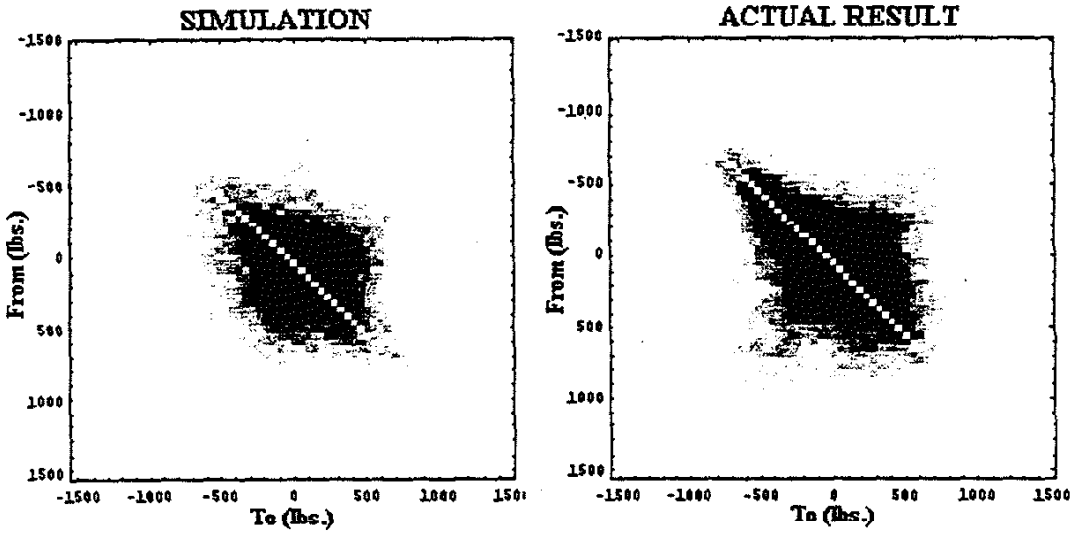


Figure 4.24 Simulation of the second most damaging histogram in a set of 54 when given 15 histograms, versus the actual result.

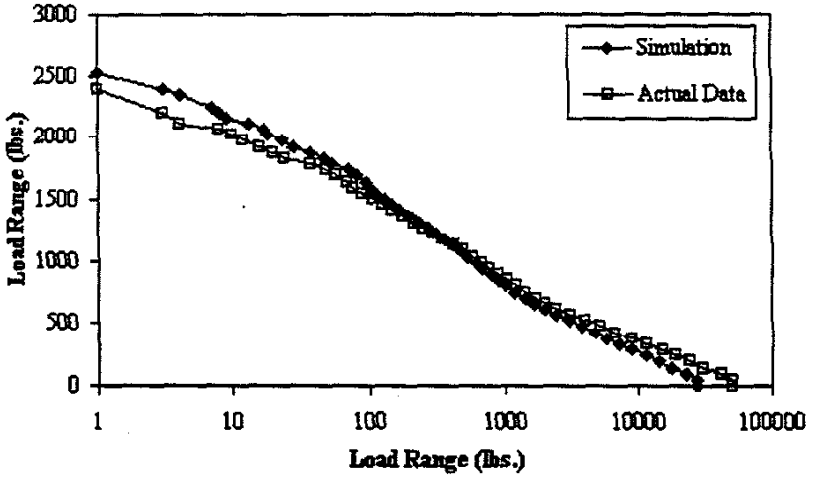


Figure 4.25 Exceedance diagrams of the histograms in Figure 4.24

While the predicted histogram of the 53rd most damaging test appears to be a better prediction of the actual histogram than is the predicted histogram of the most damaging test, the exceedance diagram of the first prediction appears to be a closer approximation than does Figure 4.25. It is difficult to judge which result is more accurate, but both results are very encouraging.

4.3 Discussion of Results and Future Recommendations

The results presented in this chapter are very encouraging. Both models appear to be reliable, but obviously more testing must be done to confirm their validity. Constructing a model to predict random events is a difficult proposition. There will never be a totally correct model because there is never an absolute answer to the question of how a loading history will react. The key is instead to construct a model that takes into account all of the subtle nuances of the histogram, and uses these underlying statistical properties to try to predict future results. The models presented here attempted to do this thru the use of the kernel method, by using theoretical probability distributions such as the Weibull distribution, and by the use of the principles of linear correlation.

In the models presented here, there are several flaws that immediately make the model less than perfect. The first is that the kernel method has problems with bounded domains, and while the reflective technique is a satisfactory solution with the histories that were modeled in Chapter 4, it is not perfectly sound in a statistical sense. There are probably loading cases where the reflective technique will cause some distortion in the density estimate.

Another potential problem in the model is that the kernel which was used is a radially symmetric kernel, and using a radially symmetric kernel can at times cause the simulated data to take on a spherical distribution. This is because of the fact that a radially symmetric kernel is best used on a radially symmetric data set, but when the data set under investigation has an unknown distribution, such as is the case with these rainflow histograms, it is best to use a radially symmetric kernel. The spherical distribution that is at times created is visible in Figures 4.21 and 4.24, when comparing

the simulations to the actual data. One approach at correcting this problem is to pre-scale the data, and one technique in particular is to linearly transform the data set with to obtain a unit covariance matrix, then use a radially symmetric kernel to obtain the density function, and then linearly transform that density function back [13]. While this is not a trivial procedure, it would ensure that the data being analyzed is perfectly radially symmetric.

The main conclusion of this work is that using non-parametric density estimation for the purpose of statistically modeling, and thereby extrapolating, rainflow histograms is in general a valid technique.

4.3.1 Recommendations for Future Work

There are much more involved mathematical and statistical techniques that could have been explored in the construction of this model, but part of the premise of the model is that it should be a model that an engineer can understand, and it shouldn't require a degree in mathematics to be able to implement it. However, it would be worthwhile to look into using kernels of varying shapes, and see how the results differ. If more mathematical work is of interest, it would also be worthwhile to try some work with spline functions, or in particular, Triogram modeling [14]. The best recommendation, however, would be to keep working with the kernel method, and find ways to optimize the techniques that have been presented in this thesis.

5. CONCLUSIONS

Procedures for the statistical modeling of rainflow histograms have been presented, along with the theory of density estimation from which those models were constructed. The first model that was presented was designed to extrapolate a single histogram to predict the loading results that could be expected if that test were carried out over a longer time period. The second model was designed to assess the variability in a set of histograms and from that variability, construct a histogram that is representative of the most extreme histogram in a larger but similar set of histograms.

Results of both models were presented in Chapter 4, and those results are very encouraging. The first model was used to simulate a small amount of field test data, and these results were very good. While the model was tested on a small amount of data, there is no reason to believe that the results obtained here cannot be duplicated on a much larger set of data.

The second model was tested on three sets of data, varying in size from 19 tests to 334 tests, with each set having a different variability in life estimation. Again, the results obtained by this model were very good. As could be expected, the best results were obtained from the data set with the least variability, and the results from the data set with the most variability proved to be the most difficult to judge. In any event, the results obtained by this model for all three data sets were very promising.

If further testing verifies that these models are representative of the real statistical properties of rainflow histograms, the potential decrease in the need for in-service testing is certainly noteworthy.

APPENDIX

The following is Table 4.1, which was originally presented in Section 4.2.1:

Table 4.1 Results of correlation analysis for ATV data, for 99.99% histogram.

	Damage					
Test #	**Reg 1**	**Reg 2**	**Reg 3**	**Reg 4**	**Reg 5**	**TOTAL**
1	1.69E-07	1.31E-06	4.90E-06	7.54E-06	0	**1.39E-05**
2	9.61E-08	6.48E-07	4.15E-06	2.55E-06	3.47E-06	**1.09E-05**
3	8.61E-08	7.74E-07	2.59E-06	5.34E-06	2.15E-05	**3.03E-05**
4	1.48E-07	1.84E-06	7.46E-06	1.23E-05	5.49E-05	**7.66E-05**
5	6.02E-08	4.74E-07	5.30E-06	1.92E-05	2.68E-05	**5.18E-05**
Avg	1.12E-07	1.01E-06	4.88E-06	9.38E-06	2.13E-05	3.67E-05
StDev	4.52E-08	5.59E-07	1.78E-06	6.53E-06	2.20E-05	2.76E-05

	Cycles					
Test #	**Reg 1**	**Reg 2**	**Reg 3**	**Reg 4**	**Reg 5**	**TOTAL**
1	5605	32	14	3	0	**5654**
2	3114	30	14	3	1	**3162**
3	3911	38	15	8	4	**3976**
4	6238	71	28	13	7	**6357**
5	3352	40	24	11	2	**3429**
Avg	4444	42.2	19	7.6	2.8	4515.6
StDev	1397.48	16.62	6.56	4.56	2.77	1413.41

Correlation Coefficient:	0.43066	0.9215	0.9715	0.9686	0.8426
	Reg 1	**Reg 2**	**Reg 3**	**Reg 4**	**Reg 5**
Prediction of Cycles:	7430	118	51	30	14

Estimated Total Damage at 99.99%:	**0.000174**

The values in the upper half of the table, that is both the damage values and the number of cycles, are determined directly from the data in the histogram. Then the linear correlation between the total damage, and the number of cycles in each region is determined using equation 3.2.3. For instance, the correlation coefficient between the total damage and the number of cycles in region 5 is determined in the following manner:

$$\rho_{region5} = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{N}}{\sqrt{\left(\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{N}\right)\left(\sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{N}\right)}}$$

where X_i are the total damage values, and Y_i are the number of cycles in region 5.

Using descriptive statistics, as presented in Section 2.3.1.1,

$$\sum X_i Y_i = (1.39E-5*0)+(1.09E-5*1)+(3.03E-5*4)+(7.66E-5*7)+(5.18E-5*2)$$
$$= 7.72E-4$$

$$\sum Xi = 1.39E-5+1.09E-5+3.03E-5+7.66E-5+5.18E-5 = 1.84E-4$$

$$\sum X_i^2 = (1.39E-5)^2+(1.09E-5)^2+(3.03E-5)^2+(7.66E-5)^2+(5.18E-5)^2 = 9.78E-9$$

$$\sum Y_i = 0+1+4+7+2 = 14$$

$$\sum Y_i^2 = (0)^2+(1)^2+(4)^2+(7)^2+(2)^2 = 70$$

and N=5.

Finally,

$$\rho_{region5} = \frac{\left[7.72E-4-\frac{(1.84E-4)(14)}{5}\right]}{\sqrt{\left(9.78E-9-\frac{(1.84E-4)^2}{5}\right)\left(70-\frac{(14)^2}{5}\right)}} = 0.84.$$

Once the correlation coefficient is calculated for each region, its value is used in equation 3.2.4.

$$E(Y|X=x) = \mu_Y + \rho\left(\frac{\sigma_Y}{\sigma_X}\right)\left(x-\mu_X\right) \qquad (3.2.4)$$

This equation tells us what the average number of cycles will be in each region, given that the total damage is known. In this calculation, since we are attempting to describe more

than just the 5 tests that were make up the sampled population, we use the standard deviation as described in Section 2.3.1.1.

For the example shown in Table 4.1, which is an estimate of the 99.99% usage, the total damage estimate is made using a Weibull distribution, and in this example, the estimated total damage is .000174. Then the estimate of the number of cycles in each region is made. An example of this calculation is shown below. We know the following:

$$\rho = 0.84$$

$$\mu_Y = \overline{\text{number of cycles in region 5}} = 2.8$$

$$\mu_X = \overline{\text{total damage values}} = 3.68E-5$$

σ_Y = StDev(number of cycles in region 5) = 2.77 (shown in Table 4.1)

σ_X = StDev(total damage values) = 2.76E-5, and

x = estimated damage = .000174 .

So, using equation 3.2.4, we calculate the number of cycles in region 5:

$$E(Y|x = .000174) = 2.8 + (0.84)\left(\frac{2.77}{2.76E-5}\right)(.000174 - 3.68E-5) \approx 14$$

This same procedure is carried out for each of the regions, to obtain the predicted number of cycles in each region, which is in the boxed region in Table 4.1.

LIST OF REFERENCES

[1] A. H-S. Ang and W. H. Tang. *Probability Concepts in Engineering planning and Design. Volume I - Basic Principles*. Wiley, 1975.

[2] R. M. Bethea, and R. R. Rhinehart. *Applied Engineering Statistics*. Marcel Dekker, 1991.

[3] D. W. Scott. *Multivariate Density Estimation. Theory, Practice and Visualization*. Wiley, 1992.

[4] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[5] T. Cacoullos. "Estimation of a Multivariate Density," *Ann. Inst. Statist. Math.*, **18**, 179-789, 1966.

[6] V. A. Epanechnikov. "Nonparametric Estimation of a Multidimensional Probability Density," *Theor, Probab. Appl.*, **14**, 153-158, 1969.

[7] I. S. Abramson. "On Bandwidth Variation in Kernel Estimators - A Square Root Law," *Ann. Statist.*, **10**, 1217-1223, 1982.

[8] D. F. Socie and K. Park. "Analytical Descriptions of Service Loading Suitable for Fatigue Analysis," Proceedings of the Tenth International Conference on Vehicle Structural Mechanics and CAE, SAE P308, 203-206, 1997.

[9] K. Dreßler, B. Gründer, M. Hack, and V. B. Köttgen. "Extrapolation of Rainflow Matrices." TECMATH GmbH, Germany. Presented at SAE'96, Detroit, USA, Feb 1996.

[10] J. A. Bannantine, J. J. Comer, and J. L. Handrock. *Fundamentals of Metal Fatigue Analysis*. Prentice-Hall, 1990.

[11] K. Park. Modeling Variability in Vehicle Service Loading Histories. Master's Thesis, University of Illinois, 1997.

[12] N. R. Mann, R. E. Schafer and N. D. Singpurwalla. *Methods for Statistical Analysis of Reliability and Life Data*, Wiley, 1974.

[13] K Fukunaga. *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.

[14] M. Hansen, C. Kooperberg and S. Sardy. "Triogram Models," *Journ. Amer. Stat. Assoc.*, (Currently in Print, 1998).