A Statistical Analysis
Of Construction Equipment Repair Costs
Using Field Data & The Cumulative Cost Model

Zane W. Mitchell, Jr.

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Civil Engineering

Michael C. Vorster, Chair
Yvan J. Beliveau
Jesus M. de la Garza
Mario G. Karfakis
Julio C. Martinez
Robert S. Schulman

April 28, 1998
Blacksburg, Virginia

Keywords: Construction Equipment, Earthmoving Equipment, Equipment Economics,
Economic Models, Economic Forecasting

A Statistical Analysis of Construction Equipment Repair Costs
Using Field Data & The Cumulative Cost Model

Zane W. Mitchell, Jr.

(ABSTRACT)

The management of heavy construction equipment is a difficult task. Equipment managers are often called upon to make complex economic decisions involving the machines in their charge. These decisions include those concerning acquisitions, maintenance, repairs, rebuilds, replacements, and retirements. The equipment manager must also be able to forecast internal rental rates for their machinery. Repair and maintenance expenditures can have significant impacts on these economic decisions and forecasts. The purpose of this research was to identify a regression model that can adequately represent repair costs in terms of machine age in cumulative hours of use. The study was conducted using field data on 270 heavy construction machines from four different companies. Nineteen different linear and transformed non-linear models were evaluated. A second-order polynomial expression was selected as the best. It was demonstrated how this expression could be incorporated in the *Cumulative Cost Model* developed by Vorster where it can be used to identify optimum economic decisions. It was also demonstrated how equipment managers could form their own regression equations using standard spreadsheet and database software.

Dedication

For my family

Acknowledgements

CONTENTS

FIGURES

Tables

Terms

CCI = Cumulative Cost Index

CCM = Cumulative Cost Model

D = depreciation

DMCL = Defender's Minimum Cost Life

EMCL = Equivalent Marginal Cost Life

E_p = expenditures for the period

GEL = Gross Expenditure Line

L* = optimum economic life

L_t = machine age at time t

NEL = Net Expenditure Line

PP_0 = Purchase Price

R_p = revenues for the period

S_t = salvage value at time t

t = the time of interest

T* = minimum value for URL gradient, or optimum average cost

TRL = Total Revenue Line

URL = Uniform Recovery Line

CHAPTER 1: INTRODUCTION

The management of heavy construction equipment is a difficult task. The equipment manager is called upon to serve as leader, resource manager, accountant, engineer, arbitrator, policy maker, and seer. The goal of this research is to identify and describe decision support tools that the equipment manager can use to reduce some of the uncertainty in decisions made concerning heavy equipment. By doing this, it is hoped that some of the seemingly "crystal ball" based decisions occurring in the day-to-day management of equipment operations can be replace with modern, statistically sound techniques. Valuable insight into the way that construction equipment deteriorates with use can also be obtained.

The purpose of this chapter is to provide the reader with an introduction to the topic of the dissertation. The problem will be introduced and defined. The hypotheses, objectives, methodology, scope, limitations, and assumptions of the research will be briefly discussed. Finally, an outline of the dissertation will be presented.

1.1   THE TOPIC

It is important that the reader have an understanding of basics concerning the management of heavy construction equipment. This section will provide an introduction to the principles and vernacular of the field. The discussion will funnel from the general to the specific. Three areas that are of particular concern to this dissertation are: Construction Equipment, Equipment Economics, and Equipment Data.

1.1.1   Construction Equipment

The function of heavy earthmoving equipment is to move or assist in the moving of soil and rock from point A to point B. The purchase of this equipment constitutes a particularly large investment on the part of the buyer. One cannot get into the business of owning this type of equipment without substantial cash reserves and/or financial backing. Most machines cost at

least $100,000—the largest pieces of equipment can cost millions of dollars. Owners of this equipment have a vested interest in insuring that it is properly used, maintained, and managed. Firms that use heavy earthmoving equipment fall into two major categories: mining companies and construction companies. Although the applications these machines perform within these two types of companies may seem similar, the conditions are very different. Mining machines perform the same task under pretty much the same conditions—day in and day out. Operations and management of the equipment usually take place in the same geographic location. Things are different in the construction industry. The machines can be called upon to do varied tasks in different locations under dissimilar conditions. Construction equipment can sit idle in a storage yard if its owner has not won the bid for any projects for it to work on—this usually does not happen in mining ventures. Most construction firms have some sort of centralized equipment management function, but actual operations are widely scattered—in some cases spanning the entire country. This research will focus on construction equipment. Parallels may be drawn to earthmoving machines that are used in mines, but that is not the purpose of this study.

Construction equipment is not a fixed asset—its value is consumed in the production of work. The ultimate goal of this work is to make a profit for the owner—if there is no profit, there is no point in owning the equipment. There are a finite number of passes that an excavator can make and a finite number trips a dump truck can make and still make profits for their owners. Machines are routinely bought, operated, and sold during the normal course of business.

There is an endless cycle of decisions that must be made with respect to equipment ownership. The equipment manager must decide how much and how often regarding routine *preventive maintenance*. Preventive maintenance is defined as those routine, periodic actions undertaken to minimize repair costs or extend the life of the machine—oil changes are a good example. *Repair* decisions occur on the next level. When the machine or one of its components breaks down during the normal course of business, it must be fixed to regain operational status. *Rebuild* decisions concern major mechanical refurbishments that extend the life of the machine. When a machine is nearing the end of its profitable life, the equipment manager must make a *replace* decision. Most of these decisions are multi-faceted. They will be explained in greater detail in Chapter 3.

The decisions described above are of an *economic* nature. They fall under the purview of making the investment as profitable as possible. There are two other classes of decisions that are often made concerning heavy equipment. The first class contains those decisions of an *operational* nature—how to get the most production out of the equipment. The second class is that of *mechanical* decisions—how to ensure the reliability of the equipment. This dissertation will focus primarily on equipment economics.

1.1.2 Equipment Economics

As mentioned above, there are three phases in the life cycle of an earthmoving machine: **buy, operate, and sell**. The buy decision comes once in the life of each machine—the equipment manager should strive to buy as infrequently as possible due the tremendous capital expense involved. Operate decisions occur on a frequent basis after the purchase of the machine—the goal is to operate the equipment as cheaply as possible maintaining suitable productivity. The sell decision may be evaluated more than once, but is only taken to "yes" one time in the life of each machine—the machine should be sold at as high a price as possible.

Taken individually, the three separate economic decisions might not be too difficult to comprehend and process. But, there is a complex dynamic between the three. Each can have a tremendous impact on the others. Even though it is very expensive to buy new machinery, operating costs are very low early in a machine's life. As operating costs increase, the sell decision should start to be considered. There is no simple answer.

The buy and sell decision combine to help define *owning* costs. Owning costs are those costs that accrue or have accrued just to have the potential of using a machine. Other inputs besides buy and sell are costs such as insurance or taxes. Owning costs are best characterized on a calendar basis—they accrue whether or not the machine is used. The longer a piece of equipment is kept, the cheaper the average owning cost per period becomes. Conversely, if the machine is kept a short period of time, the average cost of ownership per period can be relatively large due to the fact that new machines loose value very quickly in the early periods.

The use of piece of equipment generates a constant stream of *operating* costs. These are costs that occur on a day-to-day basis in the course of running a machine. If the machine sits idle, operating costs can go to almost nil. If the machine is used heavily, operating costs can climb quite high. These costs are best-defined using some metric that characterizes units of work. Typically, they are tracked by hours of operation. Some operating costs are frequent and small, such as fuel and maintenance. Other expenditures occur on a more periodic basis and can be fairly big—like tires, repairs, and rebuilds. Average operating costs are low when a machine is new. As it ages average operating costs tend to climb.

The decrease in owning costs with the concurrent increase in operating costs gives rise to the notion of *economic life*. There is, theoretically, an optimum age at which to replace a machine. This age is the age at which point the combination of average owning and operating costs is minimized. To properly analyze economic life, one must be armed with detailed knowledge of the composition and behavior of owning and operating costs. Owning costs are not that difficult to understand and quantify. They are composed of purchase price, resale price, licenses, insurance, taxes, and interest. Operating costs are complex and very data intensive. There is a constant stream of data associated with the operating cost of each piece of equipment. If this stream is properly tracked and analyzed, it can be a reliable input into the economic modeling process.

1.1.3   Equipment Data

Nearly all firms that use heavy equipment have some means of tracking its costs and usage. Specific data formats vary greatly from company to company, but there are some key elements of data that are kept in one form or another by nearly all companies. The initial data associated with the purchase of a machine is usually quite easy to record and extract. The purchase price is known before the machine is purchased and all other owning costs are tracked by the accounting function of the firm.

All periodic operating costs are normally recorded in one form or another—this is a necessary part of doing business. In order to run a business well, expenses must be tracked in order to subtract them from revenue when tax time comes. If expenses aren't well tracked the company could pay more taxes than it should and hence make less profit than it should. Usually parts and

labor involved with repairing a machine are tracked in separate accounts. Some firms break expenses down into further subdivided accounts that correspond to the major components of the machines. Expenses are usually recorded when they occur but are reported on a monthly basis.

Most firms also track "hours" worked for each machine. The definition of "hours" varies from company to company and will be discussed in more detail in Chapter 4. Also, there is usually some measure of the reliability of the machine that is tracked. Often, this is in the form of down hours, which is the time during which the machine was unavailable for production because of a mechanical problem.

Data collection methods are as varied as the companies that use them. Some use detailed computerized work-order systems that track every expense related to a machine, which components or sub-components were repaired, who performed the repairs, and how long it took. These work orders are sent to the main computer as they are closed out. Other companies rely on weekly faxes from field mechanics to let them know the quantity of parts and labor costs that should be charged to each machine. Some require that actual hour meter readings are taken on a periodic basis—others rely on hours of use that are reported from each job superintendent on a weekly basis.

Eventually, all these data find their way into large accounting databases. This is the root of most problems that equipment managers have with their data management systems. The systems were designed for accountants, not equipment managers. Mainframe computers that have huge storage capacities usually host these programs. Access to the databases is strictly controlled. Typically, 2-3 years worth of data associated with every aspect of the company is maintained on the mainframe computer. Older data are archived on tape reels or (more recently) CD-ROMs for later retrieval if needed.

Data retrieval is accomplished via an interface with the host computer. One must be conversant in the language of the mainframe computer or have at their disposal someone who is. As mentioned above, the databases were designed with accountants in mind—not equipment managers. All costs are pigeonholed into tidy accounts, but sometimes these accounts can contribute little to the effective management of construction equipment. If the data that are needed have been placed in

archives someone must go to the storage location and retrieve them. Sometimes the costs associated with obtaining archived data are higher than the benefits that can be obtained by using them.

Once they are recovered, using the data for other than standard accounting-type functions usually requires a great deal of spreadsheet gymnastics. Often, accounting reports come in two extremes—the very generalized report that is so general trends are hard to spot and the very detailed accounting code report that is detailed to the point that the data make little sense. But, the chain of expenditures that comprises operating costs can usually be reconstructed with varying degrees of effort. The topic of this dissertation is how to better use the data products available in the course of making economic decisions concerning heavy equipment.

1.2   THE PROBLEM

It has been shown that the economic decisions equipment managers are faced with can be quite complex. There is an interactive effect between owning costs and operating costs that cannot be ignored when searching for an optimal solution. Operating costs are important to consider, both in their timing and in their magnitude. The periodic usage and accounting data maintained by construction companies can be used to produce a stream of data that defines operating costs.

This understanding aside, there is still considerable debate about the life and cost of construction equipment. The economic models proposed in the literature are very simplistic, very old, and very broad in scope. Additionally, the statistical bases for most of these models are unknown. These models are seldom, if ever, used in practice.

Most equipment managers are very knowledgeable about the management of equipment, but don't truly know how to make the most of the information they have. Unfortunately, data does not equate to information. Many equipment managers can make little use of the vast resource of data that is at their disposal. They are "data rich but information poor" (Kapoor, 1996). Instead of applying sound economic theories and using statistical trends to their full capabilities, they rely more upon rules-of-thumb and good judgement. This is not meant as an affront to experience and

good judgement. Some equipment managers are quite successful in the economic decisions they make on a daily basis—economic models are seldom a suitable replacement for common sense.

The point of this dissertation is that it can be done better. Tools can be developed and employed which will improve the economic decision making capabilities of equipment managers. This topic is relevant—it contributes to knowledge and it addresses real world problems.

1.3 THE CHALLENGE

There are really two challenges associated with improving the decision-making tools that are in place for equipment managers. The first challenge is a theoretical one. A sound conceptual model must exist that can be applied across a spectrum of economic decisions. The second challenge is to develop a statistically sound methodology to support the model. The methodology should allow construction companies to employ the data they already collect to quantify variables in the model.

The first challenge has been largely met. The *Cumulative Cost Model* proposed by Vorster (1980) is a valid economic model that can serve this purpose. It can be manipulated to provide numeric and easy to understand graphical solutions to nearly every economic decision that equipment managers must make. This model is described in-depth in chapter 3 of this dissertation.

The second challenge will form the bulk of the contribution that this dissertation makes to the body of knowledge. Specifically, a methodology will be developed that will enable equipment managers to quantify variables which describe how operating costs vary over time.

1.4 HYPOTHESES

This dissertation will test three different hypotheses. These hypotheses are interrelated—they build upon each other. The validity of the first is a precondition for the validity of the second just as the validity of the second is a precondition for the validity of the third. It is a building block approach to a complex problem. Figure 1-1 shows how each of the three hypotheses relate to each other.

1.4.1   Hypothesis #1

A mathematical relationship exists between repair costs and age of heavy earthmoving equipment.

This relationship can be described in a relatively simple form, such as:

$$C_r = a + bx + cx^2 + dx^3 ... e^x$$

Equation 1-1

Where:

C_r = cumulative cost of repairs

a, b, c, d = numeric coefficients

x = age of machine

e = base of natural logarithms

The equation listed above is an example. The true equation will be developed in the dissertation and may be of a different form.

1.4.2   Hypothesis #2

It is possible to approximate the true equation for the relationship between cost and age by using linear regression techniques on existing data.

Actual data from construction firms that use earthmoving equipment will be used in a rigorous statistical analysis to determine which regressor terms are important to describing the behavior of costs with age. Terms that are not important will be eliminated. The study will be limited to linear models or non-linear models that can be transformed into linear models.
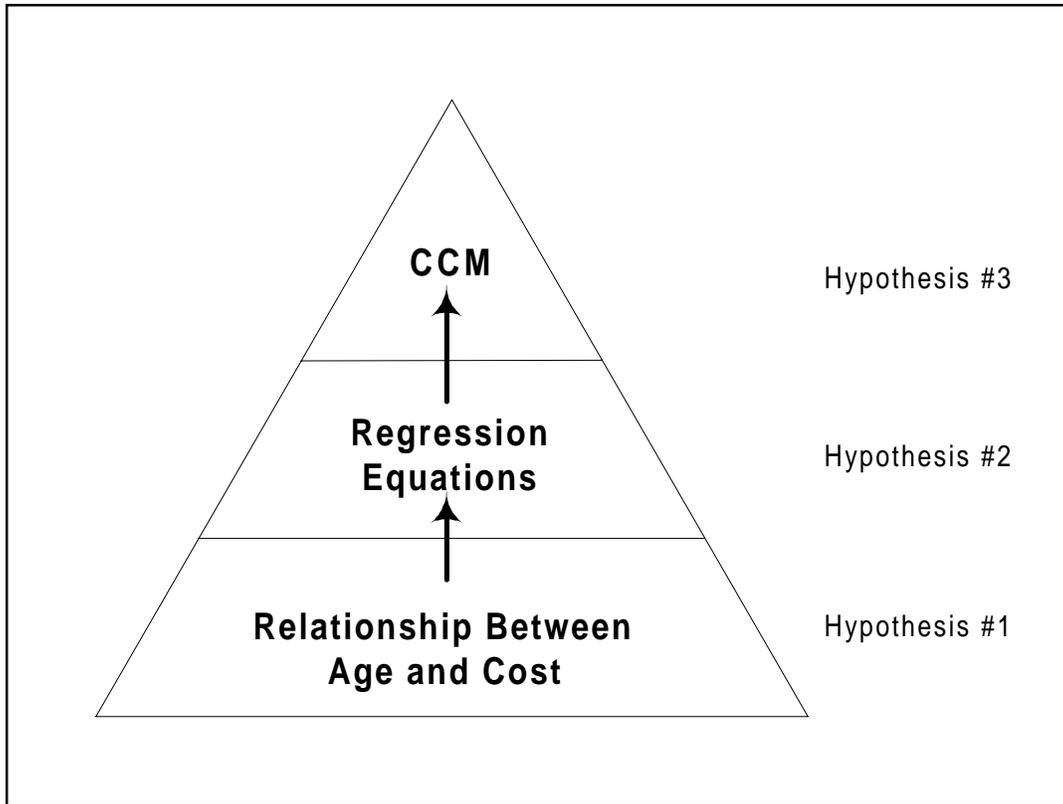
Figure 1-1: Objectives of the Dissertation

1.4.3 Hypothesis #3

It is possible to incorporate repair cost regression equations into the Cumulative Cost Model (CCM).

The CCM cannot be properly used until its basic components are defined. By combining the regression repair cost equations with other known economic costs associated with owning and operating equipment, equations for heavy equipment can be obtained.

1.5 RESEARCH OBJECTIVES

There are four objectives that will be attained to accomplish this research:

1. Data pertaining to maintenance and repair of heavy construction equipment will be collected and normalized.

2. A statistical methodology will be developed which:

   —uses the field data collected

   —shows which regressors are important when defining repair costs in terms of machine age

   —determines the values of those regressors that are significant

3. A methodology for incorporating the regression equations into the CCM will be developed and described. This will make it possible to describe the algebraic expression for the *Cumulative Cost Index* (CCI) where :

$$CCI_t = \frac{\sum_{0}^{t} \text{Gross Expenditures}}{\text{Purchase Price}_0}$$ Equation 1-2

   The line described by the above equation is also the *Gross Expenditure Line* (GEL) of the CCM in terms of the CCI.

4. It will be illustrated how the CCM can be used to aid in the decision making process concerning equipment economics.

The first objective is a routine requirement. The second objective—to develop and test a methodology—is the primary objective of the research. The third and fourth objectives—to implement the methodology in the CCM are secondary objectives.

 The dissertation will not define industry standard norms for the values of the regressor variables. Some comparisons will be drawn concerning whether different companies have similar equations and whether different equipment types and sizes have different equations. It will be shown how the methodology can be converted to practice.
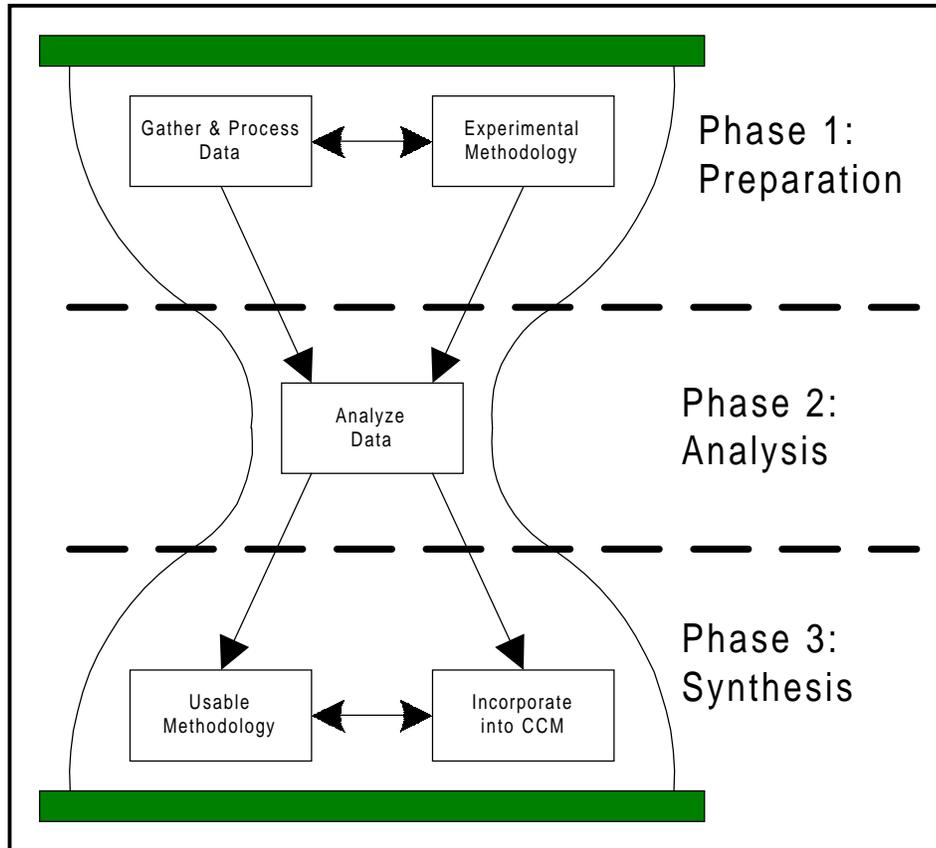
Figure 1-2: Flow of the Research

1.6   METHODOLOGY

The methodology that will be used to accomplish the objectives listed above can be divided into three distinct phases comprising five distinct tasks. The phases are: *preparation, analysis, and synthesis.* The tasks are: *gather and process data, develop test methodology, analysis, develop usable methodology, and incorporated into CCM.* These phases, tasks, and how they relate to each other are depicted in Figure 1-2.

1.6.1   Preparation

There is a certain amount of ground work that must be accomplished before any analysis can really get underway. Two steps that must be accomplished: first, the data must be gathered and processed to put it into a form suitable for analysis and, second, a test methodology must be defined for the use of this data.

In gathering and preparing the data, it is important to acknowledge up front that this is a *field study*. Since this research is based on a field study rather than a laboratory study it must be recognized that there will be a certain amount of "noise" present in the data. Had the study been conducted under laboratory conditions, much of the spurious information could have been eliminated. A tradeoff is made when choosing a field study over a laboratory study. The field study should yield a model that is closer to the way things are in reality, but variables over which the researcher has no control over can have an influence on the data. The laboratory study would have yielded a model in which all parameters could have been controlled--everything that had an impact on the data could have been quantified. However, the laboratory study may not have yielded models that are reflections of the way things really happen.

There are structural and statistical issues concerning the data that require resolution and explanation, these will be covered in Chapter 4. This step cannot really proceed without the granting of access to the data from the desired companies. Once permission is obtained, the origins and limitations of the data will be investigated.

The test methodology must be sufficiently rigorous to give a good statistical feel for how well the various models perform on the field data. This test methodology will be discussed in detail in Chapter 5.

The above two tasks are highly inter-related. The test methodology must be designed so that it can make the best use of the field data that is available. On the flip side, sound statistical practice should not be abandoned to come up with a methodology that is appropriate for sub-standard data. If a company's data do not meet some minimum structural requirements, they will not be considered in the primary analysis.

1.6.2   Analysis

Although there is only one major task that is a portion of this phase of the research (analysis), it can be further divided into two sub-tasks: preliminary analysis and secondary analysis.

Before any analyses can take place, the data must be placed into the proper format. This formidable task will be described in Chapter 6.

The preliminary analysis will be concerned with finding out what regression equations best characterize the growth of costs with respect to increase in age. This will be done through a variety of different regressions and tests on the prepared field data. This part of the analysis will be discussed in Chapter 7.

The secondary analysis will be to draw inferences concerning the results obtained in the preliminary analysis. Are there differences between different types of equipment within a company? Are there differences among similar types of equipment between companies? Is there one set of parameter values that fits every machine in every company? These questions will be answered in Chapter 8. Additionally, comparisons will be made to hypothetical results that would have been achieved using other methods of cost forecasting described in literature.

1.6.3   Synthesis

The purpose of the synthesis is to take the analysis to a different plane. There are two major tasks in the synthesis phase of the research: defining a usable methodology and incorporation into the CCM.

The usable methodology must be defined such that equipment managers can develop cumulative cost curves using commonly available applications for personal computers. The process will be described in general and developed in detail for one spreadsheet program. The usable methodology should approximate the results of the experimental methodology. For companies that do not have good data collection processes, a database and data collection scheme will be described. These topics will be discussed in detail in Chapter 9.

The final task in the methodology is the incorporation of the curves into the CCM. Suggestions for combining the operating cost curves with other costs will be provided. The use of the CCM to solve equipment related problems will be demonstrated. This will also be covered in Chapter 9.

The usable methodology must be developed so that it produces equations that are compatible with the cumulative cost model. The incorporation and usage of the equations within the CCM is highly dependent upon their accuracy.

1.7   SCOPE & LIMITATIONS

1.7.1   Scope

In order to achieve the objectives listed in section 1.5, four different companies were visited and data was gathered on their equipment fleets. An equipment fleet is defined as a group of machines of the same size and type within the same company. The data were analyzed, equations were produced that related the direct costs of maintenance and repair to cumulative hours of use, and appropriate comparisons were made. A complete methodology was documented for use by construction companies for the replication of this process and the production of their own equations. The methodology for incorporating these equations into the cumulative cost model was also documented.

1.7.2   Limitations

This dissertation will not address every aspect of the maintenance and repair cost estimating problem. It will only investigate the relationship between repair costs and machine age. As such only two variables will be part of the regression equations: machine age in hours and direct costs expressed within the CCI. Other important aspects, such as quantifying the cost of downtime will not be covered.

This work is also limited in that it will analyze historical data from a relatively small number of companies. The companies have been chosen to provide a cross-section of heavy construction firms in the United States. This does not necessarily mean that every firm type, size, geographic

region, or management style is represented. Every construction company is unique. The study is limited to the construction industry—mining applications will not be investigated.

Not all equipment categories will be modeled. Equipment categories describe their general function, or type. Not all classes within each category will be modeled. Classes describe the weight, horsepower, or size of equipment within its category. The categories and classes that will be analyzed are machines that are fairly common throughout the industry. The CCI values will be calculated for machines that are like types. Like types are not exactly similar. To allow for differing purchase prices, the GEL will be expressed in terms of the CCI as expressed in equation 1-2. Only one definition of CCI will be used.

Industry standard parameters will not be developed. Inferences will be drawn concerning some equipment types and sizes but these will be observations and are not intended to be definitive.

1.8 ASSUMPTIONS

The following assumptions were made at the beginning of this project. All are reasonable and define the context within which this work should be taken. Detailed explanations of the assumptions follow the listing.

1. The data are representative of construction equipment in general and the given type or group in particular.

2. The data were collected in a reliable manner.

3. Each company is striving for the same level of service from their equipment.

4. The cumulative hours of use on the machines is the only regressor variable.

5. The response variable, cumulative maintenance and repair cost, follows a normal statistical distribution centered about the regression equation over the range of cumulative hours worked that is investigated.

6. The variance of the response variable is assumed constant throughout the lifespan of the equipment in those cases where not enough data are present to justify a variance analysis study.

7. The cumulative repair costs on a given machine are zero when there are zero cumulative hours of use on the machine.

The data are representative of the equipment in general. Statistics is not an exact science. No statistical tool can consistently predict exact results for specific observations. The best that can be hoped for is a model that will estimate *average* repair costs for a group of machines consistently over the lifespan of these machines. Trends of individual machines can be analyzed, but it must be recognized that it is possible for individual machines to fall outside the confidence intervals developed for classes of machinery. Any inferences drawn or conclusions made rest upon the assumption that the models developed can be applied to all machines that are similar to a given type or group.

The data were collected in a reliable manner. In a perfect world, researchers would have enough money and time to operate their own fleets of equipment in carefully monitored environments to control every aspect of their experiments. This type of experiment is not possible within the scope of this research. However, a number of companies have shown a willingness to provide access to the data that *they* have collected. In essence, the experiment has already been completed. It must be assumed that the data collected by the companies are complete and accurate. It is not possible to go back in time and verify all expenditures—the records that exist have to be trusted. The trustworthiness of these records will be verified by visiting the companies involved.

Each company is striving for the same level of service from their equipment. It is reasonable to assume that each of the companies investigated is in business to make a profit. Given that they are in business to make a profit, they should each be striving for essentially the same level of service from the equipment that they own. This does not mean that each company has the same equipment maintenance policy. It simply means that they each adhere to some minimum standard

of preventive maintenance. This is what will allow a comparison of classes of equipment between different companies.

The cumulative hours of use on the machines is the only regressor variable. There are many variables that can be factors in estimating maintenance and repair costs. The regression analyses being performed are done assuming that all of these other factors are constant for all the machines in the group being studied. This simplification is necessary in order to be able to accomplish the analyses. It is certainly a reasonable assumption for machines from the same company that worked the same region of the country. It may not be as reasonable for comparing machines that came from different companies.

The response variable, cumulative maintenance and repair cost, is normally distributed throughout the range of cumulative hours worked that is investigated. Normality of the data is an assumption that must be valid in order to perform normal hypothesis testing and construction of confidence intervals. Many of the fleets that we will be analyzing are relatively small. They are so small that tests for normality of data may be inconclusive. The normality assumption is reasonable—many processes that occur naturally come close to being normally distributed (Schulman, 1996).

The variance of the response variable is assumed constant in those cases where not enough data are present to justify a variance analysis study. As will be discussed later in this document, it is expected that the data to be analyzed will have variance that increases with increasing cumulative hours worked. Simply put, this means that all new machines will have almost the same hourly repair costs but old machines will have repair costs that can differ quite a bit from machine to machine. Accurately quantifying this variance function can be very difficult with small data sets. The nature of regression analysis is such that applying the wrong variance correction factors can be much worse than applying no correction factors at all. Because of this, with small fleets we will assume constant variance.

The cumulative repair costs on a given machine are zero when there are zero cumulative hours of use on the machine. This assumption is reasonable and necessary. If for some reason a brand

new machine had required repairs before its first job, the cost of these repairs should have been covered by the manufacturer or by insurance.

1.9   ORGANIZATION OF THE DISSERTATION

This dissertation is organized into four distinct but interrelated parts. Figure 1-3 depicts these four parts as they relate to each other and the chapters of the dissertation.

1.9.1   Part I: Understanding the Challenge

Part I provides the frame of reference and context for the dissertation. It consists of the first two chapters of the dissertation.

- *Chapter 1* is the introduction.

- *Chapter 2* is the literature review. The literature review is fairly extensive in that it covers both the history of economic replacement models and the estimation of maintenance and repair costs and. This chapter is crucial to the research that is undertaken—without context it has little meaning.

- *Chapter 3* is a detailed discussion of the cumulative cost model. The basic model to be used in this study is presented and explained.

The outcome at the completion of this block will be an understanding of the aspects of equipment management, economic forecasting, and economic modeling that are pertinent to this research.

1.9.2   Part II: Defining The Work

Part II focuses on the model building and analysis definition aspects of this dissertation. The statistical analyses should produce valid results if conclusions based on those results are to be of merit.

Part I	Understanding the Challenge	1. Introduction 2. Literature Review 3. The Cumulative Cost Model
Part II	Defining The Work	4. The Data 5. Statistical Methodology
Part III	The Work	6. Data Gathering 7. Analysis 8. Results
Part IV	The Benefits	9. Integration 10. Conclusion & Recommendations
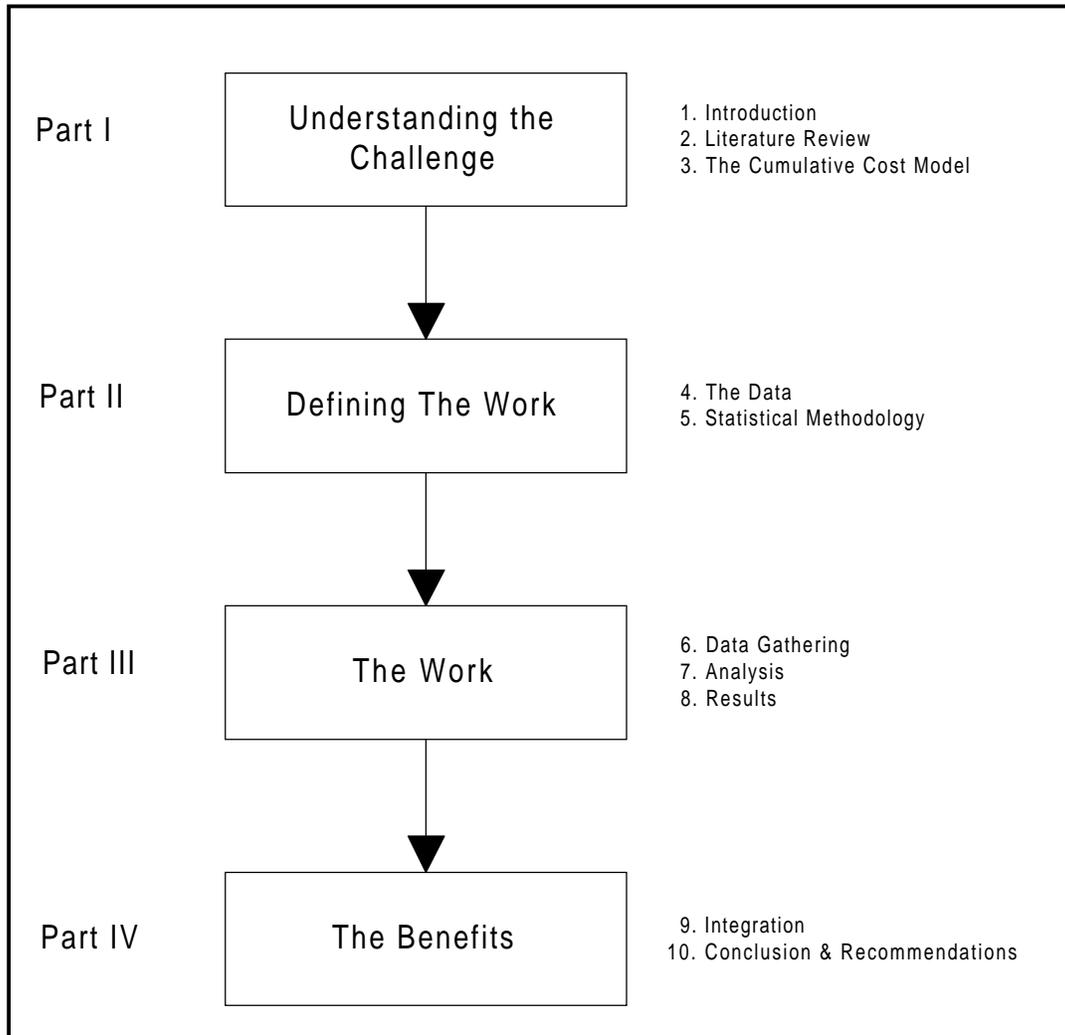
Figure 1-3: The Organization of the Dissertation

- *Chapter 4* provides the reader with a detailed understanding of the data involved with this study.

- *Chapter 5* provides the statistical theory and methodology used to analyze the data.

This section will provide the understanding needed for the statistical analysis to proceed.

1.9.3   Part III: The Work

This section of the dissertation describes the work undertaken to perform the statistical analysis and produce the results obtained. It consists of three chapters:

- *Chapter 6* highlights the data gathering operation.

- *Chapter 7* describes the analyses that took place.

- *Chapter 8* analyzes the results with respect to actual performance and other forecasting methods.

The outcome of this section of the dissertation will be an understanding of the nature of regression equations relating repair cost to equipment age. It contributes to the body of knowledge by defining and testing a statistically sound methodology for determining the equation for the Gross Expenditure Line (GEL) of the CCM.

1.9.4   Part IV: The Benefits

This is the portion of the project upon which the rest of the project is judged. Part IV synthesizes the results obtained in Part III.

- *Chapter 9* provides detailed instructions on how to use the GELs derived in Chapter 8 in the CCM to make strategic decisions concerning heavy equipment. This chapter also explains how companies can apply the cumulative repair cost equations described in Chapter 8 to define the GELs for their own equipment

- *Chapter 10* summarizes and recaps the dissertation. Areas for further study are described.

The outcome of this section will be the dissertation's contribution to the body of knowledge concerning equipment economics.

1.10   SUMMARY

This chapter was meant to serve as an introduction and a road map of the work that follows. It is the first step in link to understanding the challenge. There are probably many questions remaining in the reader's mind about the specifics of this research. These questions will hopefully be answered in the following chapters.

The next chapter is the Literature Review. In that chapter, the reader will be given detailed background information on replacement economic models and repair cost forecasting. It is the second and pivotal chapter of the section *Understanding the Challenge*. Comprehension of the basic theories involved is critical to full understanding of the impact of this research.

CHAPTER 2: LITERATURE REVIEW

This chapter provides an understanding of the basic aspects of the problems involved with making economic decisions by reviewing work that has already been accomplished in this arena.

This chapter will follow the following format:

- The historical development of engineering economic analyses that led to the genesis of the cumulative cost model (CCM) will be discussed.

- The literature that exists concerning the forecasting of equipment repair costs will be documented.

- The literature concerning the forecasting of maintenance and repair costs will be discussed.

2.1   ECONOMIC REPLACEMENT THEORY

Decisions about heavy equipment should be made based on sound economic principles, not emotions or intuition (Douglas, 1975). Economic replacement theory models attempt to answer the question: "What is the optimum economic life of this piece of equipment?" The goal is to find an optimum length of service for a given machine. After this time has expired, there is at least one other alternative (replace, retire, rebuild, etc.) which is more economical than keeping the machine in its present state. The models attempt to find the optimum length of service by using a variety of techniques based on the science of economics.

There are three basic theories in the field of economic replacement that are relevant to an understanding of this dissertation.   They are: *the cost minimization model, the profit maximization model, and the repair limit model*. There are many other names for equipment

replacement models in the literature (Jaafari and Matteffy, 1991), but most of them can be categorized as an offshoot of either cost minimization or profit maximization. Cost minimization and profit maximization theories developed on parallel paths beginning in the 1920's. Repair limit theory is relatively new—it was first published in the 1960's.

Throughout this section, the terms "Defender" and "Challenger" will be used (Terborgh, 1949). The Defender is the machine that is currently under study by the company. The Challenger is a new machine that could serve the same purpose as the Defender.

2.1.1   Cost Minimization

The theory of cost minimization can be explained quite well graphically. As mentioned in Chapter 1, most costs associated with a machine can be placed in one of two categories: ownership costs and operating costs. The average *cost of ownership* for a given machine should decrease the longer it is kept. This is because most of the capital costs involved with owning a machine are incurred as soon as it is purchased. As time goes on, the initial purchase price is spread over a longer time span and thus the average cost decreases. The average *cost of operating* a given machine should increase the longer it is kept. For example, when the machine is new repair costs should be relatively small and infrequent. As a machine is operated, repairs become more frequent—and sometimes more costly. Cost minimization strives to find a balance point between decreasing ownership costs and increasing operating costs. The specific components of owning and operating costs will be discussed in detail in Chapter 4. The cost minimization model is depicted graphically in Figure 2-1.

There are three curves depicted: average ownership cost, average operating cost, and average total cost.

$$\text{Average Ownership Cost} = \frac{P_0 - S_t}{L_t} \qquad\qquad \textbf{Equation 2-1}$$

$$\text{Average Operating Cost} = \frac{\sum_0^t E_p}{L_t} \qquad\qquad \textbf{Equation 2-2}$$

Average Cost per period at age $L_t = \dfrac{P_0 + \sum_0^t E_p - S_t}{L_t}$ **Equation 2-3**

Where:

P_0 = initial purchase price

E_p = expenditures for the period

S_t = salvage value at time t

L_t = machine age at time t
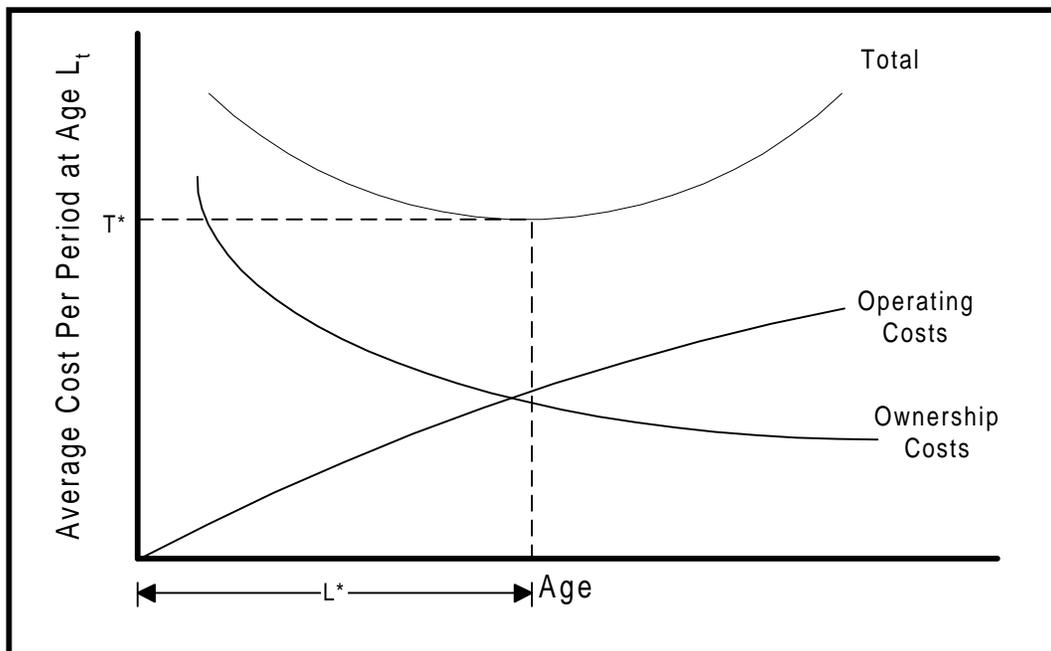
Figure 2-1: The Cost Minimization Model

Average costs are calculated by taking the cumulative costs incurred up to a given point in time and dividing these costs by machine age. Average cost curves are developed for ownership costs and for operating costs. The sum of these two curves, the *average total cost* curve, slopes downward initially when operating costs are low and the average cost of capital is decreasing. The minimum value of average total cost is T*, the point where the slope of the curve is zero. The optimum economic life, L*, is that period which ends when the sum of owning and operating

costs reaches a minimum. Note that the abscissa is labeled "age." Age is a generic term that is well suited to the diverse situations that can present themselves when conducting economic replacement analyses. This concept will be fully developed in Chapter 4.

2.1.2   The Profit Maximization Basic Model

An alternate method to the solution of replacement problems is *profit maximization* (Hotelling, 1925). Figure 2-2 is a graphic depiction of the profit maximization model. Again, three lines are depicted on the chart. They are the average total cost, the average revenue, and the average profit. The *average total cost* line is as described in Section 2.1.1. The *average revenue* is the average amount of income generated by the asset. *Average profit* is determined by subtracting the average cost from average revenue. This results in a curve that is nearly a mirror image of the average cost curve. The optimum economic life occurs at the apex of the average profit curve. If average revenue were constant, the average profit curve would be an *exact* mirror image of the average cost curve and the profit maximization economic life would be the same as the cost minimization economic life. However, the amount of revenue generated by an asset often declines with use as the machine suffers from both deterioration and obsolescence as it ages. For this reason, the economic lives for profit maximization and cost minimization are not always the same (Douglas, 1975). The equations associated with this model are:

$$\text{Average Revenue} = \frac{\sum_{0}^{t} R_p}{L_t} \qquad\qquad \textbf{Equation 2-4}$$

$$\text{Average Profit at time } L_t = \frac{\sum_{0}^{t} R_p - (P_0 + \sum_{0}^{t} E_p - S_t)}{L_t} \qquad\qquad \textbf{Equation 2-5}$$

Where:

R_p = revenues for the period
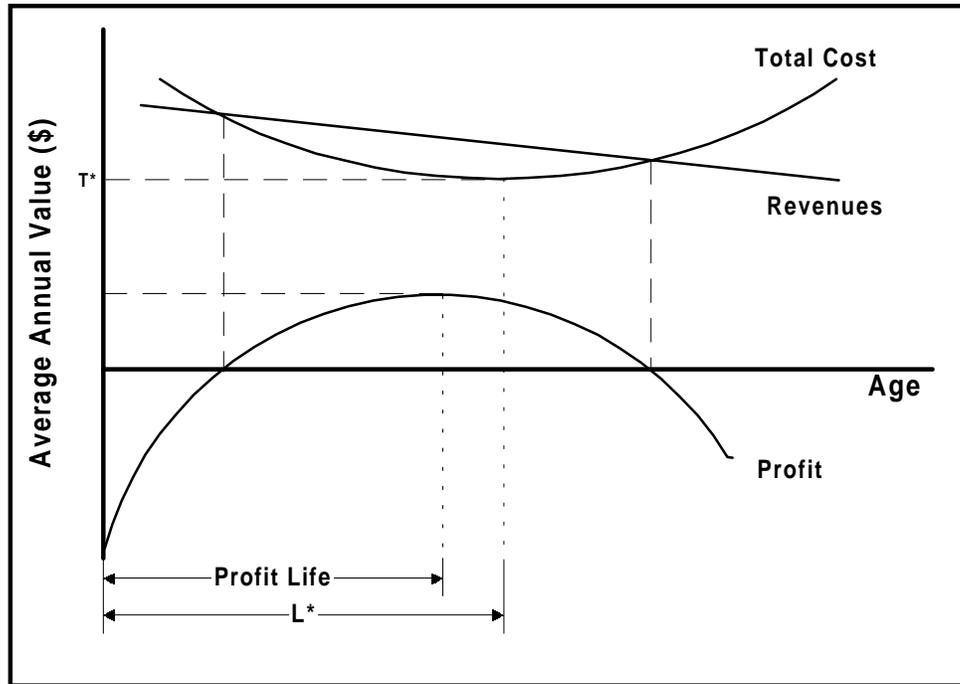
Figure 2-2: The Profit Maximization Model

The minimum average annual cost, T*, and the optimum economic life for cost minimization, L* are also depicted in Figure 2-2. It can be seen that in the case of declining revenues, the optimum life for profit maximization (Profit Life) will be less than L*. The converse is also true.

2.1.3   The Repair Limit Theory

A different way of looking at the economic replacement decision was presented in Drinkwater and Hastings' repair limit theory (1967.) The repair limit was defined as follows:

> "The repair limit is a limit on the amount of money which can be spent on the repair of a vehicle at any particular job. The values of the repair limit are dependent on the type, age, and in some cases on the location of the vehicle."

Repair limit theory is not applied until a machine has broken. The concept behind repair limit theory is that there exists some amount, ro_t, below which it economically sound to repair the machine. If the estimated cost of the repair is greater than ro_t, the repair should not be undertaken and the machine should be discarded or replaced.

The following quantity represents the future cost per year if the machine is repaired (Drinkwater and Hastings, 1967):

$$\frac{r + m(t)}{g(t)}$$

Equation 2-6

where:

r = the cost of the repair in question

m(t) = the expected total cost of future repairs from time t forward

g(t) = the expected remaining life of the machine from time t

t = the time in the machine's life at which point the repair limit evaluation is taking place

If a failed machine is scrapped then the future cost per year is θ, which is found by determining the average future annual cost of the replacement system. The replacement system is either a new copy of the Defender, or a Challenger that is different. The quantity obtained in equation 9 is compared to θ.

If

$$\frac{r + m(t)}{g(t)} < \theta$$

Equation 2-7

The machine should be repaired and returned to service as soon as possible. If the inequality is not true, then the machine should be scrapped and replaced. The repair limit is that value of r for which both sides of the inequality would be equal. Solving for r, the repair limit becomes (Drinkwater and Hastings, 1967):

$$r_0(t) = \left(\theta \times g(t)\right) - m(t)$$

Equation 2-8

This is graphically depicted in Figure 2-3 (Drinkwater and Hastings, 1967). Drinkwater and Hastings also introduced a *cumulative cost curve* to graph economic replacement models. Line

OAD is cumulative repair cost vs. age of the machine. The quantity OA represents the original capital cost. The curve AQPD represents the cumulative repair costs over time. The slope, θ, represents the average cost of similar machines against which the machine of interest is to be judged. This straight, sloping line is tangential to the cumulative cost curve at the point P. The average cumulative repair cost at any point on the cumulative cost curve is given by the slope of the line drawn from that point to origin of the plot.
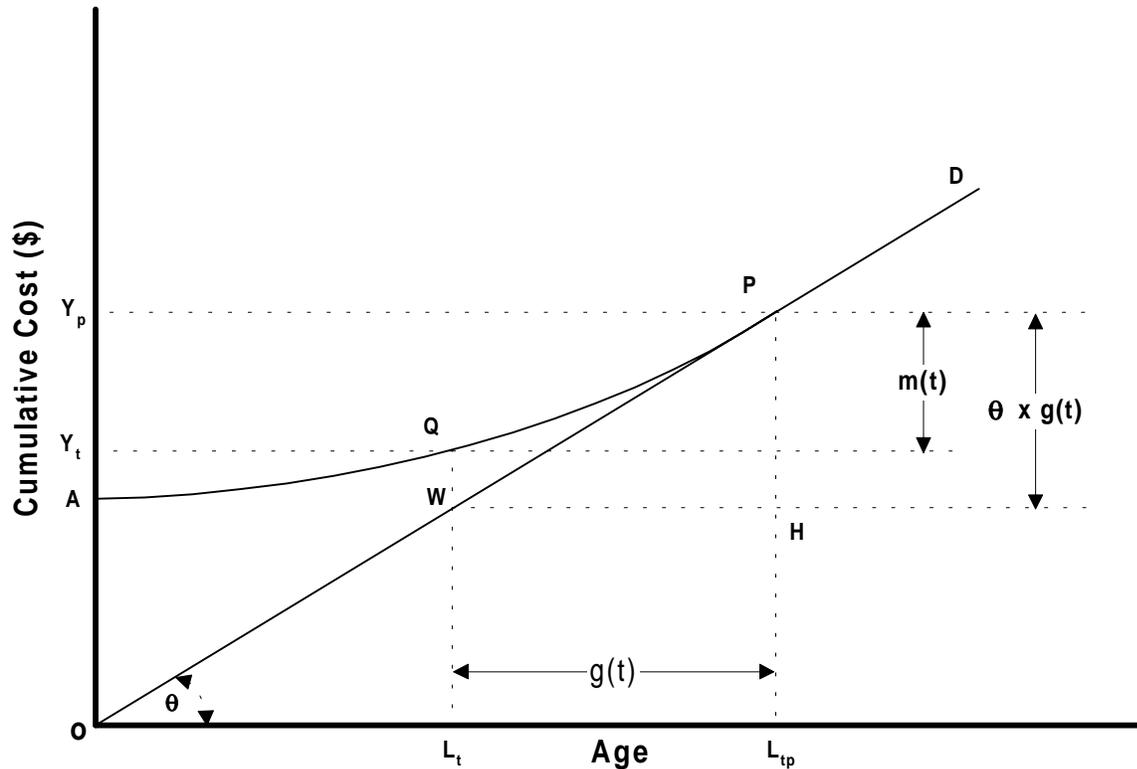
Figure 2-3: The Repair Limit Model (after Drinkwater and Hastings, 1967)

the origin. At a point in the machine's life, L_t, the line QW represents the repair limit. Beyond age L_{tp} the repair limit is zero. The following relationships are depicted on the figure:

$$g(t) = L_{tp} - L_t$$

<div align="right">**Equation 2-9**</div>

$$m(t) = Y_p - Y_t$$

<div align="right">**Equation 2-10**</div>

Repair limit theory is limited in that the model supports only one type of decision. The theory cannot be applied until a machine breaks down. Repair limit theory was revisited by Mahon and Bailey (1975) but the basic concept remains unchanged.

2.1.4 Summary

Three different economic models have been reviewed in this section. The output of each is distinct. One seeks to minimize costs, one seeks to maximize profits, and the third seeks to define a function that specifies a repair spending cap at any given point in a machine's life. None of the three is particularly well-suited to accommodating the rationale and mechanics of the others. Despite the differences, all three attempt to answer the same question: "What is the optimum time to sell?"

A model exists that can be used to emulate the mechanics of all three of the above mentioned models. The Cumulative Cost Model combines the important concepts developed in each of these theories into one package. Chapter 3 will provide an in-depth discussion of this model as developed by Vorster (1980).

2.2   IMPORTANT WORKS CONCERNING REPLACEMENT THEORY

Section 2.1 provided a basic understanding of the mechanics of economic replacement theory. This section is meant to expand upon that understanding with discussions of particularly influential works in the arena.

2.2.1   Taylor

Taylor published the paper that forms the nucleus of most modern day economic replacement theory in 1923. He defined useful (economic) life of a machine as the period of time that minimizes the unit cost of production for that machine. If a machine is sold before or after that period has expired, the average unit cost of production will be greater than the optimum unit cost. The equations developed by Taylor for average unit cost, x, over n years are:

$$x = \frac{O_1 + O_2 + \ldots O_n + W_n}{Y_1 + Y_2 + \ldots + Y_n}$$

Equation 2-11

$$W_n = C - S_n$$

Equation 2-12

where:

O_1, O_2, O_n = operating expenses for the 1st, 2nd, and n^{th} year (includes labor, repairs, fuel, etc.)

Y_1, Y_2, Y_n = number of units of output for the 1st, 2nd, and n^{th} year

W_n = cost of the machine new less the salvage value of the machine at the end of the nth year

C = cost of the machine new

S_n = salvage value of the machine in the nth year

To determine the minimum unit cost, equation 6 is applied over each successive year of operation of the machine. The value of x will first decrease then increase. The point at which the value of x has reached its minimum defines the economic life of the machine.

Taylor also presented a parallel analysis called "unit cost plus (interest)" that allowed for the calculation of minimum unit cost while accounting for interest (the time value of money.) This method defined useful (economic) life as the period that at ends at the point in time where "unit cost plus" is minimized. It is interesting to note that Taylor developed these analyses in an attempt to better describe depreciation—defining the concept of economic life was merely a means to that end.

Taylor's analyses were formed with the intent of replacing the Defender with an identical machine—no provisions were made for comparing the Defender to a different machine. Replacement of the Defender would take place when the minimum unit cost (or cost plus) was realized. Taylor implied that his equations could be used for different replacement alternatives (Challengers), but did not articulate how this could be done (Preinreich, 1940).

2.2.2   Hotelling

Hotelling (1925) was the first proponent of profit maximization.   He proposed profit maximization not as a replacement for cost minimization, but as an alternative to cost minimization.   The quantity that Hotelling sought to maximize was the value of the output (revenues) minus the cost associated with producing that output and plus the salvage the value of the machine.   He called this the *value* of the machine.   Hotelling used discounted cash flow techniques to determine this *value*.   A good discussion of discounting techniques can be found in *"Principles of Engineering Economy"* (Grant, et. al., 1990).   Hotelling's equation for *value* in a constant interest scenario is as follows:

$$V(t) = \int_{t}^{n} [xY(\tau) - O(\tau)]v^{\tau-t}d\tau + S(n)v^{n-t}$$

Equation 2-13

where:

    t = the time of interest

    τ = integration variable representing time

    V(t) = the value of the machine at time *t*

    n = the useful life of the machine (this corresponds with L* as described in Section 2.1.2)

    x = the theoretical selling price of a unit of output

    $Y(\tau)$ = output rate of the machine (function of time)

    $O(\tau)$ = operating costs of the machine (function of time)

    v = 1/(1 + i) where *i* is equal to the interest rate

    S(n) = salvage value of the machine (function of the useful life)

Hotelling refined Taylor's approach in a number of ways.   He introduced the use of integral calculus in lieu of algebraic summation to streamline calculations.   Hotelling was the first to discuss obsolescence of machines—although he was particularly vague about how to calculate obsolescence cost.   Like Taylor, Hotelling developed his methodology in the hopes of better

defining the concept of depreciation—determining the useful life of the machine was only described in as much as it furthered that goal.

2.2.3   Preinreich

Preinreich (1940) revisited Taylor's and Hotelling's theories; he also made some important contributions of his own. As with the previous articles, Preinreich was concerned with industrial equipment in general and did not write specifically of construction equipment. However, unlike Taylor and Hotelling, Preinreich overtly addressed the issue of equipment replacement instead of discussing it under the auspices of depreciation. In his words:

> "Replacement is the basic problem, because it actually affects the composition and productivity of a plant. Calculations of depreciation are mere figures entered into books, the significance of which depends entirely on the use to which they are put."

Preinreich recognized that replacement problems are not always as simple as one machine being replaced by another of the same type. He categorized the scope of replacement decisions in five distinct categories:

1. Single machine

2. Finite chain of replacement machines

3. Infinite chain of replacement machines

4. Numerous parallel chains

5. A large plant composed of a number of smaller machines that are replaced as they wear out

The one category of these five that had the biggest impact on the field was that of the infinite chain of replacements. In an infinite chain, the assumption is made that there will be a future need for a specific machine and that all of the replacements for this machine will have similar lives and economics associated with them. This means that the economic life of the current Defender is

impacted upon by the economics of future Defenders (or Challengers.) Preinreich also briefly explained how to account for a technologically improved machine (the Challenger.) However, his method did not provide the means to make a decision between a Challenger and a Defender—his method assumed that the Defender was obsolete and the Challenger was the only replacement option.

2.2.4   Terborgh

George Terborgh (1949) took cost minimization a step further. Terborgh better defined the concepts of deterioration and obsolescence in addition to the aforementioned Defender/Challenger concept. *Deterioration* is the measure of decreased performance of the Defender in relation to a brand new Defender as the equipment gradually wears out. *Obsolescence* is a measure of the lower performance of a brand new Defender in relation to a brand new Challenger. Deterioration and obsolescence taken together form the *inferiority gradient.* The inferiority gradient is essentially Terborgh's version of operating costs as described in Section 2.1.1.

This sum of the operating inferiority and the capital cost is averaged each year of the machine's life—the point at which this sum is minimized is known as the *adverse minimum*. This point corresponds with T* as defined in Section 2.1.1. The period of time that passes between the machine's purchase and the adverse minimum is the optimum economic life and corresponds with L*. If a machine is currently older than its optimum economic life, the new adverse minimum becomes the average sum of the inferiority gradient and cost of capital for the *current* year. Simply put, after the adverse minimum has been reached the best time to replace a machine is the present.

To compare the Defender and the Challenger, the adverse minimum for each is calculated. The machine with the lowest adverse minimum is the one that should be chosen. Terborgh intended that the analysis be accomplished on a yearly incremental basis. Provisions for an analysis with a horizon of greater than one year into the future are described, but not fully developed.   Terborgh revisited and expanded upon this method in 1958 and 1967 (Grant, et. al, 1990.) The revisions consisted of applications other than the replacement decision and a change of methodology to a comparison of internal rates of return instead of adverse minimums.

Terborgh was the first author to articulate the idea of units of production for mobile equipment. Previous authors defined production in very generic terms. Terborgh defined production in terms that can be applied to the construction industry. "Cost per acre cultivated" can easily be translated to cost per cubic yard excavated. "Cost per mile" is similar to cost per meter hour.

2.2.5 Douglas

Douglas (1975) wrote the first book dedicated specifically to construction equipment management. He provided descriptions of three different ways to arrive at a replacement decision: intuition, cost minimization, and profit maximization. Intuition, he said, is the most prevalent method for making replacement decisions. The use of this method has no basis in economic principles; instead it relies on the judgement and experience of the person making the decision. Good judgement and a wealth of experience are very desirable characteristics in an equipment manager. There are certainly some wizened equipment managers in the industry that can consistently make the right choices based on their "gut feel." Analytical methods can complement the intuitive abilities of the best equipment managers. Douglas downplays the validity of decisions reached using intuition. He explained that many equipment managers who make decisions based on intuition are more influenced by the initial purchase price of an item than by the long-term operating costs and reliability.

The second method Douglas describes is cost minimization. Economic life is defined as described in Section 2.1.1 for the basic cost minimization model. Replacement normally occurs at the point in time where the average cumulative cost of the Defender exceeds the minimum average cumulative cost of the Challenger. Although Douglas fully develops an example problem using the cost minimization method, he makes the statement "this (the cost minimization method) is an easy way out and considered by some to be more scientific than the method described above (intuition)." Douglas understood the mechanics of cost minimization, but he did not think very highly of it. He was concerned that those who applied cost minimization theory were not accounting for all costs.

The final method Douglas describes is profit maximization. Annual costs are subtracted from annual revenue to calculate annual profits as described in Section 2.1.2. The average cumulative

profit for each year of analysis is then calculated. When maximizing profits, the optimum economic life is defined as that year in which average cumulative profit is maximized. Figure 2-2 shows this graphically.

Douglas described three different methods for executing the profit maximization technique. The preferred method employed a computer program developed by Douglas specifically for that purpose. Another method enlisted the aid of a slide rule or calculator to help with mathematical calculation. Douglas estimated that it would take almost two years for one person to go through all the calculations to solve one problem manually. Douglas' third method employed look-up tables to ease the burden of some of the manual calculations. According to Douglas, an experienced user could make it through a problem in around two hours using his tables.

Douglas recommended profit maximization over cost minimization and intuition. He implied that a profit maximization policy is better for business than a cost minimization policy. He went further to say that a cost minimization policy should be used only when profits cannot adequately be determined.

2.2.6   Collier and Jacques

Many other authors have attempted to refine the cost minimization/profit maximization model over the years. Most of these refinements have focused on the mechanics of the calculations involved and the definitions of the costs involved.

Collier and Jacques (1984) developed the "Geometric Gradient-to-Infinite-Horizon Method." This method explains in detail how to handle the time-value-of-money calculations for many different cost categories. Cost categories used by the Geometric Gradient-to-Infinite-Horizon Method repair cost, maintenance cost, tire cost, downtime cost, obsolescence cost, accessory cost, taxes and insurance cost, decline in salvage value, and overhaul cost. The types of costs will be described in Chapter 4. Using this method, many of these expenditures are defined in terms of geometric gradients.

Equations are developed to find the net present value of all the life cycle costs associated with the existing Defender, the first replacement Challenger, and all future replacement Challengers. These

costs are summed to find an overall net present value. When this combined net present value is minimized, the optimum replacement strategy has been found. Two components are varied in the net present value equations. These are the remaining life of the Defender, N, and the assumed life of the Challenger, L. The minimum net present value is found through an iterative process that tests all reasonable combinations of values for N and L. The iteration process easily lends itself to computer applications. Jaafari and Mateffy (1991) further refined this method and developed a computer program to implement it.

All in all, the three basic economic theories presented in Section 2.1 are still valid today. The focus of most of the literature over the years has been that of bringing practice closer to theory. The reality is that the theory has not been brought into practice. Much has been written, but little has been applied. Some basic concepts are in use, but no one model has gained industry-wide acceptance. A goal of this research is to develop a format that is more easily understood and applied by practitioners of equipment management.

2.3   ECONOMIC FORECASTING

There are two main aspects to every forecasting problem, the forecasting aspect itself and a planning aspect (Makridakis, et. al., 1989). According to Makridakis, a *forecast* is simply a prediction of what will happen—it is an input into the planning process. A *plan* is something a decision-maker devises with the intent of shaping future events into a favorable outcome. The forecast can be key to the success or failure of the plan, but it is not an end in itself. To understand why it is important to have an accurate way of predicting equipment costs, it is necessary to first have an understanding of what these forecast costs can be used for.

The economic models and their enhancements described in Sections 2.1 and 2.2 make up the planning portion of the decision making process that was described in the introduction of this chapter. The second body of knowledge that must be understood to fully comprehend this dissertation is that of economic forecasting. Economic forecasting is "the study of historical data to discover their underlying tendencies and patterns" (Hanke et. al., 1995). This section will discuss the mechanics of forecasting in general. This will lay the groundwork for detailed

discussions of forecasting as it relates to maintenance and repair costs. Section 2.4 will delve into the particulars of applications developed specifically for equipment management.

Although forecasting has been a science for over a century, the advent of computers has really made forecasting a mainstream activity. It is only recently that personal computers have made it possible for managers at nearly every level of business to analyze data and make forecasts. In the past, these functions were relegated to mainframe computers—before that to the statisticians and bookkeepers.

2.3.1   Uses of Economic Forecasts

Forecasts can be used *as business planning tools, process control devices, and communications vehicles* (Wilson et. al., 1994). The bulk of what was discussed in Section 2.1 concerns itself mainly with the *business planning* aspects of forecasting. It is important that maintenance and repair costs can be adequately forecast to develop the cost curves from which strategic decisions can be made concerning fleet management and make-up.

Another benefit of having accurate maintenance and repair forecasts is the fact that they can be used to identify "problem" machines. If a machine has had a particularly bad repair history, something should be done to rectify the problem—this is the *process control* side of forecasting.

 Additionally, maintenance and repair forecasts make up a portion of internal rental rates for heavy equipment. These rates are used by project estimators to prepare bids and by project managers on the job. In this capacity, forecasts serve as a communication vehicle. The predictions of the equipment manager/committee are communicated to the rest of the company for use in different functional areas.

2.3.2   Types

The types of forecasting available to today's managers are quite numerous. They are best categorized for the purposes of this discussion as qualitative and quantitative methods.

2.3.2.1 Qualitative Methods

Qualitative methods can most closely be associated with the intuitive approach to replacement economics described in Section 2.2.5 by Douglas. The person making the forecast does so on the basis of judgement and intuition (Makridakis et. al., 1989).   Intuitive approaches have been used for forecasts along the entire continuum of time horizons ranging from the immediate to the long term (Makridakis et. al., 1989).

A more formalized qualitative method that also has a place in the equipment management arena is the *Jury of Executive Opinion* (Wilson et. al., 1994). The jury is composed of all the company's top executives that have a stake in the outcome of the forecast. By combining their specialized knowledge and experience, the jury can (hopefully) derive a better forecast than any one individual of the jury could have. This method is best suited to forecast time horizons of three months to two years (Makridakis et. al., 1989) but can be applied to other time horizons as well. The mechanics of running the jury can vary (Wilson et. al., 1994). Often, the jury physically meets in one place, discusses the issues involved, and makes the forecast. Sometimes, especially when dealing with conflicting personalities, one person individually visits each of the jury members, takes in the information, and makes a decision. Some construction companies employ the Jury of Executive Opinion method when setting their internal rental rates—the jury goes by other names, but the concept remains the same.

Qualitative methods do not require an in-depth understanding of mathematical methods on the parts of the participants. Individuals and firms that use qualitative methods tend to like them. Eighty-two percent of the firms familiar with forecasting techniques use the Jury of Executive Opinion (Makridakis et. al., 1989).   Qualitative techniques are most valuable when there is a lack of hard data that can be used for quantitative techniques or when the time horizon of the forecast is far into the future (Kim, 1989).

There are disadvantages to these techniques also. Chase summarized these disadvantages: "(1) they are almost always biased; (2) they are not consistently accurate over time; (3) it takes years of experience for someone to learn how to convert intuitive judgement into good forecasts" (Chase, 1991).   Makridakis (1989) stated that people are generally overoptimistic in the

preparation of subjective forecasts. He also pointed out that it is generally more expensive to employ qualitative techniques than quantitative methods. This is primarily due to the amount of time that executives have to put into making forecasting decisions.

These disadvantages aside, there is always some amount of subjectivity involved when making a forecast. As will be seen in Chapter 6, determinations must be made as to which statistical model is the best for a given situation. These determinations are qualitative decisions on the part of the researcher and, when implemented by industry, will be qualitative decisions on the part of the maintenance manager.

2.3.2.2 Quantitative Methods

Quantitative methods are better suited to the prediction of maintenance and repair costs (Kim, 1989). In general, there are volumes of data available on these costs and used properly these data should be able to provide a reasonably accurate forecast. Quantitative methods that could be applied to equipment management include naïve, moving average, exponential smoothing, time-series analysis, and regression (Makridakis et. al., 1989).

"Naïve" when used in reference to numerical forecasting techniques refers to the simplicity of the forecast—not the abilities of the forecaster. The approaches can be quite simplistic (Hanke et. al., 1995). The quickest of naïve forecasts merely assumes that the future value will be equal to the present actual value. Other naïve forecasts include using the trend for the last two actual values to predict the future value or multiplying the current actual value by a subjective growth factor (e.g. the future value will be 1.05 times the current value). Naïve forecasts are best suited to short term forecasting horizons (less three months).

Moving averages are also well suited to short-term forecasts. They are slightly more quantitative than the naïve methods. The benefit of using a moving average over a naïve method is the partial elimination of errors induced by randomness (Makridakis et. al., 1989). Random events can cause one value to unusually higher or lower than it would normally be. These random events could throw off Naïve forecasts. Moving averages attempt to mitigate randomness by basing the forecast on an average of values for a specified period of time. The number of observations

included in the moving average is *n*. The *n* most current observations are averaged to produce the number that the forecast will be based upon.

$$\text{Moving Average} = \frac{\sum\limits_{t-n}^{t} x}{n}$$ **Equation 2-14**

Where:

       t = the present time

       n = the number of observations in the average

       x = the value of the forecasting parameter

As with the naïve methods, the moving average can be applied as-is or multiplied by a qualitative growth factor. Moving averages play a significant role in the software product *Fleet Information System*® (FIS) by M. Vorster and M. Kapoor. Equipment managers are supplied with moving averages for repair costs and downtime, among other things. In addition to reducing the impact of randomness, the moving averages in FIS help to reduce the seasonal and cyclical nature of the construction industry.

Exponential smoothing is yet another tool available for short-term forecasting. It is similar in concept to the moving average except that the more recent observations are given greater weight in determining the forecast. A weighting factor, α, is chosen such that $0 < \alpha < 1$. As α approaches 1, greater weight is assigned to the most current observations. The equation for simple exponential smoothing is (Hanke et.al., 1995):

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t$$ **Equation 2-15**

Where:

      $\hat{Y}_{t+1}$ = the smoothed forecast
      α = the smoothing parameter
      Y_t = the current actual value
      $\hat{Y}_t$ = the forecast for the current value

There are other exponential smoothing techniques available that are more complex but better suited for the analysis of data with trends or seasonality. The basic concept remains the same.

Time-Series Analysis comprises a variety of techniques whereby patterns in streams of data are identified as they relate to the passage of time. Once the patterns have been identified, they are applied through the forecasting horizon to come up with a forecast future value. Time-series techniques are especially good at characterizing trended, seasonal, and cyclical data streams. The most popular time series techniques in use today are time-series decomposition and the auto-regressive integrated moving average (ARIMA, or Box-Jenkins) techniques. Typically, time-series methods are best suited to short-term forecasts (Makridakis, 1989).

Time series decomposition consists of attempting to identify the separate components that make up a stream of data. A time series decomposition equation takes the form (Wilson et. al., 1994):

$$\mathbf{Y} = \mathbf{T} \times \mathbf{S} \times \mathbf{C} \times \mathbf{I}$$ **Equation 2-16**

Where:

> Y = forecast variable
>
> T = long-term trend in the data
>
> S = seasonal trend in the data
>
> C = cyclical trend in the data
>
> I = random variations

The Box-Jenkins methodology is an iterative process in which the data are compared to a series of models to determine which model provides the best fit. The model that provides the best fit is the one that is chosen to complete the forecast. The underlying assumption is that future values of the forecast variable are related to the past values of the forecast variable. There is no causative relationship between the forecast variable and time. In reality, most data are affected by time somewhat so the data must be transformed so that it does not show a time specific trend. A good discussion of the Box-Jenkins methodology can be found in *Time Series Analysis: Forecasting and Control* (Box et. al., 1994).

Regression techniques are appropriate for up to medium range forecast horizons (up to two years)(Makridakis et. al., 1989). In nearly all texts, a distinction is made between regression models based on trends and regression models based on cause. Regression of a trend over time is the formulation of an equation that expresses the forecast variable as a function of time. Regression of a cause is the formulation of an equation that expresses the forecast variable as a function of one more things that cause the forecast variable to fluctuate. The mechanics of regression are the same for both cases and will be discussed at length in Chapter 4 of this dissertation. It is important to note that in the study undertaken for this dissertation, *time* is the *cause* of the fluctuations in the response variable.

Advances in computer software and hardware have put easy-to-understand regression tools at the fingertips of today's managers. Basic regression can be accomplished within the confines of many spreadsheet programs.   Regression is hungrier for data than the other quantitative methods mentioned, but maintenance managers typically have an abundance of data.

Although quantitative methods of forecasting have been shown to be consistently more accurate than those of qualitative methods (Makridakis et. al., 1989), they do have their shortcomings. One of the biggest of these is the fact that they depend on past events to predict the future. The extrapolative capability of any of the quantitative methods for long-rang forecasting is questionable. All quantitative methods require a data source. Some require a good deal more data than others do. All quantitative methods also require the use of some analytical capabilities both by the forecaster and by the forecast user.

This being said, there is more than enough equipment data available to support a quantitative solution. The fact that the forecast desired is one appropriate for a medium-range time horizon eliminates all methods besides ARIMA and regression. The difference between regression and time-series analysis regarding our problem is almost philosophical in nature. Regression seeks to define the behavior of a response variable in relation to some causative event. ARIMA seeks to define that behavior as a function of simply the passage of time. What is being sought in this research is a linkage between the age of a machine (in cumulative hours of use) and its cumulative repair cost. This is a causative relationship. Although hours of use are definitely a measure of the

passage of time, our interest in them is more as a measure of the amount of work performed. Work *causes* a machine to have more expensive or more frequent breakdowns; time is simply the best way we have of measuring this cause. Regression will be the methodology of choice. Chapter 5 will cover the details of how regression will be used.

2.4   MAINTENANCE AND REPAIR COST FORECASTING

There are many methods in use today to forecast equipment repair costs. Most of them are empirical, not data driven. This topic has been revisited sporadically over the past fifty years by many of the leading authorities in the construction field. What follows in this section is a recounting of these methods along with their strong and weak points.

2.4.1   Straight-line Methods

Most of the equipment maintenance and repair cost estimation techniques described in literature use a constant repair cost over the life of the machine. If a plot of cumulative repair cost vs. cumulative hours of use is formed for any of these methods, the plot reveals a straight line. The slope of the line is the hourly repair cost. This is depicted in Figure 2-4.

Two different methods proposed by Nichols utilize constant hourly repair costs over the life of the machine (Nichols, 1976). In the first of these, repair costs per hour are estimated as a percentage of the straight-line depreciation of the machine. This is particularly useful when a machine is actually depreciated by the straight-line method because ownership cost calculations are simplified. Straight-line depreciation is rarely used in practice today.

A second constant repair cost method described by Nichols estimates hourly repair cost as a percentage of purchase price. This percentage is determined by using numbers published by the Associated General Contractors of America. Peurifoy et. al., recommend a similar method with the percentage determined by company-specific historical records (Peurifoy, et. al., 1995).
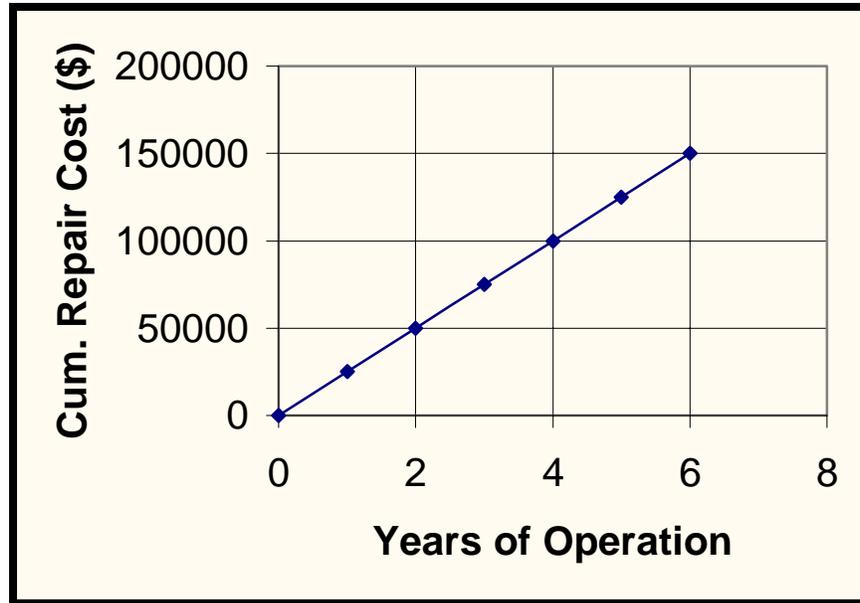
Figure 2-4: Repair Cost as a Percentage of Depreciation

In the *Handbook of Heavy Construction*, E. A. Cox also recommends estimating equipment repair costs as a percentage of purchase price (Cox, 1971). Cox modifies this approach slightly by including multiplication factors for type of service (easy, medium, or severe). The Caterpillar earthmoving equipment manufacturing company recommends an approach nearly identical to that proposed by Cox (Caterpillar, 1995). Caterpillar adds an additional factor for machines that will be used for more than 10,000 hours, but this factor is applied over the entire lifespan of the machine. Terex (Terex, 1981) and Fiatallis' (Fiatallis, 1981) approaches are only slightly different from Caterpillar's. The U. S. Army Corps of Engineers modifies this approach by adding factors to account for regional price variations and inflation (EP 1110-1-8, 1995).

The value of the constant repair cost models is their simplicity. They are very straightforward and easy to apply. They may, however, be an oversimplification of a process that should be represented by a more complex model. Additionally, changes in equipment design, manufacturing processes, and quality may have had an impact on these models.

2.4.2   Terborgh

One of the pioneers of modern equipment management, George Terborgh, recognized nearly fifty years ago that the relationship between repair costs and accumulated hours of use was non-linear (Terborgh, 1949). Data that Terborgh analyzed showed a trend for repair costs per unit of output that increased more rapidly during the early part of a machine's life. Costs tended to reach a static value after many hours or years of service. Because the curvature of the repair rate curves seemed so slight, Terborgh suggested the curve could be replaced with a straight line.

It should be noted that Terborgh studied many different classes of equipment, from textile machinery to farm implements. Unfortunately, he did not study heavy construction equipment— the closest category was farm implements. These implements ranged in age from new to twenty years old. The data on this equipment were collected in 1936. Some of the machines studied were built as early as 1916. The accumulated hours on the machines at the twenty year point were extremely low when compared to what heavy construction equipment would accumulate in that time frame.

The average hours of cumulative use for the twenty-year-old farm machinery was 1,000—some pieces of construction equipment accumulate more than twice that many hours in just one year. Some of the farm machinery studied had been in service for twenty or more years. Construction equipment is seldom kept more than ten years. Since Terborgh was plotting repair cost per unit of output and this research is more concerned with cumulative repair costs, a transition needs to be made. The straight line that Terborgh mentioned in reference to growth of repair costs would translate into a quadratic line if it were a plot of cumulative repair costs. His slightly curved line would translate in some function higher than a quadratic (possibly a cubic or exponential).

As an exercise, a hypothetical cumulative repair cost curve was constructed using numbers obtained from Terborgh's data on farm implements. To aid in the construction of the curve, it was assumed that the acreage output per hour remained constant at one acre per hour over the life of the implement. This curve obtained is depicted in Figure 2-5. It can be seen that the curve shows a definite upward trend and can also be seen that there is indeed a slight curvature to the plot. A simple multiple regression was performed on the data to yield a quadratic equation.

Although the "x" term seems to be more significant than the "x^2" term, the fact that the "x^2" term is present lends credence to the idea that there is curvature to cumulative cost curves and that some type of optimization should be possible.
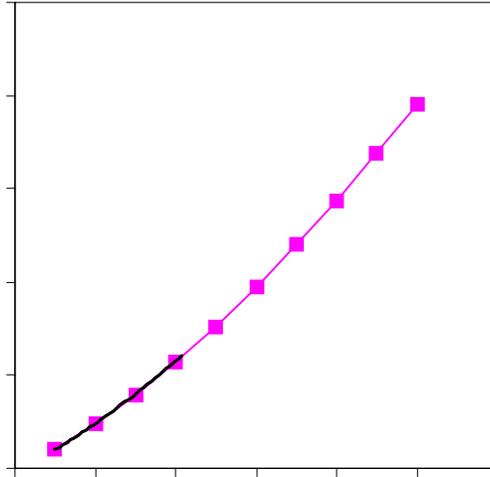
Figure 2-5: Cumulative Repair Cost

Since heavy construction equipment works more hours over a much shorter time than the farm implements of Terborgh's study, Terborgh's graphs may not accurately reflect how construction equipment behaves. Another problem with Terborgh's study is that it directly compared new machines with machines that were twenty years old. Many advances in technology were implemented between 1916 and 1936 (most notably the assembly line). There are also inflation considerations that were not addressed. It is promising, however, that non-linear trends were identified as part of the way equipment ages over 60 years ago.

2.4.3   Nichols

Herbert Nichols proposed a detailed method of estimating repair costs in his book, *Moving the Earth* (Nichols, 1976). An hourly repair cost is obtained by multiplying factors for type of equipment, total hours of use, years of useful life, temperature, work conditions, maintenance quality, type of use, operator style, luck, equipment quality, and pace of work. These factors are multiplied together and then multiplied by 1/10,000th of the purchase price of the machine to

obtain an hourly cost. Nichols' repair cost multipliers increase almost linearly as a function of cumulative hours of use (Figure 2-6). These factors are designed to be used by all types of construction equipment. They are not tailored for any particular category or group of equipment. But, they are scaled when the type of equipment multiplier is applied. This essentially increases or decreases the slope of the line shown in Figure 2-6.
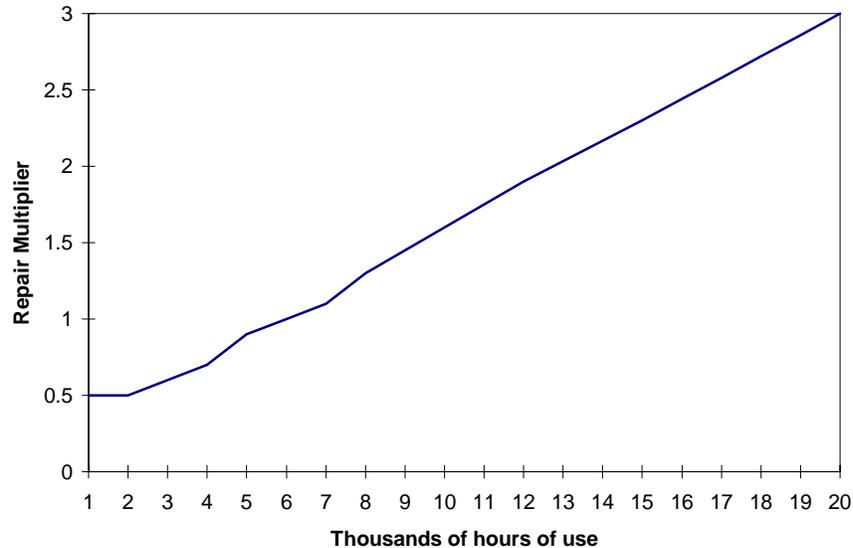
Figure 2-6: Repair Multiplier vs. Cumulative Use (Nichols, 1976)

As a precursor to Nichols' discussion of his repair factors, Nichols recommended that company-specific data be used as a primary means of estimating repair costs—but he failed to explain how. Of all the methods of predicting equipment repair and maintenance costs documented in literature, Nichols' repair factor method stands alone as one that attempts to account for the increasing rate of repair costs with increasing accumulated hours of use.

2.4.4   Nunnally

S. W. Nunnally proposed a method of estimating repair costs that is similar to the sum-of-the-years digits method of depreciation accounting (Nunnally, 1993). In sum-of-the-years'-digits depreciation accounting, the depreciation in a given year is calculated by the following formula:

$$D = \frac{(N - m + 1)(P - S)}{\left[\dfrac{N(N+1)}{2} \right]}$$

Equation 2-17

Where:

D = depreciation

N = total number of years in the economic life of the machine

m = the current year in the life of the machine

P = the initial purchase price of the machine

S = the salvage value

When using the sum-of-the-years'-digits method of depreciation, the amount of depreciation claimed is large at first and tapers off as the machine gets older. This accounting procedure is more beneficial to companies than constant depreciation when figuring taxes (it is, however, no longer an acceptable method of tax accounting in the United States).

The method put forth by Nunnally gives a function that is essentially the inverse of the depreciation function described above. It is given by the equation:

$$C = \frac{m}{\left[\dfrac{N(N+1)}{2} \right]} * \frac{\text{Lifetime Repair Cost}}{\text{Hours Operated}}$$

Equation 2-18

Where:

C = the hourly repair cost

With this formula, it can be seen that the cost will increase with an increase in m, the current year, instead of decrease. The lifetime repair costs are expressed as a percentage of original purchase price. This percentage is based on operating conditions (favorable, average, or severe). These factors range from 40% for a dragline operating in favorable conditions to 105% for a scraper operating in severe conditions. Nunnally provides a table that gives these percentages for each category of equipment. The hours operated in the equation are the total hours that a company

expects to operate a given machine during its life. This should not be construed to mean that Nunnally's theory accounts for accumulated hours of use—it does not. This theory assumes that the total accumulated repair costs at the end of a machine's life will be the same, regardless of whether it has worked 1,000 hours or 10,000 hours. The repair cost multiplier (the first fraction in the formula given above) increases incrementally each year of a machine's life (Figure 2-7). This figure was made for a hypothetical machine that had an expected economic life of six years.
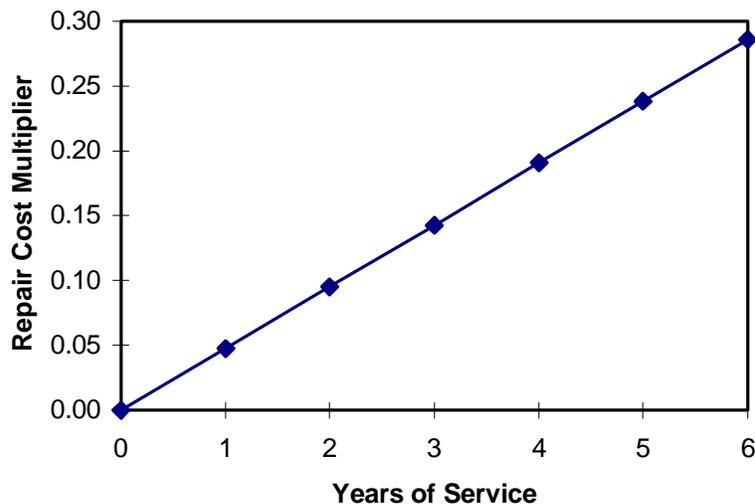
Figure 2-7: Repair Cost Multiplier vs. Years of Service (Nunnally, 1993)

Assuming the above machine cost $250,000, the cumulative repair costs of the machine can be charted (Figure 2-8). The function charted appears to be quadratic, but it fails to take into account the accumulated hours of use on the machine. The accumulated repair costs at the end of each year would be same, regardless of how many hours the machine had worked. A dilemma results when applying Nunnally's method across years in which production varies. Because of the way equation 2-18 works, a machine could conceivably have a *lower* hourly repair cost as it ages. All it would take would be an exceptionally low usage rate in an early year followed by an exceptionally high usage rate. This scenario is realistic in the world of heavy construction. If a company fails to win jobs in one year and succeeds in bidding many jobs the next, they could be in a position where the machines are relatively idle one year and working double shifts the next.

Common sense dictates that forecasting models are not followed blindly, but the point is made that there are many assumptions made in empirical models.
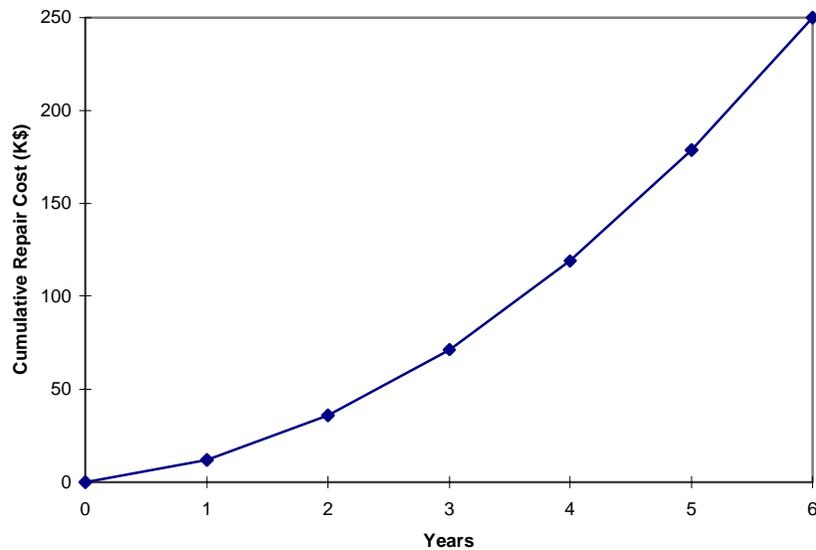
Figure 2-8: Cumulative Repair Cost vs. Years (Nunnally, 1993)

Nunnally made an attempt to account for the increasing nature of repair costs as machines age. Nunnally, however, defined machine age in terms of the passage of time on a calendar. Calendar time may not be the best way to characterize the growth of repair costs for earthmoving equipment.

2.4.5   Kim

In his dissertation, Yong Hwan Kim developed a statistical method of estimating repair costs based on the combined failure distributions of major components (1989).   The study was conducted on a large fleet of U.S. Army trucks. The trucks ranged in age from 8 to 24 years old. Twenty critical components on the trucks were selected and distributions for their failure times were developed. By combining failure curves for the critical 20 components, Kim developed a time-series model of the average repair costs for these trucks.   This model was statistically consistent with the actual repair costs incurred by the army during those times.   Kim found that for equipment with very long life spans, repair costs increase monotonically up to a point, after

which they decrease slightly and then level out. In the late years of an Army truck's life, it will have a relatively constant repair rate per year. Kim concluded that his methods were applicable to all long-life machinery.

The direct applicability of this study to the construction industry is questionable. Most Army mechanized equipment wears out due to physical aging, not due to use. Interviews with the U.S. Army's TACOM (Tank Command) confirm this (Mitchell, 1997). The trucks in Kim's study accumulated an average of less than 2,000 miles of travel per year. Assuming an average speed of 40 M.P.H., these trucks were in operation less than 50 hours per year—this is not even close to the number of hours construction equipment is used (typically at least 1,000 hours per year). Additionally, Kim's trucks are subject to relatively static load conditions and are driven mostly on smooth surfaced roads. Most units of construction equipment are subject to dynamically changing load conditions and are operated on rough surfaces. Another problem with the applicability of Kim's research to the construction industry was the assumption of a steady state of component replacement with an infinite life of the frame and other systems supporting those components. This does not hold true for construction equipment.

2.4.6   Observations

Numerous methods of estimating equipment repair costs are described in literature. The cover a broad spectrum, ranging from over-simplistic empirical formulas to a difficult to employ time series method based on individual component failure. No data-driven approach was described that could easily be applied to data already collected by construction firms.

2.5   SUMMARY

This purpose of this chapter was to present a literature review that will set the stage for a fuller understanding of the chapters that lie ahead. First, literature on economic replacement models was reviewed to provide a broad context for the need for cumulative repair cost equations. Cost minimization, profit maximization, and repair limit theories were discussed. Following the discussion of economic theories in general, important contributions and enhancements that have

been published over the years were discussed. This provides the background needed to understand the Cumulative Cost Model which will be discussed in Chapter 3.

After discussing economic theories, forecasting methodologies were described. First the general tenets and methods of forecasting were put forth, then some methods intended specifically for machinery were discussed. This provides the background needed to understand the development of the test methodology (Chapter 5). It will also give the background to understand comparisons made to other forecasting methods which will be accomplished in Chapter 8.

CHAPTER 3: THE CUMULATIVE COST MODEL

Chapter 2, the Literature Review, provided the background needed to understand the theories involved with the Cumulative Cost Model (CCM). This chapter will discuss this model and its uses in detail. The CCM is the model best suited for economic decision making within the equipment management environment.

3.1   THE BASIC MODEL

An explanation of why the CCM was chosen as the primary economic model is in order. The three economic replacement models discussed in Chapter 2 allow for some type of numeric solution to economic replacement problems. The numeric solution is very important because the issue at hand is the development of an economic replacement policy that will provide the greatest financial benefit to the companies involved.

The economic replacement models discussed can also be graphically depicted. Many authors fail to communicate the importance of a graphical solution. Some neglect to discuss it altogether. Graphical solutions enable the decision-maker to better conceptualize the problem at hand. Costs (and revenues if applicable) are depicted as curves on a two-dimensional chart. Drawing tangents to the applicable lines depicts the optimization functions.   By having clearer problem definitions, equipment managers can understand exactly what the optimization process does for them.

The cumulative cost model provides a valid numerical solution and an intuitive graphical depiction of the problem being analyzed. It also provides things that the other models do not (Vorster, 1980). With the cumulative cost model, it is possible to depict and understand changes in total costs, average costs, and marginal costs. The cumulative cost model is the only one of the economic replacement models that incorporates both classic economic replacement theory and

repair limit theory. The cumulative cost model can be used to minimize costs or to maximize profits—it is not implicitly tied to one method or the other. It is also possible to explicitly show the three basic steps of buy, operate, and sell at any point in the machine's life. The cumulative cost model allows for more than one definition of economic life for heavy construction equipment.
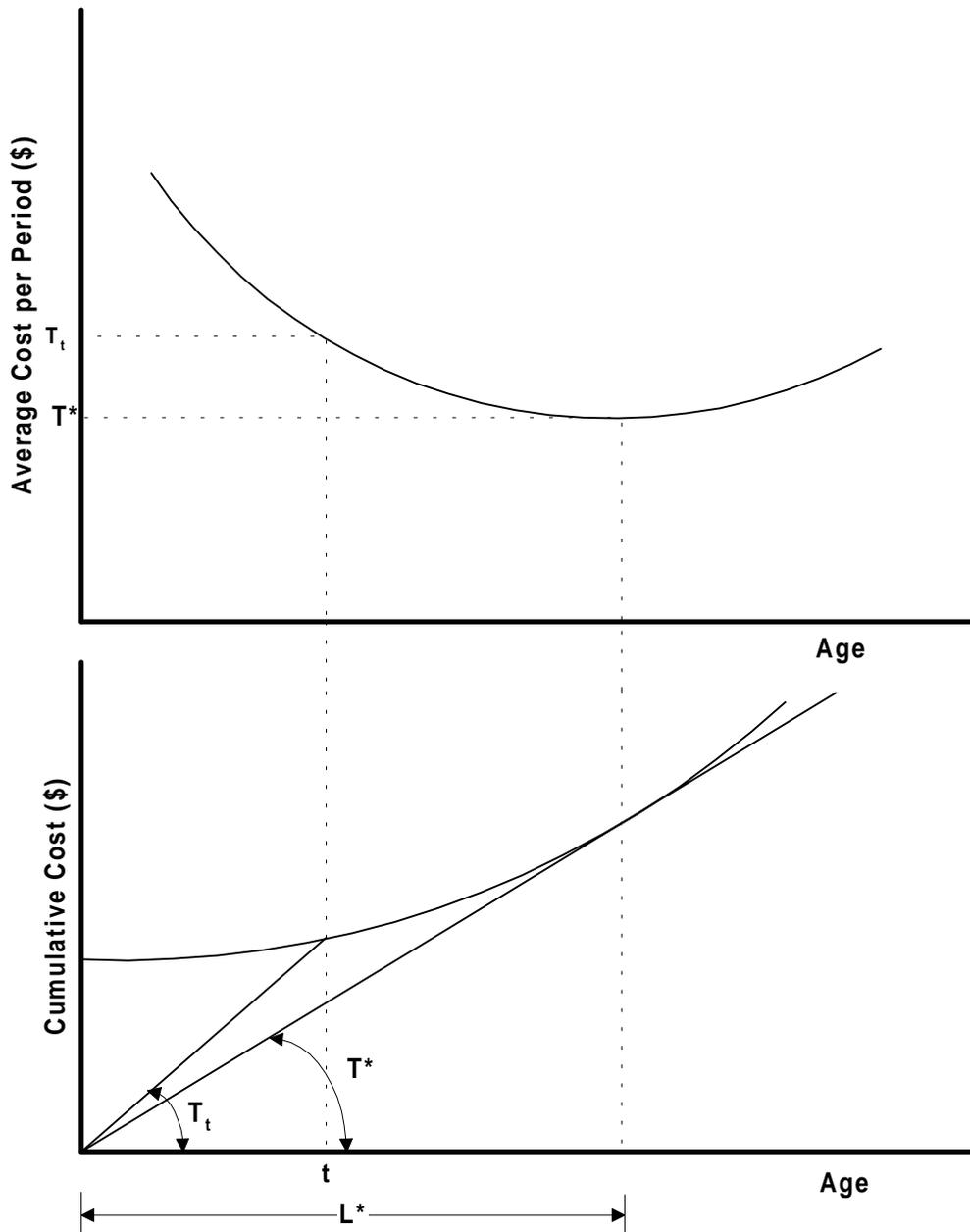
Figure 3-1: The Cumulative Cost Model vs. Cost Minimization (Vorster, 1980)

Figure 3-1 (Vorster, 1980) is a geometric comparison of the cumulative cost model and the cost minimization model. It can be seen that both of the models can be used to show the optimization function. Both optimum points are defined by geometric tangents to the cost curves. The cost minimization method uses a horizontal tangent to the total average cost curve to define T*, the minimum average annual cost and L*, the optimum economic life. The cumulative cost model uses a tangent to the cumulative cost curve drawn that has its intercept fixed at the origin. This tangent defines the same optimum point that the horizontal tangent defines for the cost minimization model. T* and L* have the same meaning as in the cost minimization model, but T* is defined a little differently. Instead of being the vertical coordinate of the optimum point, T* is the slope of the tangent line drawn to the optimum point.

The average operating cost, T_t, for a given time, t, can be found graphically for each of the two models by drawing lines. For the cost minimization model, T_t is found by drawing a vertical line from the ordinate at the time of interest that bisects the average total cost curve. A horizontal line is then drawn from the point of bisection to the abscissa. The point where this horizontal line joins the abscissa is T_t. For the cumulative cost model, a straight line is drawn directly from the origin to the point where the vertical line corresponding to the time of interest bisects the cumulative cost curve. The slope of this line is T_t

The abscissa of the CCM is age. Units are not specified at this point to highlight the flexibility of the model. Age can take the form of calendar age, age in cumulative hours of use, or age in units of production. The definitions of these various ages will be presented in Chapter 4.

The ordinate of the cumulative cost model is *cumulative cost*, normally expressed as either the sum of or net present value of all transactions to date. All owning and operating costs *can* be depicted in the CCM. Costs will be discussed in more detail in Chapter 4.
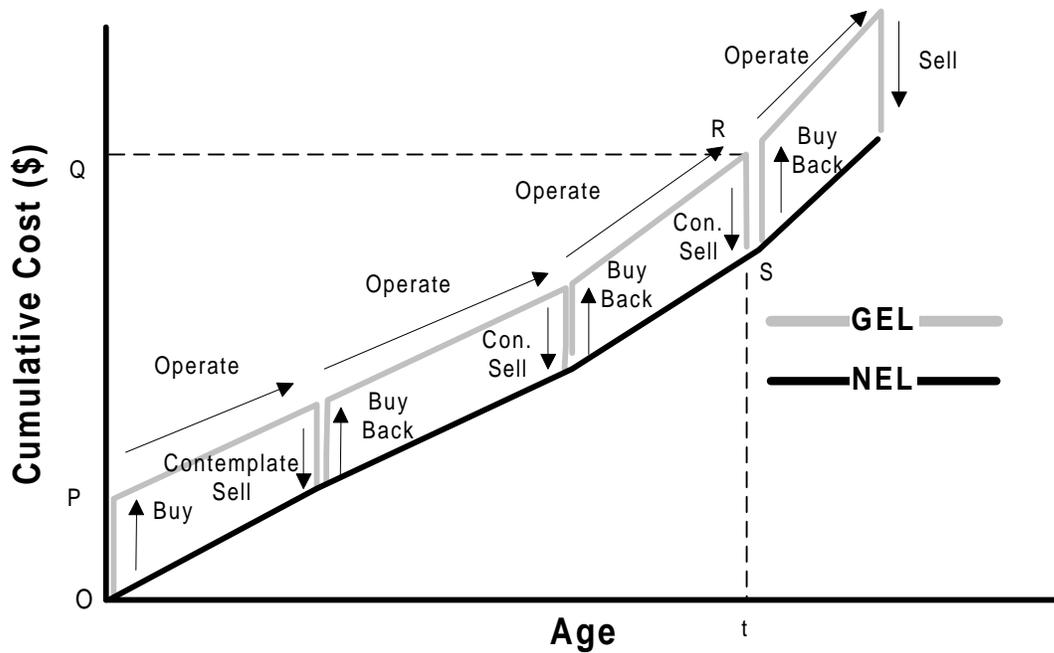
Figure 3-2: The Cumulative Cost Model--Detail

Figure 3-2 shows a simplified version of the cumulative cost model in detail for four periods. Straight lines are used in place of curves for the purposes of illustration only. It can be seen that the entire life cycle of the machine is depicted on this graphic. The four periods shown indicate four times the sell decision was contemplated. The fourth time, the machine was sold. Line OPRS, which is the Gross Expenditure Line (GEL), goes sharply upward at time zero to reflect the initial purchase of the machine. It then rises slowly as costs are incurred over the life of the machine. It finally drops abruptly when the machine is sold or when a sale is contemplated. Line OS is the Net Expenditure Line (NEL). The following definitions apply to some of the other line segments:

OP = original capital cost (P_o)

PQ = expenses for the period (ΣE_p)

RS = Salvage value at time t (S_t)

St = Net expense for the period $\{P_0 + \Sigma E_p - S_t\}$

OS = Uniform recovery line (URL)

3.2   THE CCM IN DEPTH

The CCM can be discussed in greater depth now that the basics are known. Figure 3-3 fills in the details lacking in Figure 3-2 and provides the basis for the following definitions:

P, R_1, R_2, ... R_t = Gross Expenditure Line (GEL)

O, S_1, S_2, ... S_t = Net Expenditure Line (NEL)

OS_1, OS_2, ... OS_t = Uniform recovery lines (URLs)

Tan tOS_t = URL gradient at time t, or the average cost to time t (T_t)

T* = minimum value for URL gradient, or the optimum average cost

L* = optimum economic life
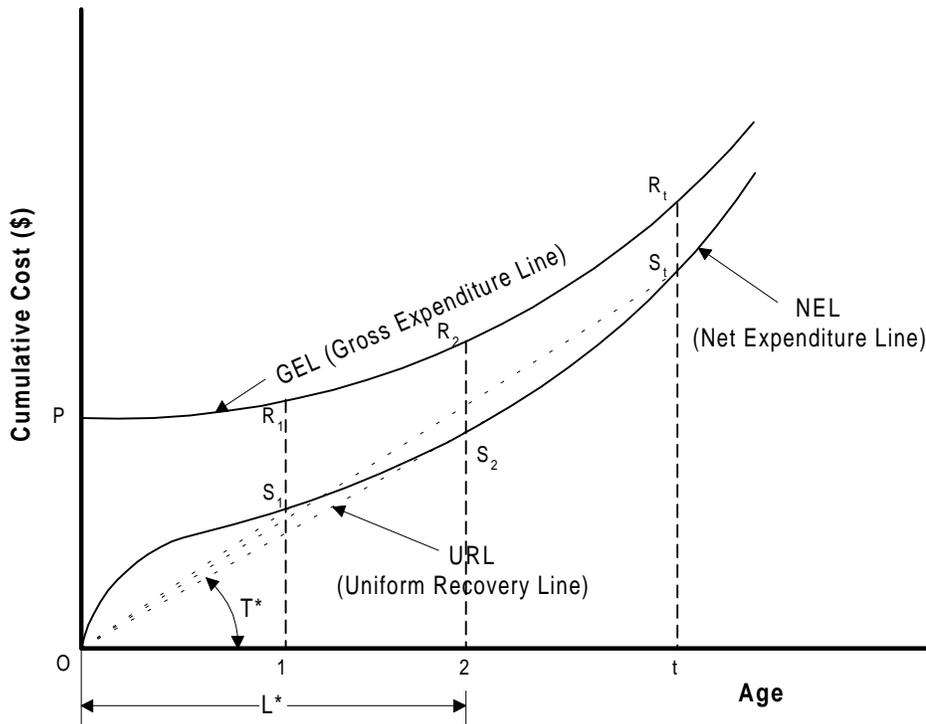
Figure 3-3: The Cumulative Cost Model (Vorster, 1980)

The NEL is equal to the GEL minus the salvage value of the machine at time t. Salvage value is sometimes referred to as residual value—it is the amount of money that the machine could be sold for at a particular point in time. The reason for the "hump" in the NEL is the rapid decline in salvage value early in the life of an asset. As the residual value decreases, the NEL converges with the GEL. According to Drinkwater and Hastings (1967), the residual value of a machine at
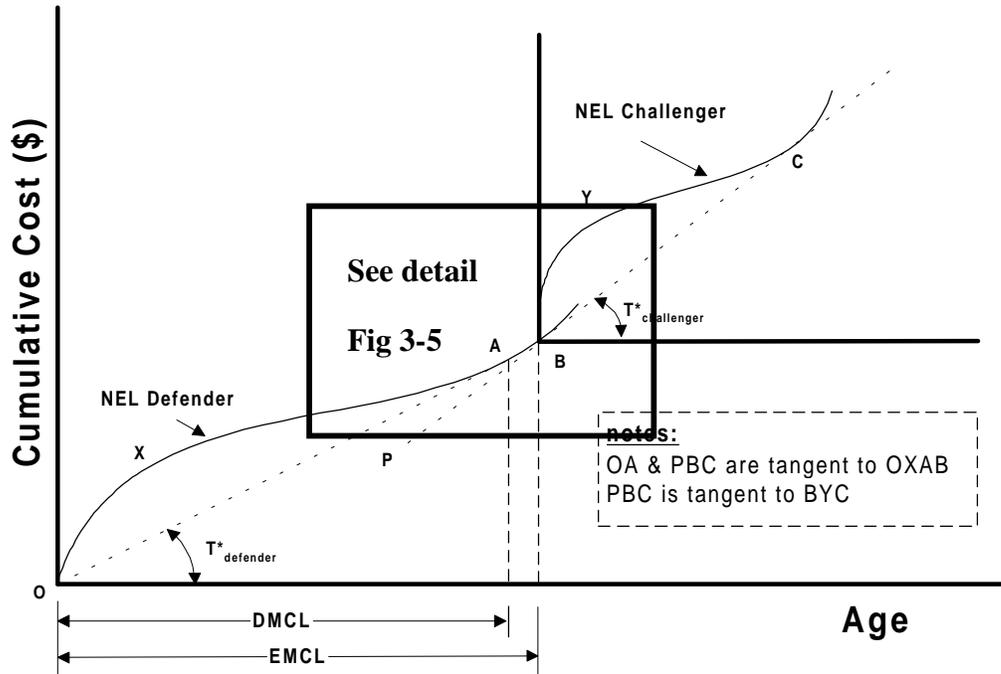
Figure 3-4: Definitions for Economic Life

any given time should approximate the repair limit at that time. It can also be seen that the minimum T_t is reached when the uniform recovery line (URL) is tangent to the NEL. This is the gradient T^* that was discussed earlier. The cumulative age defined by the bisection of the NEL by the URL is L^*.

There are two definitions of economic life that are of importance in the cumulative cost model:

1. The Defender's minimum cost life, DMCL: defined as that period which ends when the average annual cost of a Defender reaches a minimum. This is equivalent to L^* for the Defender.

2. The equal marginal cost life, EMCL: defined as that period which ends when the marginal cost of keeping the Defender one more period systematically exceeds the minimum average annual cost which can be expected from an equivalent Challenger.
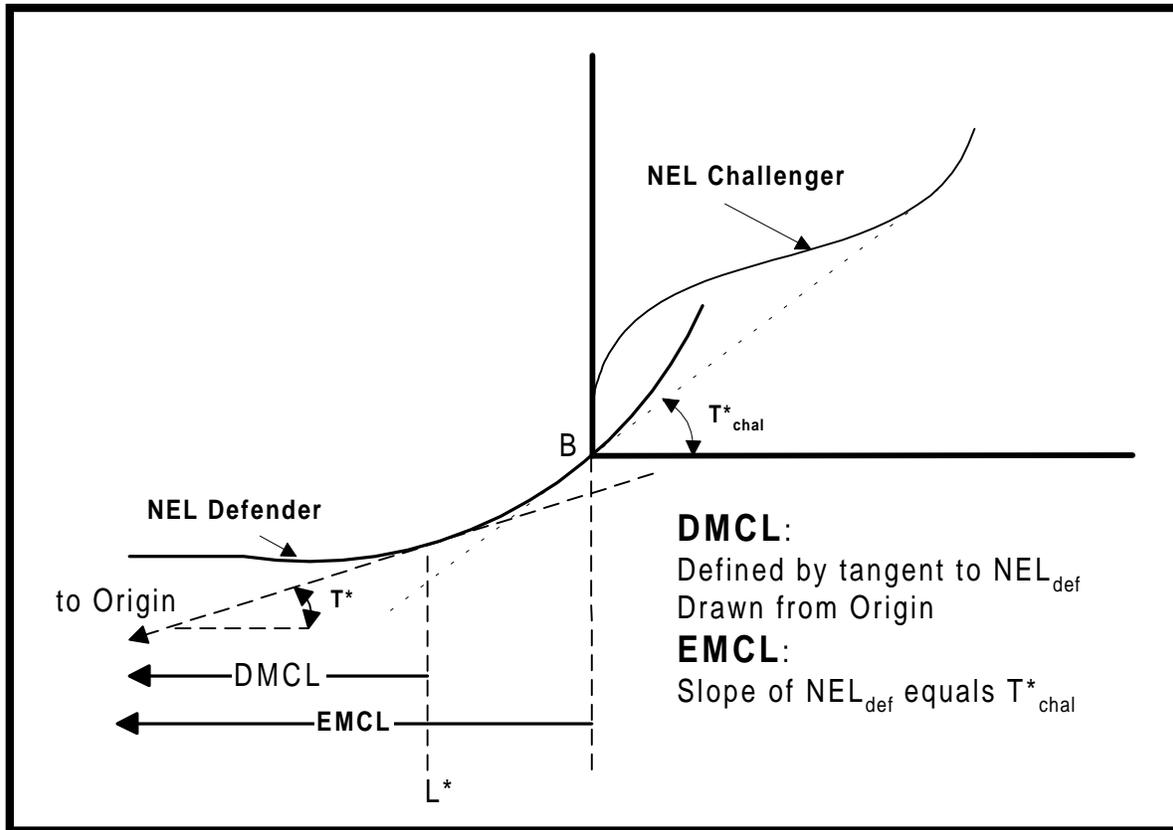
Figure 3-5: Detail of Figure 3-4

These definitions are depicted graphically in Figure 3-4. This figure also shows how successive machines can be depicted on one figure. This helps the equipment manager better visualize exactly what happens to a given asset as time progresses. The Challenger is depicted on the figure with its own age and cumulative cost axes. These axes must be of the same scale as those of the Defender. The graph depicting the Challenger is then laid on top of the Defender's graph with the point of interest on the NEL of the Defender serving as the locus for the origin of the Challenger's graph. The Challenger's graph can be depicted at any point along the Defender's NEL. The point corresponding to the EMCL was chosen for Figure 3-4 to aid in the visualization of economic life.

3.3   USING THE CCM

Using the CCM is not difficult. Geometric or conceptual solutions are easy and intuitive. Numerical solutions are more involved and rely on a knowledge of the equations that define the

NEL and GEL. Optimizing the equations is a matter of simple calculus as will be demonstrated in Chapters 8 and 9. However, the quality and validity of the results obtained from the CCM cannot exceed the accuracy of the algebraic expressions used to describe the NEL and GEL.

The first curve that must be defined is the Gross Expenditure Line. The GEL should reflect all components of owning and operating cost if perfect accuracy is desired. These components were discussed briefly in Chapter 1 and will be discussed in more detail in Chapter 4. With a few notable exceptions, it should not be too difficult to construct the GEL up to the present is a company keeps good cost accounting records. The matter is, however, complicated by the consequential costs of obsolescence and deterioration which are well accepted but difficult to quantify (Vorster and de la Garza, 1990). Using the GEL will be described in Chapter 9.

Defining the NEL is extremely difficult. Although it is simply GEL less the salvage value, salvage value is dependent upon many factors. These include, but are not limited to: the hours on the machine, the calendar age of the machine, the timing of the machine's major rebuilds, the machine's exterior appearance, the region of the country in which the sale is to be made, the time of year in which the sale is made, and the market conditions which affect the demand for the particular type of machine being sold.

However, as cumulative hours increase the residual values decrease and the NEL converges with the GEL. After a machine reaches a certain age, its residual value is not that great and is not influenced as much by the factors listed above. A method for approximating the NEL will be discussed in Chapter 9.

3.4   DECISIONS SUPPORTED BY THE CCM

One of the main reasons the CCM is so attractive lies in the scope of decisions that it supports. Most of the models discussed presented a solution for only one type of decision: like-for-like equipment replacement. An example of this would be the replacement of an aging scraper (Defender) with a new scraper (Challenger). The Challenger typically will have some sort of advantage, be it longer expected life or improved production, which causes it to be more

economical than the Defender.   The goal of the economic replacement model is to find the optimum point in time to replace the Defender with the Challenger.

A survey of equipment managers (Mitchell, 1997) was undertaken to determine which type of economic replacement decisions they make most often.   Although like-for-like replacement had the highest rating (36%), it was clear that a model that supported other types of decisions would be useful.   The scope of decisions supported by the cumulative cost model are listed below:

1. *Purchase*: This is the initial purchase of a piece of equipment for a fleet. The purpose of the purchase is not to replace an existing asset. The purpose is to expand opportunities, increase production capacity, or perform a task that the current fleet cannot perform. Usually, a decision must be made between two or more alternative machines, each with one or more associated methods of finance.

2. *Maintain*: Maintain decisions are those that pertain to the money invested in preventive maintenance (PM) in an effort to minimize repair expenditures or extend the life of a machine. Decisions on the types and timing of PM should be made by the equipment manager for each type of machine owned.

3. *Repair*: Repair decisions are those decisions concerning whether or not to repair a machine that has failed while in service. Repairs do not extend the life of a machine—they merely bring it back to an operational state. Most firms that own equipment delegate repair decisions to the field with some caveats. These caveats usually take the form of a price ceiling above which the decision to repair or not is deferred to the next higher management level.

4. *Rebuild*: Rebuild decisions are distinguished from repair decisions in that a rebuild extends the service life of a machine. The rebuild can be accomplished on the whole machine or just on critical components like the drivetrain. Usually, rebuilds represent a significant investment in the machine.   Capital spent on rebuilds can be partially recovered through depreciation. Rebuild decisions can be made at any time and are not driven by the fact that the machine has failed.

5. *Like-for-like replacement*: This decision was explained above. While it is the capital decision that equipment managers most often face, it is by no means the only type of capital decision that they must make. This is the only capital decision addressed by most authors in the literature.

6. *Production capacity replacement*: Production capacity replacement was the second most frequent capital decision problem among the respondents to our survey. In this type of replacement problem, one general category of equipment is replaced with another general category of equipment to have no net change in production capacity. An example of this is the replacement of scrapers with articulated dump trucks because the articulated dump truck/excavator combination is seen to be more versatile and cost effective than the scraper/push dozer combination. Production capacity replacement problems are usually more subjective than like-for-like replacement problems. Collateral costs can be more heavily involved when making a decision to switch equipment types.

7. *Retire*: Retire decisions are made when it is desirable to remove a machine from service. The proceeds from the retirement sale can either be removed from the equipment division completely or reinvested in unrelated equipment types. Old equipment is sold, the money is made available for new purchases, and the equipment manager is once again faced with a *purchase* decision.

Although it may be possible to modify the usage of other models to accommodate *some* of the decisions listed above besides like-to-like replacement, no single model except for the CCM can support all of the above decision types.

The remainder of this section will discuss exactly how the CCM can be used to support all of the above decisions. Additionally, the mechanisms for using the CCM to determine impact on profit will be covered.

3.4.1   Purchase

The decision to purchase a new piece of equipment that is not a replacement of something already in the company's inventory is a strategic decision that should be made by the company's top

management. The company should have well-developed strategic goals and select its core fleet composition based on those goals. After the decision to purchase has been made, the cumulative cost model can aid in the identification of the best suitable candidate for the job. Usually, there will be two or more machines that must be evaluated. Each of those machines will have a number of financing options (e.g. purchase, lease, lease/purchase, etc.) A decision will be made on not only which machine to purchase, but how to finance it as well.

The first step in the process is to obtain reliable information concerning historical resale values, operating costs, and financing options from the manufacturers of the prospective machines and the lending institutions with whom the firm does business. The information on operating costs and financing options will be used to estimate the GEL for each alternative using cumulative hours of use as the abscissa. The data on historical resale values will be subtracted from the GEL to derive the NEL. It is important to note that if a lease option is being evaluated the GEL and the NEL will be one in the same. Once the NEL has been derived, a tangent URL from the origin is drawn (

Figure 3-6).

This tangent point defines the ending of the DMCL as defined above. The slope of this URL, T*, is the minimum average cost per hour for owning and operating the machine. Future machines should have the same T*, the replacement should be made when the DMCL is reached. All other things being equal, the alternative with the smallest T* should be the machine and finance method chosen. If there are mitigating circumstances (i.e. one machine has a higher productivity than the others do), appropriate collateral costs can be assessed to the weaker machines.

RULE: *When making a purchase decision, alternatives should be evaluated in terms of their minimum URL gradient. All other things being equal, the machine with the lowest minimum URL gradient (T*) should be selected.*
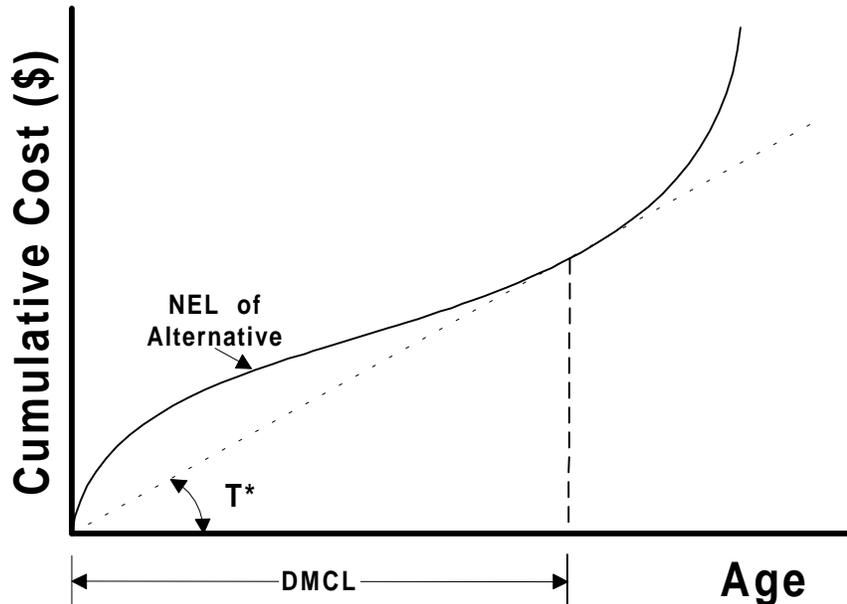
Figure 3-6: The Purchase Decision

3.4.2   Maintain

Almost all heavy equipment companies have some sort of preventive maintenance policy. It is very easy to quantify the direct costs associated with a given policy and their timing. The cost of the maintenance is easy to calculate because the items accomplished are spelled out in the policy. The intervals between maintenance visits are also specified in the policy. A difficult thing to quantify is the results of a given policy. These quantities and their timing can be determined from manufacturer's data or from actual field data obtained by testing various policies.

To compare policies, URL's for *identical* machines under *different* maintenance policies are compared. Graphically and conceptually, there is little difference in the mechanics between this procedure and the one described above relating to machine purchase. The only difference is that instead of comparing different machines, the manager is comparing different maintenance policies. The NELs and URLs for each policy are drawn and T* for each policy alternative is determined.

The policy that provides the lowest T* is the one that should be chosen *for the machine of interest*. This policy *may not* be the best policy for the rest of the fleet.

RULE: *When making a maintain decision, preventive maintenance policy alternatives should be evaluated in terms of their minimum URL gradients. All other things being equal, the policy with the lowest minimum URL gradient (T*) should be selected.*

3.4.3   Repair

The format of the cumulative cost model is similar to that of the repair limit model developed by Drinkwater and Hastings (1967). The basic concept of repair limit theory is that there exists some dollar amount, the repair limit, below which it economically sound to repair the machine. If a machine breaks down and the estimated cost of the repair is greater than the repair limit, the repair is too expensive and the machine should be retired or replaced.

According to Hastings (1969), in a perfect market the optimum repair limit of an item is equal to its resale value. In terms of the cumulative cost model, this is the difference between the GEL and the NEL. This was depicted graphically in Figure 2-3. The line segments R_1S_1, R_2S_2, and R_tS_t represent the repair limit at the respective points in the machine's life. To apply repair limit theory, obtain a good estimate of the cost of repair, r_e. Compare this estimate to the difference between the GEL and the NEL at the appropriate point in the machine's life.

The application of repair limit theory to every minor repair that takes place on a machine would be counterproductive. A valid strategy would be for the equipment manager to periodically evaluate the repair limits of all the machines in the fleet. From this evaluation, monetary limits could be set for field repairs. These limits would be somewhat less than the repair limits. If the cost of the repair is less than the field limit, the field mechanics can perform the repair without approval from the equipment manager. If the cost of the repair is greater than the field limit, the equipment manager should evaluate the repair in terms of the actual repair limit.

RULE: *When making a repair decision, the machine should be replaced or retired if $r_e \geq$ (GEL-NEL). If $r_e <$ (GEL-NEL), the machine should be repaired.*

3.4.4   Capital Rebuild

Rebuilds are undertaken to extend the life of the equipment in question. Although a rebuild may extend the physical life of a system, it will not necessarily make the average operating cost cheaper. There are two general rebuild scenarios: planned rebuild and rebuild due to failure. At any point in time there are three options available to the equipment manager concerning a planned rebuild: the machine should continue to operate in its current condition, the machine should be rebuilt, or the machine should be replaced. The option that yields the lowest URL gradient is the one that should be chosen. If the machine has failed, the option of continuing the operation of the machine is not available.

Deciding which URL gradients to use can be a bit tricky when making a rebuild decision. If the Defender's DMCL has not expired, the URL gradient used for the comparison should be T* for the Defender before the rebuild. Otherwise, the current URL gradient should be used. In the case of a failed Defender, the URL gradient of the Defender is not used.

The URL gradient for the rebuilt Defender is found by drawing a line tangent to the NEL of the rebuilt Defender (Figure 3-7). This gradient will be referred to as $T^*_{rebuild}$. The NEL for the rebuild Defender is a continuation of the NEL for the Defender. The point at which the rebuild is accomplished will show a sudden vertical increase in the NEL to account for the initial cost of the rebuild. The NEL should then resume a flattened and gradually increasing path. It is important to note that $T^*_{rebuild}$ is determined using the origin in relation to the original purchase of the machine, not the displaced origin with respect to the timing of the rebuild. The URL gradient to use for the Challenger is T* for the Challenger. These three URL gradients (or two in the case of a failed machine) are then compared. The option with the lowest URL gradient is the one that should be chosen.

RULE: *To make a positive rebuild decision, two conditions must be satisfied: first, $T^*_{rebuild}$ of the rebuilt Defender must be less than the URL gradient (or T* if DMCL has not expired) of the Defender before the rebuild and second, $T^*_{rebuild}$ for the rebuilt Defender must be less than the T* of the Challenger.*
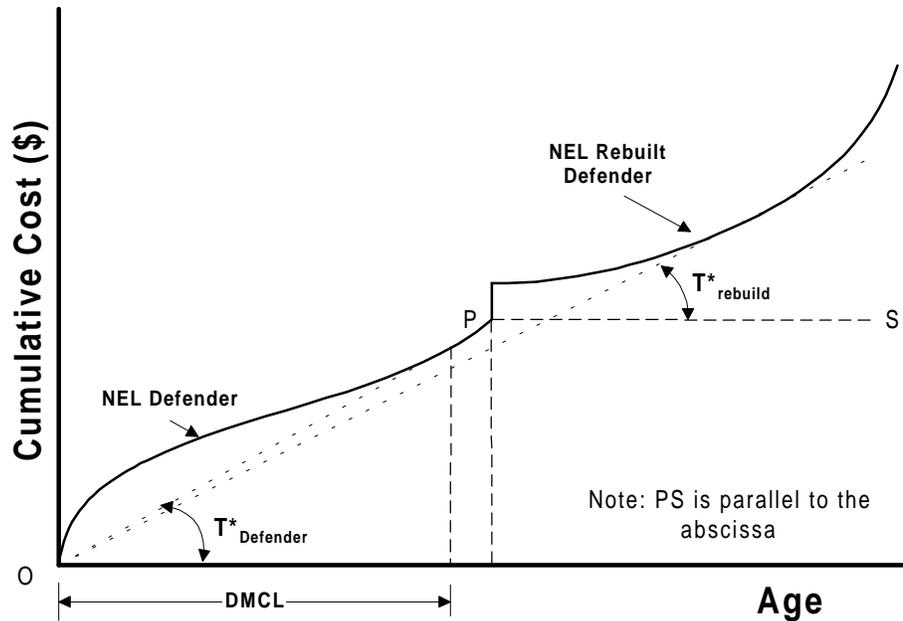
Figure 3-7: The Rebuild Decision

3.4.5   Like for Like Replacement

If a decision has been made to replace the Defender, the best time to do it is at the point when it becomes cheaper to own and operate the Challenger. To continue to operate the Defender beyond that point results in additional costs that would not have been incurred had the replacement action occurred sooner. Unnecessary costs are also incurred if the Defender is replaced too soon.

The model associated with like for like replacement decisions is depicted in Figure 3-8. The slope of the URL represents the average hourly operating cost. Graphically, the Defender is operated until the URL of the Challenger is tangent to the Defender's NEL. Algebraically this occurs when the marginal cost of the Defender is equal to T* of the Challenger. When the EMCL has expired, the Defender should be sold and the Challenger should be purchased.

RULE: *When making a like-for-like replacement decision, the Defender should be replaced when its marginal cost systematically exceeds T* of the Challenger.*
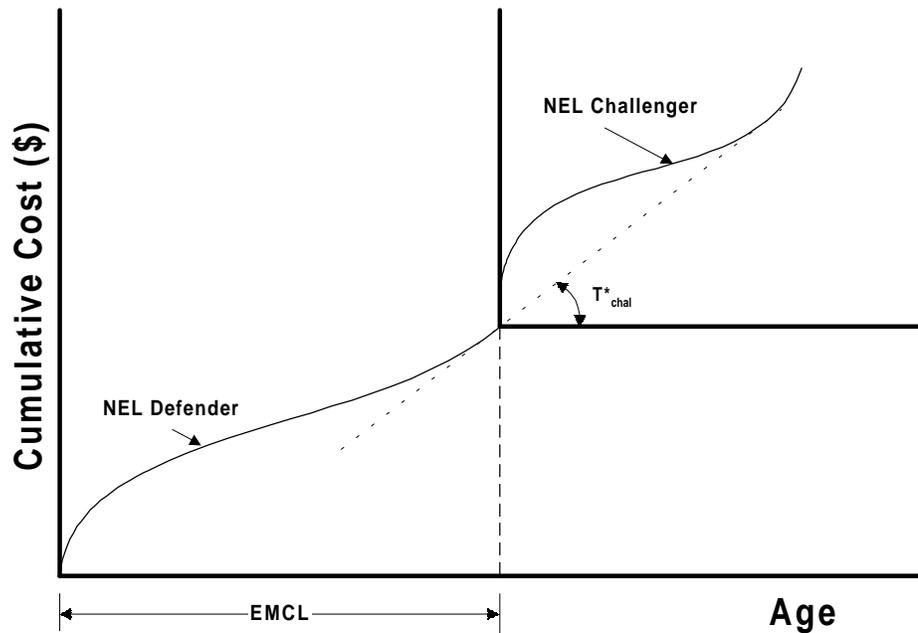
Figure 3-8: Like-for-Like Replacement

3.4.6   Production Capacity Replacement

The production capacity replacement decision is the most complex of the capital decisions that will be discussed.   The analysis becomes more of a method comparison than a one-to-one economic competition.   However, the decision can still be likened to the classic Defender vs. Challenger.   The old method is the Defender; the new is the Challenger.   The machine age that makes the most sense to use in this type of analysis is machine age in units of production.   This makes the slope of the URL equivalent to the average unit production cost.   The graphic model will be identical to that used in the like-for-like replacement decision (Figure 3-8) with the exception of the units on the abscissa.   Before starting the comparison, it is essential that the equipment manager possess good production data on both the Defender and the Challenger.

What complicates this type of decision more than the others is the team nature of generating units of production.   When a production capacity replacement decision is made, more than just one type of equipment is affected and decisions should be based on cost and production for *systems* rather than for individual units.   An example of a production system would be an excavator with its assigned articulated dump truck and all of the other equipment necessary to maintain the haul

roads and compact the fill. When making a production capacity replacement decision the NEL's that are compared must be expressed not only in terms of the costs directly associated with the prime movers in question, but also with indirect costs that reflect the team nature of the production effort.

To accomplish this, URL's for the assisting units must also be calculated in terms of machine age in units of production assistance. Assisting units include not only the dozers and the excavators, but any graders, compactors, water trucks, or other equipment associated with the production. The appropriate optimum average cost per unit of assistance is then added to the NEL of each of the production machines. After this is done, the comparison can be made and a decision obtained.

RULE: *When making a production capacity replacement decision, replace the Defending System with the Challenging System if the URL gradient for the Defending System is greater than T* for the Challenging System.*

3.4.7   Retire

The retire decision is the final decision to discuss. This takes place when the equipment manager has made a decision to sell a given piece of equipment and not replace it. On the surface, this problem seems fairly straightforward. There is no Challenger to forecast costs for. There is no fleet of equipment that must be taken into consideration. If there are no other external influences, the decision model is quite simple. The cumulative cost model should be developed using machine age in cumulative hours of use. The DMCL for the machine should be determined. The machine should be sold when the DMCL has expired. If the DMCL has already passed, the machine should be sold immediately. Graphically, the evaluation is identical to that which was used for the evaluation of alternatives in the initial acquisition decision (

Figure 3-6).

A complicating factor is the fact that the residual value of the machine could be put to work elsewhere in the company, contributing to profitability. The equipment manager can take this into consideration by including a collateral cost that accounts for the difference in revenue generating potential between the Defender and other investment opportunities.

RULE: *A machine selected for retirement should be removed from service when its URL gradient reaches a minimum.*

3.4.8   Profit Maximization: The Retire Decision

It was mentioned earlier that the cumulative cost model could also be used assist in decision making on the basis of profit maximization. This is done by superimposing a Total Revenue Line (TRL) on the model (Figure 3-9.) The TRL represents the cumulative revenues generated by the
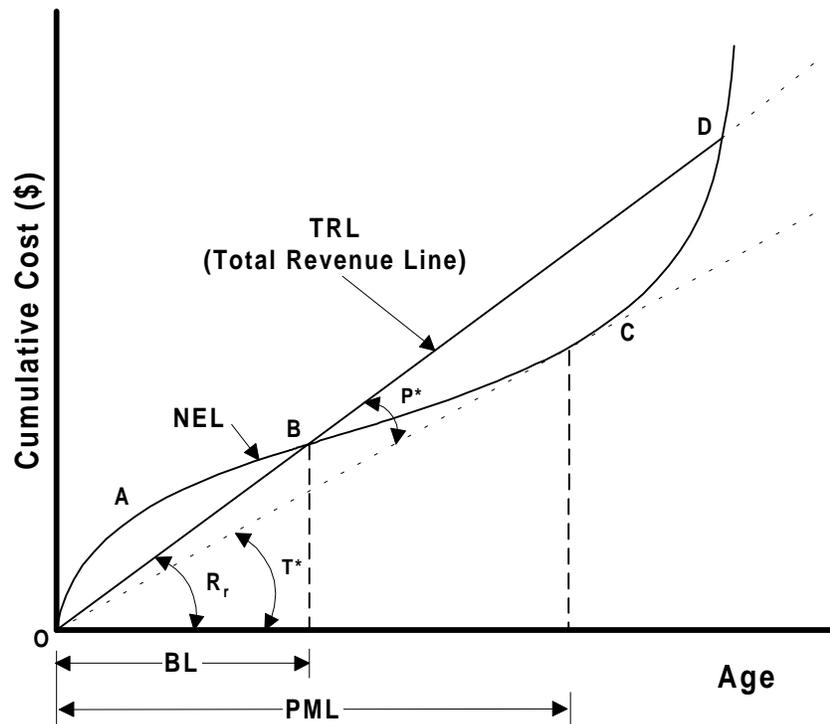
Figure 3-9: Retire (Profit Maximization)

given piece of equipment. The slope of the TRL, R_r, represents the average revenues generated per unit of age. The angular difference between the TRL and the URL at a specific point on the

NEL equates to the marginal profit generated per unit of age at that point in the machine's life. It follows from this that the Profit Maximization Life (PML) is that period which ends when the angular difference between the TRL and the URL is the greatest.

For illustration purposes, the retire decision will be revisited using profit maximization methodology. If the machine is retired very early in its life, profits would be negative. Point B is the breakeven point—average profit would be equal to zero at this point. The breakeven life is represented by BL. For simplification purposes, the slope of the TRL in this example is constant. Since the slope of the TRL is constant, the URL that yields the greatest angular difference will be the one associated with T*. Since the slopes of both the TRL and URL are constant, the optimum marginal profit per unit of age, P*, is also the average profit per unit of age. Also, since the TRL is constant the PML is equivalent to the DMCL as described earlier in the discussion of theory.

RULE: *The optimum lifespan for retiring a machine based on profit maximization occurs when the angular difference between the slopes of the TRL and the URL is maximized.*

3.5   SUMMARY

In this chapter, the Cumulative Cost Model was presented and discussed in detail. It was shown that the model is ingenious, intuitive, and flexible in the scope of decisions that it supports. Although the cumulative cost model has tremendous practical and academic potential, the key to its successful implementation is the accurate definition of the NEL. The assumption has been made that the NEL is concave. This is the basis of the optimization function. If the NEL is not concave—if the average owning, operating, and collateral costs on a machine do not increase with age, the cumulative cost model (and most other models presented to date) are invalid. The equations that make up the NEL must be fully developed and defined to ensure the cumulative cost model yields valid results. This dissertation is the start of that process.

This concludes Part I of this dissertation, *Understanding the Challenge*. The topic has been introduced, the literature has been reviewed, and the basic model has been defined. The reader

should now be ready to understand Part II, *Defining the Work*. Part II commences with Chapter 4 wherein the structural and statistical issues concerning the data will be discussed.

CHAPTER 4: THE DATA

The first step of *Defining the Work* for this dissertation is gaining an understanding of the characteristics of the data to be analyzed. After the data is understood, a methodology can be developed and the analyses can be performed. A hypothetical data set has been formulated to illustrate many of the peculiarities of repair cost data. This data set is depicted in Table 4-1. The data set consists of five machines. The machines are of two different types, were purchased at different times, and have differing data collection methods. This data set consists of cumulative hours of use and cumulative repair costs for the machines. It will be used throughout this chapter.

There are two general categories of issues concerning the data. They are structural and statistical issues. The structural issues pertain to getting the data into a usable format. The statistical issues pertain to making this study statistically valid.

4.1   STRUCTURAL ISSUES

There are number of structural issues with the data that must be addressed prior to formulating any plan for analysis. These include:

- Use of field data

- Differing machines

- Differing times

- Data collection periods

- Machine age

- Cost

- Data pairing

- Confidentiality

Table 4-1: Illustrative Data Set

	Machine #A1		Machine #A2		Machine #A3		Machine #B1		Machine #B2	
List Price	$ 150,000		$ 125,000		$ 157,000		$ 350,000		$ 350,000	
Month	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost
JAN	0	$0	4634	$34,126						
FEB	192	$135		$34,618						
MAR	202	$636	4969	$34,746						
APR	202	$1,453		$35,485						
MAY	202	$1,921	5294	$36,230						
JUN	554	$2,844		$36,377			0	$0	0	$0
JUL	705	$3,825		$37,307			128	$1,030	139	$720
AUG	764	$4,268		$37,984			324	$40,325	287	$2,315
SEP	818	$5,024		$38,903			453	$41,651	442	$4,309
OCT	829	$5,937		$39,664	0	$0	586	$43,591	602	$4,573
NOV	914	$6,198	5817	$39,757	113	$888	754	$44,117	750	$6,351
DEC	950	$6,631		$39,779	245	$1,644	883	$45,377	878	$6,501
JAN	1112	$6,886		$40,096	266	$2,089	1064	$46,199	1011	$7,442
FEB	1176	$7,834		$40,532	424	$2,399	1243	$46,704	1194	$8,944
MAR	1230	$8,787		$40,623	534	$3,167	1436	$48,378	1365	$10,190
APR	1263	$9,600		$40,663	641	$2,075	1581	$49,836	1558	$11,327
MAY	1444	$10,464	6186	$41,488	837	$2,294	1770	$49,873	1753	$11,517
JUN	1574	$11,381		$42,140	990	$2,817	1898	$50,414	1872	$13,323
JUL	1767	$11,637		$42,624	1187	$3,653	2085	$50,530	1979	$14,763
AUG	1955	$12,629	285	$42,887	1215	$4,247	2269	$51,427	2091	$15,719
SEP	2147	$13,463		$43,186	1321	$4,335	2409	$51,606	2221	$15,924
OCT	2310	$13,838		$43,875	1367	$4,455	2552	$52,874	2415	$16,448
NOV	2352	$14,053		$44,399	1564	$4,794	2661	$54,774	2517	$17,127
DEC	2508	$14,887	635	$45,077	1720	$5,188	2764	$55,256	2675	$17,363
JAN	2564	$15,088		$45,308	1871	$5,964	2950	$56,032	2871	$19,248

This section will discuss each of these issues as they pertain to this research. The structural issues are reflections of things that must be done to the data to get it to the condition where it is suitable for analysis.

4.1.1   Field Data

The fact that this study will be based on field data was introduced in Chapter 1. It is not possible to obtain laboratory data that is suitable for this study. Field data should provide a realistic

picture of how repair costs escalate over time. The downside of field data is that it can contain "noise". Things can happen that distort the reliability and contribution of the data associated with a particular machine. The more machines that are part of the study, the less of an effect these distortions will have.

The data were collected by a wide variety of people in a wide variety of organizational positions. Each company had its own unique set of data collection procedures. In some cases, the data would pass through multiple hands in hard copy format before its entry into the accounting databases. Each company involved was visited and their data collection processes were investigated. The purpose of these visits was to validate the accuracy of the collection effort. Although occasional glitches in the data were encountered, the collection of cost data by all of the companies involved was deemed reliable. The collection of data concerning hours of use was a different matter. Some companies only tracked billable hours for their machines in their databases. They did not track the actual hours of usage. For these companies, a solution was devised that used their oil change records. This solution will be further discussed in Section 4.1.3.

Despite the fact that all of the companies had reliable cost data collection methods, there were still some observed errors in the data that were obtained. The most common error was that of negative repair costs for a given month. This is illustrated in area "A" of Table 4-2. When the question of negative costs was posed the companies involved, the answer obtained was that the negative charges were due to either overcharges or mistaken charges that occurred in an earlier month. To fix this error, the negative charges were removed from the preceding month (or months) and the negative charge was eliminated from the data set. This is illustrated in area "A" of Table 4-3. The reason for doing this was to eliminate the false fluctuations in cumulative repair cost induced by adding and subtracting charges that should not have been there in the first place.

Another error, though not as common as negative charges, was that of replaced hour meters. This was obvious in that either the cumulative cost at time zero was not zero or the cumulative hours associated with a given machine went down with the passage of time. This is illustrated in area "B" of Table 4-2. The fix for this problem was to first confirm that the meter had been replaced. After confirmation of meter replacement was obtained, the cumulative hours at time of

Table 4-2: Data Problems

Month	Machine #A1		Machine #A2		Machine #A3		Machine #B1		Machine #B2	
List Price	$ 150,000		$ 125,000		$ 157,000		$ [D] 00		$ 350,000	
	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost
JAN	0	$0	4634	$34,126						
FEB	192	$135		$34,618						
MAR	202	$636	4969	$34,746						
APR	202	$1,453		$35,485						
MAY	202	$1,921	5294	$36,230						
JUN	554	$2,844		$36,377			0	$0	0	$0
JUL	705	$3,825		$37,307			128	$1,030	139	$720
AUG	764	$4,268		$37,984			324	$40,325	287	$2,315
SEP	818	$5,024		$38,903			453	$41,651	442	$4,309
OCT	829	$5,937		$39,664	0	$0	586	$43,591	602	$4,573
NOV	914	$6,198	5817	$39,757	113	$888	754	$44,117	750	$6,351
DEC	950	$6,631		$39,779	245	$1,644		$45,377	878	$6,501
JAN	1112	$6,886		$40,096	266	$2,089		$46,199	1011	$7,442
FEB	1176	$7,834		$40,532	424	$2,399	1243	$46,704	1194	$8,944
MAR	1230	$8,787		$40,623	534	$3,167	1436	$48,378	1365	$10,190
APR	1263	$9,600		$40,663	641	$2,075	1581	$49,836	1558	$11,327
MAY	1444	$10,464	6186	$41,488	837	$3,294	1770	$49,873	1753	$11,517
JUN	1574	$11,381		$42,140	990	$2,817	1898	$50,414	1872	$13,323
JUL	1767	$11,637		$42,624	1187	$3,653	2085	$50,530	1979	$14,763
AUG	1955	$12,629	285	$42,887	1215	$4,247	2269	$51,427	2091	$15,719
SEP	2147	$13,463		$43,186	1321	$4,335	2409	$51,606	2221	$15,924
OCT	2310	$13,838		$43,875	1367	$4,455	2552	$52,874	2415	$16,448
NOV	2352	$14,053		$44,399	1564	$4,794	2661	$54,774	2517	$17,127
DEC	2508	$14,887	635	$45,077	1720	$5,188	2764	$55,256	2675	$17,363
JAN	2564	$15,088		$45,308	1871	$5,964	2950	$56,032	2871	$19,248

replacement were used as the baseline. This fix is illustrated in area "B" of Table 4-3. In cases where the cumulative hours at time of replacement were not available, the machine was eliminated from the data set.

A third problem, which was only encountered twice, was that of machines damaged in accidents. Some companies account for accident repairs under separate codes, but others include them as part of general repairs. The accidents were noticeable by very large charges that could not be identified as rebuilds or major component overhauls. This is illustrated in area "C" of Table 4-2.

Table 4-3: Structural Solutions

Month	Machine #A1 Cum. Hours	Machine #A1 Cum. Cost	Machine #A2 Cum. Hours	Machine #A2 Cum. Cost	Machine #A3 Cum. Hours	Machine #A3 Cum. Cost	Machine #B1 Cum. Hours	Machine #B1 Cum. Cost	Machine #B2 Cum. Hours	Machine #B2 Cum. Cost
List Price	$150,000		$125,000		$157,000		$350,000		$350,000	
JAN	0	$0	4634	$34,126						
FEB	192	$135		$34,618						
MAR	202	$636	4969	$34,746						
APR				$35,485						
MAY			5294	$36,230						
JUN	554	$2,844		$36,377			0	$0	0	$0
JUL	705	$3,825		$37,307			128	$1,030	139	$720
AUG	764	$4,268		$37,984			324	$1,122	287	$2,315
SEP	818	$5,024		$38,903			459	$2,448	442	$4,309
OCT	829	$5,937		$39,664	0	$0	586	$4,388	602	$4,573
NOV	914	$6,198	5817	$39,757	113	$888	754	$4,914	750	$6,351
DEC	950	$6,631		$39,779	245	$1,644	88	$6,174	878	$6,501
JAN	1112	$6,886		$40,096	266	$2,089	10	$6,996	1011	$7,442
FEB	1176	$7,834		$40,532	424	$2,399	1243	$7,501	1194	$8,944
MAR	1230	$8,787		$40,623	534	$2,075	1436	$9,175	1365	$10,190
APR	1263	$9,600		$40,663	641	$2,075	1581	$10,633	1558	$11,327
MAY	1444	$10,464	6186	$41,488	837	$2,294	1770	$10,670	1753	$11,517
JUN	1574	$11,381		$42,140	990	$2,817	1898	$11,211	1872	$13,323
JUL	1767	$11,637		$42,624	1187	$653	2085	$11,327	1979	$14,763
AUG	1955	$12,629	6471	$42,887	1215	$247	2269	$12,224	2091	$15,719
SEP	2147	$13,463		$43,186	1321	$4,335	2409	$12,403	2221	$15,924
OCT	2310	$13,838		$43,875	1367	$4,455	2552	$13,671	2415	$16,448
NOV	2352	$14,053		$44,399	1564	$4,794	2661	$15,571	2517	$17,127
DEC	2508	$14,887	6821	$45,077	1720	$5,188	2764	$16,053	2675	$17,363
JAN	2564	$15,088		$45,308	1871	$5,964	2950	$16,829	2871	$19,248

In one case a charge of almost one-tenth the value of the machine was made in the first 500 hours of operation—a period when most repairs are covered under warranty. The fix to this problem was the elimination of the repair charges due to the accident. This is shown in area "C" of Table 4-3. Some smaller repairs due to abuse probably were not eliminated. This should be noted as a shortcoming of some of the data.

4.1.2   Differing Machines

As mentioned in Chapter 1, units of construction equipment can vary in many regards. This is why a cumulative cost index (CCI) will be used in this research instead of raw dollar figures. There can be both physical and usage differences between different machines. The physical differences are those that can be seen just by looking at the machine. These can be categorized in terms of equipment class, group, and brand. The usage differences are less apparent. Differences that fall into this category are those that relate to the application the machine normally performs and those that relate to the company that owns the machine.

The classes, or types, of construction equipment available vary considerably in what the design intent of the machine is. The various classes of equipment are well understood. A good discussion of the different classes of equipment and their uses can be found in Peurifoy, et. al. (1995). The sample data set has machines from two different classes. There are three machines from the "A" class and two from the "B" class (see Table 4-1).

Within the general classes of equipment, there are also size groupings. Track-type tractors, for example, can vary in weight from 15,000 lbs. to over 200,000 lbs. (Caterpillar Performance Handbook, 1996). The differences in purchase price can be just as extreme. Construction companies typically track their equipment by size groupings within the given types. The size groupings are at the discretion of the equipment manager. Typical groupings are those based on horsepower, weight, bucket size, and loaded capacity. Different companies group equipment differently. In this dissertation, groupings will be applied consistently across companies. It will be ensured that the equipment groups investigated include machines of the same size for each of the different companies.

Construction equipment also varies by manufacturer. There can be wide variations in purchase price, quality, and reliability among machines that are of the same type and in comparable size groups but from different manufacturers. This research will not separate equipment my manufacturer.

An attempt will be made to standardize the differences of equipment, at least in terms of price, by using the *cumulative cost index* (CCI). A method of comparing unlike pieces of equipment was

needed for the purposes of this study. Repair costs, as well as initial purchase price, can differ considerably between the different classes and groups of equipment. They can vary within a group of equipment depending on manufacturer. It is desirable to be able to compare how repair costs accumulate across different classes of equipment. One of the end products of this study will be a tool that can forecast repair costs in terms of dollars per hour. For the purposes of economic modeling within the constraints of the cumulative cost model, it is important that the quantities reflected be associated with cumulative costs—not instantaneous cost per hour.

A convenient way to compare repair costs of unlike machines is to index them to the initial price of the machine. Unlike machines and their repair costs can then be compared and equations can be developed. The formula that will be used to calculate the response variable is:

$$CCI_t = \frac{\sum_0^t (P_t + L_t + O_t) + PP_0}{PP_0}$$

Equation 4-1

Where:

   CCI_t = cumulative cost index at time t

   P_t = cost of parts at time t

   L_t = cost of labor at time t

   O_t = other maintenance costs at time t

   PP_o = new list price of the machine

Parts, labor, and other maintenance costs are cumulative costs. The CCI will provide the common ground by which comparisons can be made between non-identical machines. The CCI will form the ordinate of the Cumulative Cost Model for the purposes of the initial analyses. It should be noted that the minimum value for the cumulative cost index is one. Also, it should be noted that the cumulative cost index should not decrease with increasing machine age. It can increase or remain constant but a decrease is not normally possible.

The instantaneous hourly repair costs of a machine can be back calculated from the cumulative cost index equations by taking the derivative of the regression equation evaluated at the point in

time of interest. Alternatively, the repair costs for a job can be estimated by calculating the cumulative cost index at the start and end points of the job and taking the difference of the two. Techniques for these two manipulations will be described in Chapter 8.

The CCI is not a perfect standardization index. But, it should allow some means by which unlike types and groups of equipment can be compared. It could also serve as a means by which different makes of machine could be compared. Comparisons of differing manufacturers are not an objective of this study. In fact, many construction companies make such comparisons very difficult by the ways in which they compose their fleets. Although there are similarities among brands, it is easier and cheaper to train mechanics and stock spare parts if all machines in a particular class come from one manufacturer. This is certainly the case in nearly all of the fleets observed for this dissertation.

4.1.3 Machine Age

The abscissa of the cumulative cost model has been generically referred to as "age" up to this point. "Age" is a very general term, though. What is important is how the machine has aged. For this reason, the abscissa of the model will be referred to as *machine age*. There are three types of machine age that are worthy of discussion at this point. These are machine age in calendar terms, machine age in units of production, and machine age in cumulative hours of use.

In textbook economic replacement problems and in most equipment replacement models, the units of the abscissa would be *calendar age*. This is convenient because it is relatively easy to measure a machine's age in calendar terms. One needs merely to subtract the original purchase date from the current date and the result is the machine's calendar age. Many of the costs associated with heavy equipment are *not* best depicted by the passage of calendar time. Specifically, machine repair and maintenance costs do not accrue as a result of the passage of time (in most cases).

A separate problem with the use of calendar time as the abscissa is the cyclical nature of the construction business. Weather, the economy, and a company's success in bidding projects are each important factors in determining whether or not a machine is used.

An additional concern with the use of calendar age is that as a machine nears the end of its useful life, a company will use it less and less (Terborgh, 1949). When a machine is new, it will be used quite a bit more than when it is old. On a calendar basis, a new machine might have more repair costs incurred than an old machine—but have substantially more production associated with the accumulation of those repair costs.

Machine age in *units of production* is the measure of how much work a machine has actually accomplished. There are a number of difficulties associated with defining machine age in units of production. First is the difficulty of defining exactly what a unit of production is for a particular machine. For some machines, this is an easy task. Units of production for a haul unit could be the movement of some volumetric measure over some linear distance. For some equipment, it is very difficult to define exactly what a unit of production would be. A motor grader tasked with the maintenance of a haul road is a good example. Production could be defined in a number of ways—none of which are wrong. It could be in terms of earth that is physically moved. Or, production could be measured by the utility that results from the haul units being able to travel at greater speed the haul road.

The actual quantification of production units can also be a difficult task. In some cases, modern technology has made it very easy to ascribe specific units of production to specific machines. The Vital Information Management System (VIMS) by the Caterpillar company provides an outstanding tool for measuring the actual production of new haul units (Kannan, 1997) using electronic sensors. For other types of equipment, measuring production is usually a more difficult, manual process. Many companies track *job* productivity, or the productivity of a team of machines, but do not track the productivity of individual units.

Machine age in *cumulative hours of use* is the final type of aging that will be discussed. The distinction between calendar age and usage age was alluded to in Chapter 2. Few parts on a machine wear out or break over time even if the machine does no work. An example of one part that does is the rubber hose. Given enough time, a rubber hose can deteriorate to the point that it is unusable just from environmental exposure. What is more often the case, at least in companies

that work their machines regularly, is that parts on a machine break as a function of how much work that machine has done.

Machine age in cumulative hours of use can be likened to odometer readings on automobiles. This age is a measure of how many hours the machine physically operated. Age in cumulative hours of use dampens many of the cyclical variations in operating cost.

Machine age in cumulative hours of use provides a linkage to units of production. The linkage is not perfect, however. At times, machines may be running in idle. They may be travelling to and from the job site. Age in cumulative hours of use is blind to these situations. In that respect, it is not a perfect measure of the "hard" work that usually causes wear and tear on parts.

Considering all three of the machine ages defined above, machine age in cumulative hours of use was chosen as the abscissa for our model. It strikes a balance between the availability of data and the applicability of results. The difference between calendar age and age in cumulative hours of use can be observed in the illustrative data set (Table 4-1). There is an abundance of data concerning calendar age, but most machines do not break down primarily as a result of calendar aging. Machine age in units of production provides of a very good measure of how much work the machine has accomplished—there is a dearth of available data, however. The data for machine age in cumulative hours of use is not always easy to get—but it is available. There will be some bias between the cumulative hours of use and the actual productive hours of use, but it is felt that this bias is acceptable considering the alternatives.

There are three types of hours that could be tracked by construction companies. *Billed hours* are those hours for which a company charges a particular job for the use of a machine. These hours may or may not be an accurate reflection of how much work the machine has actually accomplished. Sometimes, jobs are charged for using a piece of equipment for 40 hours per week whether the machine is actually used 40 hours or not. In other cases, the billed hours are those hours that are reported by the site superintendent. Sometimes these numbers are intentionally under-reported to make the job appear more profitable.

Clock hours are the number of hours that a machine was actually running. They are a measurement of time. One way to track clock hours is by worksheets that equipment operators fill out on a daily basis. The amount of time they spent in their machine would be the clock hours for that machine.

Meter hours are those hours taken from a meter that is a mechanical part of the machine. Sometimes the meter is hooked to the engine, sometimes it is hooked to the transmission. Engine meters provide some measure of how many revolutions the engine has had. Transmission meters track the number of revolutions of the transmission. The distinction between the two is that engine meters run when the machine is in neutral gear, and transmission meters run only when the machine is in gear. The output of the meter is scaled to approximate "hours" of use, but the output is actually a measurement of how many mechanical revolutions there were of the engine or the transmission. Meter hours will be the measurement of choice for this study. They are not a perfect measure for how much work a machine has done, but are a better measure than either of the other two methods.

Not all companies track the accumulation of meter hours on their machines as a matter of policy. This can lead to some problems in the data collection effort. These problems are not insurmountable, though. Many companies that do not track their meter hours explicitly have the data available through other sources. Most companies participate in oil-sampling programs to one extent or another. The points in machines' lives at which these oil samples are taken are usually recorded in terms of a calendar date and in terms of meter hours. By associating the calendar date of the oil change with monthly cost data, cumulative costs for a given number of cumulative hours can be determined. This procedure will be described in greater detail in Section 4.1.7.

4.1.4   Differing Times

Two time effects on the cumulative cost index (CCI) of a given piece of machinery are cumulative hours worked and calendar age. The first effect, that of cumulative hours worked, is the effect that is most important to this research. The second, that of calendar age, is not the primary focus of this research but should not be blatantly set aside as unimportant. The impact of inflation can be a major concern when trying to make an informed business decision regarding cash flows that

take place over any appreciable length of time. Most companies keep equipment in their fleets for at least five years. During that time, the economy could be subjected to any number of twists and turns. Inflation is "the decrease in purchasing power of the medium of exchange caused primarily by governments which spend more than they can obtain through taxation or through borrowing from savers" (Schultz, 1976).

The machines in this study ranged in age from 1987 models to 1996 models. The machines were purchased at different times and operated over different periods. In actual dollars, 1987 machines were considerably cheaper to purchase than 1996 machines—in constant dollars, the list prices changed very little. Area "D" of Table 4-2 provides an illustration of this. The three different machines were purchased at different times—the initial list prices vary accordingly. In actual dollars, a repair made in 1996 cost more than a repair made in 1988—in constant dollars, the repairs cost essentially the same. All expenditures will be adjusted for inflation by indexing to a common base. The chosen base year was 1987 because this was the date of the earliest data.

4.1.5   Data Collection Periods

Data collection and reporting periods vary from company to company. Most companies collect and report their cost data on a least a monthly basis. Some do it on a weekly basis. Companies that explicitly keep track of meter hours do so on either a weekly or monthly basis. Companies that do not track meter hours usually track billed hours concurrently with their cost reporting. Because of these variations in collecting and reporting techniques, some companies will have more data points available for their machines than other companies for similar ranges of cumulative hours. In the illustrative data set (Table 4-1), Machine #A1 and Machine #A2 had their data pairs collected using different methods. It can be seen that there are considerably more data points available for Machine #A1 than Machine #A2.

Another aspect of data collection periods is that the cumulative hours of use within reporting periods can vary considerably for any given machine and from machine to machine. Once again, this can be seen by comparing Machine #A1 with Machine #A2 in the illustrative data set. What this means is that even within a given company different data points will represent different accumulations of hours. The problems with this are more of a statistical nature than of a

structural nature. This will be addressed in greater detail in Section 4.2.3. It is important to note here that the data structure will differ from company to company and from machine to machine.

A third aspect of data collection periods is that different ranges of machine age are available for each machine. Once again a comparison of type "A" machines illustrates this. The data available for Machine #A2 start at over 4000 cumulative meter hours and cover the range up to around 7000 meter hours. The data available for Machine #A1 start at zero and cover a range up to around 2500 hours. The data for Machine #A3 start at zero and only go up to 2000 meter hours. Three different machines with three different ranges of cumulative hours.

4.1.6   Cost

Many different costs have been proposed for inclusion in economic models. They range from straightforward, tangible costs such as fuel consumed to complicated, intangible costs such as the cost of obsolescence. These costs can be broken down into three broad categories: direct costs, provisional costs, and collateral costs. These are reflected in Table 4-4.

Direct costs are those costs that are simply quantified, clear, and directly related to owning, operating, maintaining, and repairing an individual item of equipment. They occur regularly within a given accounting period. Direct costs affect a company's operating budget. They are offset by the revenue stream generated by the piece of equipment in the work that it performs. Direct costs are incurred constantly over the life of the machine. The initial purchase price (P_0) is the first capital outlay that should be considered. Then come the direct costs associated with operating a piece of machinery (E_p). These costs include: fuel, oil, tracks and tires, preventive maintenance, repairs, licenses, taxes, and insurance.

Provisional costs are internal charges made to cover the anticipated cost of discrete events that occur a limited number of times during the life of the machine. Inflows into the provisional cost "money pot" come from charges for future expenditures. The costs of major repairs or rebuilds on a piece of equipment are normally handled by charging a provisional hourly rate and establishing a repair reserve which is balanced across a particular unit, group, or fleet.

Table 4-4: Cost Categories

Cost Category	Owning Costs	Operating Costs
DIRECT COSTS Continuous Inflow Level Not Critical Continuous Outflow	Licenses Insurance Tax	Fuel/Oil/Lubricants Tires & Tracks Ground Engaging Tools Repair Parts & Labor
PROVISIONAL COSTS Continuous Inflow Level IS Important Infrequent Outflow (Once or Twice in Life)	Depreciation* Provision for Replacements* *inflows	Rebuilds
COLLATERAL COSTS Imaginary Inflow??? Level Unknown Imaginary Outflow???	Inflation Cost of Capital	Downtime Failure Obsolescence Low Productivity Technology Versatility
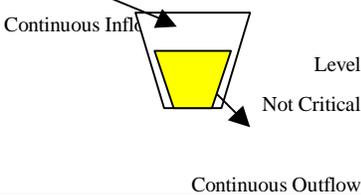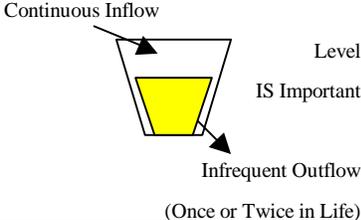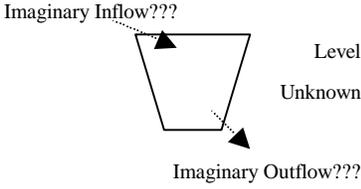
Amortization is another type of provisional cost which allows for the fact that the resale value of a machine decreases from P_0 to S_t with the passage of time. It is given by the equation:

$A_t = P_0 - S_t$ **Equation 4-2**

Where:

t = the time of interest

A_t = the amortization at time t

P_0 = the initial purchase price

S_t = the salvage value of the machine at time t

Collateral costs are more difficult to quantify and are not a part of every model in the literature. Collateral costs include obsolescence costs, associated resource impact costs, lack of readiness costs, service level impact costs, and alternative method impact costs. Obsolescence costs are those cost that "occur" as a machine ages technologically. New technologies that provide increased productivity, reliability, or versatility contribute to obsolescence costs. Obsolescence costs can take the form of higher repair or production costs for units that do not have the new technology. They can also be manifest in the form of bids for jobs that are lost because of these higher costs.

Collateral costs also occur when a machine breaks down. There are four sub-component costs associated with lack of availability and downtime (Vorster and De La Garza, 1990). *Associated resource impact costs* concern the effects of the failure on other components of the team. *Lack of readiness costs* accrue as a result of resources that could be used not being in a state of repair such that they can be used. *Service level impact costs* measure the decreased productivity of a fleet of equipment when a portion of that fleet has failed. *Alternative method impact costs* occur when a different method of production must be used due to the failure of a given component of the original production team. The prediction of the timing of most collateral costs is difficult, but the quantification of collateral costs can be even more involved.

Determining which costs to include in a model is a very complex process. Most equipment managers should be comfortable with the calculation of direct costs and provisional costs. Collateral costs can be difficult to estimate—all equipment managers may not be comfortable with them.

The costs that will be isolated for this study are those associated with equipment maintenance and repair. The components of owning and operating cost that seem to be most appropriate to maintenance considerations and this study are parts, labor, lubrication, and other miscellaneous maintenance costs. Tires, undercarriage components, and ground engaging tools are often tracked in accounts separated from general repairs to the machine and generally have much shorter lives than the equipment they are associated with. The useful lives of these "expendables" are highly dependent upon local conditions and operator skill. Bucket teeth on excavators will

last longer when used in common earth than when used to load rock. Operators who allow the tires to spin or skid on their machines will go through tires much more quickly than those who do not. To make the models developed less condition-sensitive, costs for these three items will not be considered. This is not to say that these costs are either unimportant or small. Tires and tracks especially can cost tremendous amounts of money. The costs associated with these items are worthy of investigation by themselves and are beyond the scope of this study.

Falling under the general heading of cost is the issue of initial purchase price. Different companies can be charged different amounts for the same machine. Trade-in and lease allowances can cloud the real cost of purchasing equipment. For the purposes of this study, list price will be used for all machines—regardless of what companies paid for them. This is another standardization measure that should increase the reliability of the GEL.

4.1.7 Data Pairing

Regression analysis requires both a value for the regressor (cumulative hours of use) and the response (CCI) for each point that will be part of the analysis. One of the problems with the data was that the data pairs were not always in the same database or even in the same computer. Also, sometimes one or both numbers in the data pair were missing. Some companies did have integrated databases that could provide meter hour and cost information in the same query—these were the exception rather than the rule. Other cost databases contained detailed entries about costs and the dates on which they occurred, but contained no information concerning the meter hours of the equipment on those dates.

Meter hours were obtained from different sources in these cases. Some companies maintain separate preventive maintenance databases that contain tracking information on meter hours versus the timing of oil changes. Another source for meter hour information was oil sampling databases. Many companies participate in oil analysis programs whereby samples of oil are analyzed on a periodic basis to provide warning of impending failure of specific components. When these samples are taken (usually during all preventive maintenance oil changes), the date and meter hours of the machine are recorded. For one company, the oil sampling database was not easily accessible. In this instance, meter hours on specific dates were obtained by going

through the maintenance receipts for each machine to be analyzed. This is a fairly laborious and time-intensive process.

To combine the data from the two different sources, calendar date was used as the common point. Costs from months in which oil changes took place were associated with the meter hour readings from those particular months. One problem with this was that cost data were reported in end-of-the-month increments and the oil changes did not necessarily take place at the end of the month. Oil changes that took place on or prior to the 15th of the month in question were assumed to have taken place at the beginning of the month. Oil changes that took place after the 15th of the month were assumed to have taken place at the end of the month. There is a certain amount of error induced by this assumption. The CCIs associated with oil changes that took place early in the month is probably understated because all cost that had taken place up to that oil change were not necessarily included. By the same token, the CCIs associated with oil changes that took place late in the month are probably overstated because they could include expenditures that occurred after the oil change. These errors should be offsetting in the long run.

Of utmost importance is that there be no more than one data pair per month (since the cost data were tabulated on a monthly basis) and no more than one data pair per oil change. To allow more data pairs would be to artificially fabricate data.

4.1.8   Confidentiality

Construction contracting is a very competitive business. Much of the work that companies do is obtained through the competitive bidding process. Jobs can be won or lost by very small margins. When approached for data, many companies were concerned that they could lose their competitive edge or trade secrets by participating in this study. Some of the firms that participated in this study actually compete against each other in the same markets. Their data were obtained only through their trust that this dissertation would not give their competitors insight into the way that they run their businesses.

Respecting the privacy concerns of firms involved, this study will not divulge the names of the companies that provided data. Although their management policies and practices are known and

understood, they will not be discussed in other than general terms. No raw data will be presented either in the body of or the appendices to this dissertation. Any examples or illustrations that use actual costs instead of the CCI will be composed of hypothetical data sets, not real ones.

4.1.9 Summary

As evidenced by the discussion in this section, data in its raw form will not be appropriate for analysis. There are many characteristics of the data that must be either addressed or acknowledged. Some of these characteristics are of the data taken as a whole, some are between companies, and some are within companies. Using the techniques described above and understanding the implications of the structural issues, raw data sets will be transformed into data sets that are suitable for statistical analysis. There are other issues that must be addressed before the analysis can proceed. These are of a statistical nature and will be discussed in Section 4.2.

4.2 STATISTICAL ISSUES

The statistical issues concerning the data to be studied are varied. For this study to have statistical merit, these issues should be understood and addressed. Where assumptions are made, justifications should be provided. Where there are shortcomings, they should be acknowledged.

When performing statistical analyses, there are a number of assumptions that are made about the data that enable hypothesis testing concerning the data to be valid. Violations of those assumptions do not necessarily invalidate hypothesis testing, but they can induce a little more uncertainty into the results obtained. It would be ideal if it could be said for certain that the data used violate no assumptions. This ideal may not be achievable. Linear regression makes the following assumptions concerning the data to be analyzed: they are independent and that the error terms have constant variance which is normally distributed about a mean of zero (Myers, 1990). There are additional problems associated with the data structure that have statistical ramifications: relative dominance, repeated points, and varying intervals.

4.2.1   Data Independence

The independence of data assumption requires that the residual error associated with one data pair in the regression are not related to the errors of other data pairs in the regression. Since the CCI and cumulative meter hours are both cumulative measures, each observation of meter hours vs. CCI is somewhat dependent upon the previous observation for a given machine. It logically follows that the errors associated with these dependent data points may be dependent themselves.

The dependence or independence of the data has no bearing on the quality of fit of the curve. Least-squares regression is a mathematical technique. It develops a solution that minimizes the squared distance to the regression line of all the points fed into it—whether they are independent or not. This could conceivably have an effect on hypothesis testing and in the reliability of any confidence intervals developed as a result of the analyses.

A study was conducted by Mahon and Bailey (1975) on British military vehicles. The purpose of the study was to test the feasibility of implementing a replacement policy based on repair limit theory. As part of that study, the independence assumption was tested on repair intervals and costs. It was concluded that repairs that occurred more than one year apart could reasonably be assumed to be independent. Repairs that were less than one year apart could not be assumed to be independent. If this were to be applied to construction equipment, one year's worth of usage would equate to approximately 1700 hours for an average machine.

4.2.2   Variance

The variance assumption requires that the distribution of the error terms be constant throughout the range of values of the predicted response. The distribution must also be normal with a mean value of zero. What this means is that there should be no increase in the variability of the CCI with an increase in meter hours. It will be demonstrated that this is not true. New machines do not break down that often and the costs of the repairs are not that high. Typically the CCIs remain very tightly grouped for new machines. As machines get some hours on them, they begin to break down. Some break down more than others do. The CCIs machines with higher cumulative hours of use had greater variance than those with lower cumulative hours. This will

be demonstrated in Chapter 7. Violation of the variance assumption could have an effect on the reliability of hypothesis testing and confidence intervals.

4.2.3   Relative Dominance

The problem of relative dominance is that some machines may have more data pairs than other machines. This could be due to differences in usage, dates of purchase, or data collection styles. Machines that have more data pairs can have more of an influence on the final regression equation than those with fewer pairs can. In the illustrative data set, Machine #A1 has greater relative dominance than Machine #A2 (Table 4-1).

In some ways, this could be considered good—in other ways, bad. The good part is that the machines for which the most is known have the biggest impact on the regression. The bad part is that there is no way of knowing that the machine that is permitted to have an extra influence on the regression is an "average" machine. If it has uncharacteristically high or low CCIs for the number of meter hours it has, it could skew the regression and render estimates of T* and L* less reliable.

4.2.4   Repeated Points

This is the simplest statistical issue to address and solve. The problem is that for any given machine, there may be long periods of time where it is idle. There are two primary reasons for this. First, the machine could be in the shop for major repairs. In this case, the cumulative meter hours would remain constant and the CCI would climb as the repairs are made. The other possible scenario is one in which the company just doesn't have any work for the machine. In this case, both meter hours and the CCI would remain constant. This is illustrated in area "E" of Table 4-2.

In both cases, since the machine is idle counting more than one point for the same cumulative hours is essentially fabricating data. It would be inferring that something happened when in fact, nothing happened—the machine was idle. The implication of using these repeated points is that a machine that is unemployed or hard broken could have as great an influence on the regression as a machine that is gainfully employed. This problem is not the same as cumulative hours increasing

with no corresponding increase in CCI—that scenario would simply mean the machine had a good month.

4.2.5   Data at varying intervals

This problem is similar to relative dominance. Some machines work more hours in a given month than other machines. Machines normally work in the range of 100-150 hours per month. But, in bad months they could work zero hours (described above) and in good months they could work as much as 400 hours. For companies that track meter hours on a monthly basis this means that machines that work very little could have as great an impact on the regression as machines that work a great deal—only one data pair per month is allowed regardless of how many hours the machine worked.

For companies that do not track meter hours with their cost data the problem is different, but the effect is the same. Usually oil changes occur at some set interval, every 300 hours is a good estimate. Sometimes the oil change comes late—as late as 500 hours between changes. The oil changes rarely come earlier than they should. Other times, the records available on a given machine can indicate a gap of 1000 or more hours between oil changes—in these cases what probably happened is the documentation of the oil change did not make it to machine's file in the main office. This is illustrated in area "F" of Table 4-2. This can have the same effect as described for differing monthly production. One machine could have a very small interval of cumulative hours between data pairs while another could have a very large interval of cumulative hours between points. The machines with the most points have the greatest impact on the regression.

4.3   POSSIBLE SOLUTIONS

It has been shown that there are many statistical issues concerning these data. Yet, the data can still be analyzed and can still yield meaningful results. This section will discuss how.

4.3.1   Address Independence

The only way to ensure absolute independence of the data pairs is to use only one data pair for each machine in the study. The point that makes the most sense to use is the final, or most recent, point. This will ensure that the regression covers the greatest range of hours. This solution not only solves the statistical problem of independence, it simultaneously clears up any problems relating to relative dominance, repeated points, and varying intervals. Each machine is only represented once. This is illustrated by area "A" of Table 4-5.

From a practical standpoint though, a lot of information is lost by the application of this procedure. This is especially true for companies with small fleets and for companies that have fleets of machines that are of the same age. In the small fleet case, the regression might have to proceed with only three or four data points. If the points aren't evenly spaced along the ordinate, the resulting curve may not truly represent how the CCI grows. In the case of a fleet of near-singular age, the "curve" would be a straight line between the origin and the mean value of the CCI for the points clustered around one small range of cumulative hours. Either of these situations is bad—the curve may not represent the way that CCI increases with increased meter hours.

4.3.2   Address Variance

Non-constant variance can be confirmed by looking at the residual plots from the regressions. The way to solve for non-constant variance, if it is present, is to perform *weighted regression*. Weighted regression will be described in greater detail in Chapter 5. In essence, the points in the regions of cumulative meter hours with the greatest variance would be assigned a weight that gives them less relative importance in the calculation of the least squares solution.

The problem with this is that there must be enough data to accurately assess what the weights should be. If improper weights are used, the "cure" could be worse than the "disease". According to Myers (1990), the weights should be derived using a minimum of 9 observations at each interval of interest. If less than 9 observations are present, it may not be possible to accurately describe the variance. Nine observations does not mean there should be a minimum of

Table 4-5: Regression Data Pairs

	Machine #A1		Machine #A2		Machine #A3		Machine #B1		Machine #B2	
List Price	$ 150,000		$ 125,000		$ 157,000		$ 350,000		$ 350,000	
Month	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost	Cum. Hours	Cum. Cost
JAN	0	$0	4634	$34,126						
FEB	192	$135		$34,618						
MAR	202	$636	4969	$34,746						
APR				$35,485						
MAY			5294	$36,230						
JUN	554	$2,844		$36,377			0	$0	0	$0
JUL	705	$3,825		$37,307			128	$1,030	139	$720
AUG	764	$4,268		$37,984			324	$1,122	287	$2,315
SEP	818	$5,024		$38,903			453	$2,448	442	$4,309
OCT	829	$5,937		$39,664	0	$0	586	$4,388	602	$4,573
NOV	914	$6,198	5847		113	$888	754	$4,914	750	$6,351
DEC	950	$6,631			245	$1,644	883	$6,174	878	$6,501
JAN	1112	$6,886		$40,096	266	$2,089	1064	$6,996	1011	$7,442
FEB	1176	$7,834		$40,532	424	$2,399	1243	$7,501	1194	$8,944
MAR	1230	$8,787		$40,623	534	$2,075	1436	$9,175	1365	$10,190
APR	1263	$9,600		$40,663	641	$2,075	1581	$10,633	1558	$11,327
MAY	1444	$10,464	6186	$41,488	837	$2,294	1770	$10,670	1753	$11,517
JUN	1574	$11,381		$42,140	990	$2,817	1898	$11,211	1872	$13,323
JUL	1767	$11,637		$42,624	1187	$3,653	2085	$11,327	1979	$14,763
AUG	1955	$12,629	6471	$42,887	1215	$4,247	2269	$12,224	2091	$15,719
SEP	2147	$13,463		$43,186	1321	$4,335	2409	$12,403	2221	$15,724
OCT	2310	$13,838		$43,875	1367	$4,455	2552	$13,671	2417	$16,448
NOV	2352	$14,053		$44,399	1564	$4,794	2661	$15,571	2547	$17,127
DEC	2508	$14,887	6821	$45,077	1720	$5,188	2764	$16,053	2675	$17,363
JAN	2564	$15,088		$45,308	1871	$5,964	2950	$16,829	2871	$19,248

nine machines in each data set. It means, for example, that at 2500 cumulative hours of use there must be observations for nine machines. With the data that was available, weighted regression was not advisable.

4.3.3 Address relative dominance

A way that relative dominance could be addressed besides using only the final data pair for each machine is through using average values of the CCI at discrete, evenly spaced intervals. To do this, an interval would be chosen based on how many data are available. For some companies, data might be available to support 100-hour intervals, for others (especially those that do not explicitly track meter hours) it might be a higher number. Data pairs for each machine will be interpolated for the selected intervals. This is illustrated for Machine #A1 in Table 4-6. The mechanics of this will be described in Chapter 6. What is important to emphasize here is that *there can only be one interpolated data pair between any two actual data pairs.* The interpolation of more than one data pair between data pairs would be the fabrication of data. So the interval selected must support the interpolation rule. For the data available for this study, 500-hour intervals worked best. The average value of the interpolated data pairs is then computed for each 500-hour interval. The regression is accomplished on the data pairs that represent these averages.

This only eliminates some of the relative dominance problem. But, it completely eliminates the problem of data intervals and repeated points. The reason only some of the relative dominance problem is removed is that some machines may take part in the determination of more average values than other machines do. Although not a perfect solution to the relative dominance problem, it is deemed adequate. This option would also partially solve the independence issue. Since the data pairs would not necessarily be based on data from the same machines, some of the independence problem would be solved.

This method does create some additional problems, though. Confidence and prediction intervals for such a regression will not have the same meaning as those for other regressions. The prediction intervals will describe a range of possible values for an *average* machine, not a *specific* machine. Additionally, measures of regression performance such as R^2 must necessarily be better because the regression equation is based on values that have had a good deal of their variability removed.

Table 4-6: Interpolated Data Pairs

Month	Machine #A1 $150,000 Cum. Hours	Machine #A1 $150,000 Cum. Cost	Machine #A1 $150,000 Cum. Hours	Machine #A1 $150,000 Cum. Cost
JAN	0	$0		
FEB	192	$135		
MAR	202	$636		
APR				
MAY				
JUN	554	$2,844	500	$2,566
JUL	705	$3,825		
AUG	764	$4,268		
SEP	818	$5,024		
OCT	829	$5,937		
NOV	914	$6,198		
DEC	950	$6,631		
JAN	1112	$6,886	1000	$6,710
FEB	1176	$7,834		
MAR	1230	$8,787		
APR	1263	$9,600		
MAY	1444	$10,464		
JUN	1574	$11,381	1500	$10,860
JUL	1767	$11,637		
AUG	1955	$12,629		
SEP	2147	$13,463	2000	$12,823
OCT	2310	$13,838		
NOV	2352	$14,053		
DEC	2508	$14,887	2500	$14,845
JAN	2564	$15,088		

4.3.4   Address Repeated Points

This is relatively easy to accomplish. Data pairs that have repetitive values for cumulative meter hours are identified. All but one of them is eliminated. But which one to keep? It depends on how one looks at the problem. It does matter which point is selected because in the case of broken machines, the raw CCIs of the repeated points may be different. Even if the raw CCIs are the same (such as when the machine is just idle), the CCIs adjusted for inflation will differ. The question of which point to use is almost of a philosophical nature. The first point is the one that was chosen as the one that makes the most sense to use. The CCI for a machine should include

all repair costs *up to* the point where those hours were reached. Costs incurred after the meter hour reading was reached should be ascribed to the next CCI interval. This solution is demonstrated in area "E" of Table 4-3.

4.3.5   Address Data Interval

Differences in data intervals can be addressed by using the data set described under "address relative dominance". Instead of taking the average of the points at the specified interval, all points at each specified interval would be used in the regression. This solution also partially solves the problem of independence, partially solves the problem of relative dominance, and eliminates the problem of repeated points.

Independence is partially addressed because the number of hours between data pairs is increased in most cases. Relative dominance is only partially solved because some machines could still have greater representation than others could in the regression.

4.4   DEDUCTIONS

There is no perfect solution. The nature of our field data does not permit one. A laboratory experiment is not feasible. Structurally, a number of manipulations must be done to transform a raw data set into one that is suitable for analysis. This will be done to all data sets. Statistically, there is no one clear-cut solution to the problem. To compensate for this and to attempt to find the best possible solution to the problem, the following analyses will be performed:

1. Regression of all data pairs except for repeated points. This method provides the maximum number of data points. It is illustrated in area "B" of Table 4-5.

2. Regression of data pairs at 500-hour intervals

3. Regression of average data pairs at 500-hour intervals

4. Regression of final data pairs—statistically, this is the purest solution

All four data sets will undergo the same analysis for each of the 17 fleets. After the results of the regressions are analyzed, a recommendation as to which method is the best can be made.

4.5   SUMMARY

This chapter has looked at the nature of the data to be analyzed in considerable depth. Issues relating to structure and statistics were discussed. Solutions were proposed for structural and statistical problems. A plan of attack was proposed for how to determine which statistical methodology is best suited for determining cost growth equations. A summary of the structural and statistical issues and their solutions is provided in Table 4-7.

Table 4-7: Issues and Solutions Summary

Structural Issue	Solution
Field Data	Use it and recognize limitations
Differing Machines	Cumulative Cost Index
Machine Age	Cumulative Hours of Use
Differing Times	Inflation Index
Data Collection Periods	Acknowledge the Problem, Address in Stats
Cost	Include only Direct O & O costs
Data Pairing	Oil Sampling Databases
Confidentiality	Do not release specifics
Statistical Issue	**Solution**
Data Independence	Data sets 2, 3, and 4 to varying degrees
Variance	Use weighted regression if possible
Relative Dominance	Data sets 2, 3, and 4 to varying degrees
Repeated Points	Do not use them
Varying Intervals	Data sets 3 and 4 to varying degrees

As was evident in section 4.2, due consideration of statistical matters was necessary in the discussion of the data. This is the inter-relationship between Chapter 4, "The Data", and Chapter 5, "Test Methodology". Formulating the data sets was dependent upon statistics just as developing the statistical methodology will be dependent upon the nature of the data sets.

CHAPTER 5: TEST METHODOLOGY

The second and final step of *Defining the Work* is to define the methodology that will be used for analyzing the data. The reader should now understand the nature of the data and the issues surrounding it. It is now necessary to explain how the analyses will be accomplished within the constraints of the data.

This chapter is presented in three major parts. The first part defines and explains the concept of regression and the particulars of the types of regressions that will be done for this dissertation. After this concept is explained, some data adjustments and checks that are necessary will be discussed. Finally, the analysis procedures will be explained.

5.1   REGRESSION

This section flows from a very generalized discussion on the modeling process to detailed statistical explanations of some of the techniques that will be employed in the analyses to be accomplished.

5.1.1   The Process

According to Box, et. al., model development is an iterative process (1994). The first stage that must be accomplished is the postulation of a general class of models. This is done by considering the various methods available and making a choice of which class of models is most appropriate for the needs at hand.

Two general divisions of models are deterministic models and stochastic models. Deterministic models can consistently yield an exact forecast. For the economic data that will be analyzed, this is not possible. The process is a stochastic process. In other words, there are probabilities involved that impact the accuracy of the forecast (Box, et. al., 1994).

Some review and clarification on why regression was chosen as the model for this research is in order. Subjective models were ruled out because the goal of this research is to make better use of the data that are available to equipment managers. Subjective methods are not very data-intensive. Moving averages and exponential smoothing do not do an adequate job of describing non-linear trends or of forecasting beyond medium range. They also do not yield an equation of the type that would be useful in the cumulative cost model.

The desired equation type is:

$$CCI = f(x) \hspace{4cm} \textbf{Equation 5-1}$$

Where:

       x = Cumulative Hours of Use

Time series methods were not the best choice for the data or the goals of this dissertation. The theoretical basis of time series models is that observations of a phenomenon are taken at specified intervals of time. The values of this phenomenon fluctuate, but the passage of time is not the *cause* of the fluctuations. The dissertation is attempting to describe the causal basis of the accrual of hours of use on a machine in determining the amount of money that must be spent to maintain and repair that machine. *Time* is the *cause* of the fluctuations.

The causal regression-based methodology was selected as the best choice for this modeling problem. This methodology can handle nearly all data patterns, can be used for forecasts across the spectrum of planning horizons, and requires only a moderate level of mathematical ability on the part of the user.

The type of regression that will be used is *least squares regression*. Least squares regression is the most commonly used and best-understood regression method available. The goal of this type of regression is to minimize the residual sum of squares. Residuals are the differences between what the response variable actually is and what the response variable is predicted to be. In this case, the response variable is the Cumulative Cost Index (CCI). The regressor variable is

cumulative meter hours. For a good discussion of least squares regression, refer to Myers'
Classical and Modern Regression with Applications (1990).

An illustration of what regression does is given in Figure 5-1. The figure does not represent
actual data, the lines are exaggerated for illustrative purposes. The three staircase-like lines
represent the growth of CCI for three separate machines. A step function is what actual CCI lines
look like. Although the equations under development model cash flows as continuous streams, in
reality cash flows are discrete and periodic. Since the accumulation of hours is in reality a
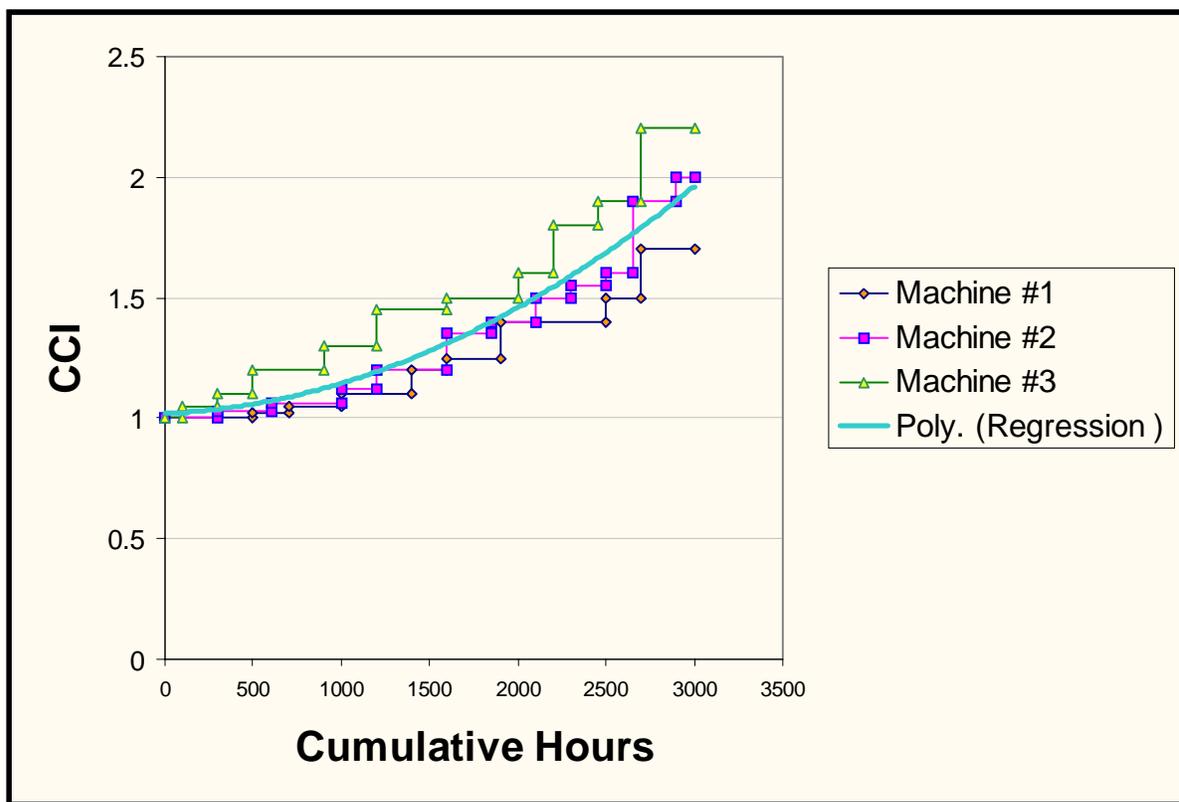
Figure 5-1: Regression

continuous function, the CCI remains constant until another cash flow occurs at which time the
cumulative cost line makes another step upward. The curved line that runs through the middle of
the data is the regression line. This line is a prediction of how an average machine should perform

5.1.2   The Models

With the decision to use regression as the modeling technique made, the functions to be modeled can be chosen. The general form an equation derived from using linear regression is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon$$ **Equation 5-2**

Where:

y = response variable

β_0 = intercept

$\beta_1, \beta_2 \ldots \beta_n$ = coefficients

x_1, x_2, x_n = regressor variables

ε = residual or error term

For this research, the term β_0 will always be equal to one—this will be explained further in the discussion later on regression through the origin. Although it was mentioned in the assumptions in Chapter 1 that there is only one regressor, cumulative meter hours, more than one function of that regressor will be evaluated. The functions of the regressor to be evaluated are: x, x^2, x^3, and e^x. The reason these four terms were chosen was they each can describe the monotonically increasing line that defines the CCI in relation to cumulative meter hours. To evaluate these four terms, the following models will be tested:

1. $y = \beta_0 + \beta_1 x + \varepsilon$

2. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

3. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

4. $y = \beta_0 + \beta_4 e^x + \varepsilon$

5. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_4 e^x + \varepsilon$

6. $y = \beta_0 + \beta_3 x^3 + \varepsilon$

7. $y = \beta_0 + \beta_1 x + \beta_3 x^3 + \varepsilon$

8. $y = \beta_0 + \beta_1 x + \beta_4 e^x + \varepsilon$

9. $y = \beta_0 + \beta_1 x + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

10. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

11. $y = \beta_0 + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

12. $y = \beta_0 + \beta_2 x^2 + \beta_4 e^x + \varepsilon$

13. $y = \beta_0 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

14. $y = \beta_0 + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

15. $y = \beta_0 + \beta_2 x^2 + \varepsilon$

These fifteen models will be tested by running the regression where y = CCI and x = cumulative meter hours. One of the requirements of linear regression is that the behavior of the β terms be linear with respect to the regressors. There are, however, some *non-linear* models that are also of interest. These models can be *transformed* into linear models by using logarithms. By transforming the data an equation of the following type can be found:

$$\ln(y) = \ln(f(x)) + \varepsilon \qquad\qquad\qquad \textbf{Equation 5-3}$$

Problems arise when the equation is converted to its original form:

$$y = f(x) \times \varepsilon \qquad\qquad\qquad \textbf{Equation 5-4}$$

The difference between Equation 5-4 and Equation 5-2 is that the residual (or error) term is *multiplicative* instead of *additive*. This creates additional statistical concerns. By transforming the error structure, additional violations of the statistical assumptions could arise. Specifically, the homogeneity of variance and the normal distribution of the errors assumptions may not hold. The following four models were selected because they are non-linear models that predict a monotonically increasing response with increases in the regressor and because they come close to approximating linear behavior (Ratkowsky, 1990). The first model is a classic non-linear descriptor of growth originally proposed by Freundlich to describe some naturally occurring processes relating to chemistry (1926). The logarithmic transformations are listed to the side of

each model. The presence of "1" in these models compensates for the fact that at zero cumulative meters hours the CCI should equal 1.

16. $y = 1 + \alpha x^\beta$ $\qquad\qquad\qquad$ $\ln(y\text{-}1) = \ln(\alpha) + \beta\ln(x)$

17. $y = 1 + x^\beta$ $\qquad\qquad\qquad$ $\ln(y\text{-}1) = \beta\ln(x)$

18. $y = 1 + e^{\beta(x)}$ $\qquad\qquad\qquad$ $\ln(y\text{-}1) = \beta x$

19. $y = 1 + \alpha e^{\beta(x)}$ $\qquad\qquad\qquad$ $\ln(y\text{-}1) = \ln(\alpha) + \beta x$

5.1.3 Regression Through the Origin

By definition, the Cumulative Cost Index (CCI) should equal one when a piece of equipment has zero cumulative hours on it. It is possible for the CCI to be other than one, but improbable. There should be no cumulative repair costs on a machine with zero hours—if there are they are probably the result of an accident and would be accounted for as described in Chapter 4. What this means in terms of the regression is that the intercept term should be fixed at "1". This was mentioned in the previous section.

Regression through the origin is one way the intercept term can be made constant. Many statistical packages do not allow the researcher to specify a desired intercept point in regression, but they will allow the regression to proceed with no intercept term. This is equivalent to forcing the regression through the point on the Cartesian x-y plane of (0,0). Since our desired intercept is the point (0,1), a value of one must be subtracted from each CCI before using it in the regression. The regression equation obtained from SAS® will be modified after completion of the regression to reflect an intercept term (β_0) of one. Adding the intercept parameter after the regression has been accomplished will not affect the validity of the regression. The curve that is fit will have the same properties, it will just be translated one unit in the positive "y" direction.

Tests will be accomplished to determine whether or not regression through the origin is an acceptable alternative. Using the *intercept* model a 90% confidence interval for the average value of y will be constructed for the case where x = 0. If the value "0" falls in that confidence interval, regression through the origin is an acceptable way to fit the curve (Hahn, 1977).

One of the pitfalls of regression through the origin is the dilution of the applicability of the coefficient of determination, R^2. In regression through the origin, R^2 is measured around the value zero. In standard regression with an intercept term, R^2 is measured about the mean of the fits (Myers, 1990). The two values are not readily comparable—the R^2 statistic as normally calculated is not a valid way to compare intercept vs. no intercept models. It can be used to compare competing no intercept models, but it should be used with the caution that the statistic is telling the researcher something a little different than the normal R^2. The numerical values of this statistic will always be higher in regression through the origin and may mislead the researcher to think that the model is better than it really is. To accommodate this problem, a *corrected* sum of squares will be used to compute the various measurements of effectiveness. The corrected sum of squares was proposed by Myers (1990). A detailed discussion of the corrected sum of squares and a comparison of it to the sum of squares for the intercept case are available in Myers (1990, 33-34).
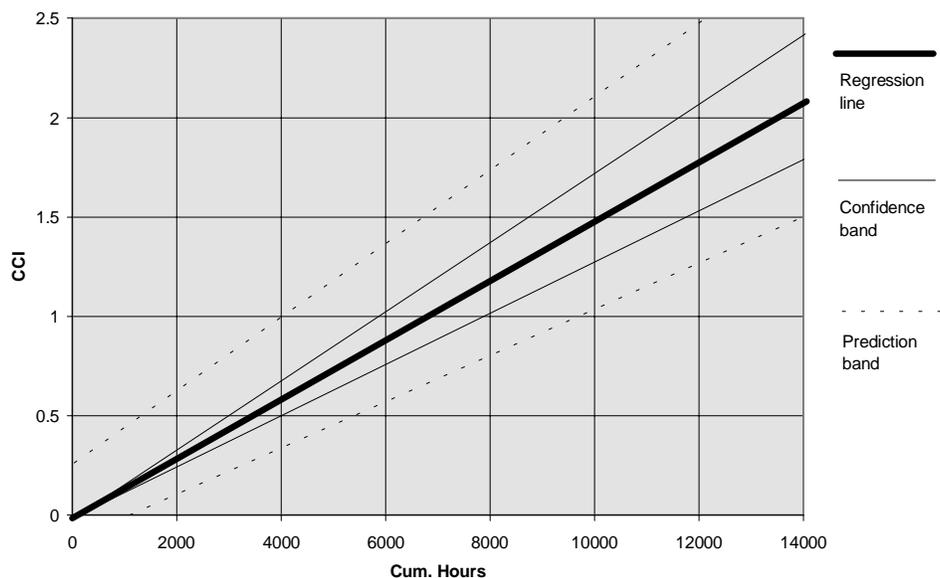
Figure 5-2 - Confidence Bands for Regression Through the Origin

Confidence intervals are also affected by using regression through the origin (Hahn, 1977). Confidence intervals for models with an intercept term typically have an hourglass shape about the regression line. The minimum width of the confidence interval of the response variable occurs at

the average value of the regressor variable. In the case of regression through the origin, the minimum confidence interval width occurs at the origin, where its width is zero. The confidence interval lines then gradually diverge from the regression line with increasing values of the regressor (Figure 10). It is important to note that the increasing width of the confidence interval could correspond with the increase in variance that was discussed in Chapter 4. The prediction intervals follow suit—they are further from the regression line than the confidence intervals, just like in normal regression.

Regression through the origin will be accomplished using the NOINT option of PROC REG in SAS. The NOINT option will not be used for all of the transformed non-linear models to be investigated. The reason for this is that the intercept term becomes part of the function of the regressor after the equation is transformed to its original form from its logarithmic form.

5.1.4 Data Options

In Chapter 4 it was concluded that four types of data sets would be analyzed for each fleet of equipment that is a part of this study. The first data set will be composed of all available data pairs except those that are repeated. The second data set will be composed of readings from each machine in the fleet interpolated to 500-hour intervals. The third data set will be composed of average data pairs derived from the average of interpolated values at those discrete intervals. The fourth and final set to be analyzed will be composed solely of the final data pair associated with each machine.

Because there is an abundance of data but not always an abundance of machines of varying cumulative hours of use, three data sets will be investigated that do not fully address the issue of data independence.

The first set of data points, all data pairs except for those that are repeated only addresses the statistical issue of repeated points—the other issues are not addressed at all. This type of regression utilizes 100% of the data than can "legally" be used. This data set produces the worst violations of the independence of data assumptions of the four sets considered. There is an additional way to use all of the data points available that was considered for this study called

growth curves. The data in this study could not be used in this methodology because cradle-to-grave data is needed on every machine to be analyzed—this was not available.

The second data set, data pairs at 500-hour intervals for all machines in the fleet, addresses some of the shortcomings of the first data set. Relative dominance, independence, and data interval are all addressed to varying degrees.

The average values of data pairs at specified intervals, the third data set, will not provide a solution that is as statistically pure as the first. The issues mentioned above are addressed a little better in this method than in the second, but they are still not fully solved. A new problem that comes up with this data set is that the confidence intervals generated by such an analysis do not provide the same information that they do in other three data sets used. This is because some of the variability is removed from the data before it is analyzed.

Regression on the final data pairs for each machine is the fourth and most statistically pure method of analysis for this dissertation. Intuitively, it may seem odd that using only one data point for each machine is an appropriate method when so much data are available on each machine. As was explained in Chapter 4 there are issues concerning the data that make this method more preferable than others do. This type of regression eliminates the concerns regarding independence, relative dominance, repeated points, and data intervals.

The reason that regression of only the last observation for each machine should work is that, barring influential observations, each final point should fall in the vicinity of the true model (regression line). If there are enough points spread over the range of values of the cumulative hours of interest, it should be possible to develop a statistically sound regression equation from those points.

5.2   PREPARATION OF DATA

Despite all the data filtering, manipulation, and analysis that was described in Chapter 4, there are still a few things that must be done to the data before the actual analyses can proceed. The data must be scaled, some of them must be reserved for cross-validation, and the variance should be assessed.

5.2.1 Data Scaling

Data scaling was found to be necessary after a few trial regressions. The problem is that if the raw value for cumulative hours of use is used as the regressor, some of the coefficients obtained through regression are of such a small magnitude that they are difficult to use and comprehend. Furthermore, because the coefficient obtained for the e^x term was such a small number (or because e^x was such a large number) the statistical computer program would only give error messages when regressions with the e^x term were included. Some graphs should facilitate the understanding of this issue.
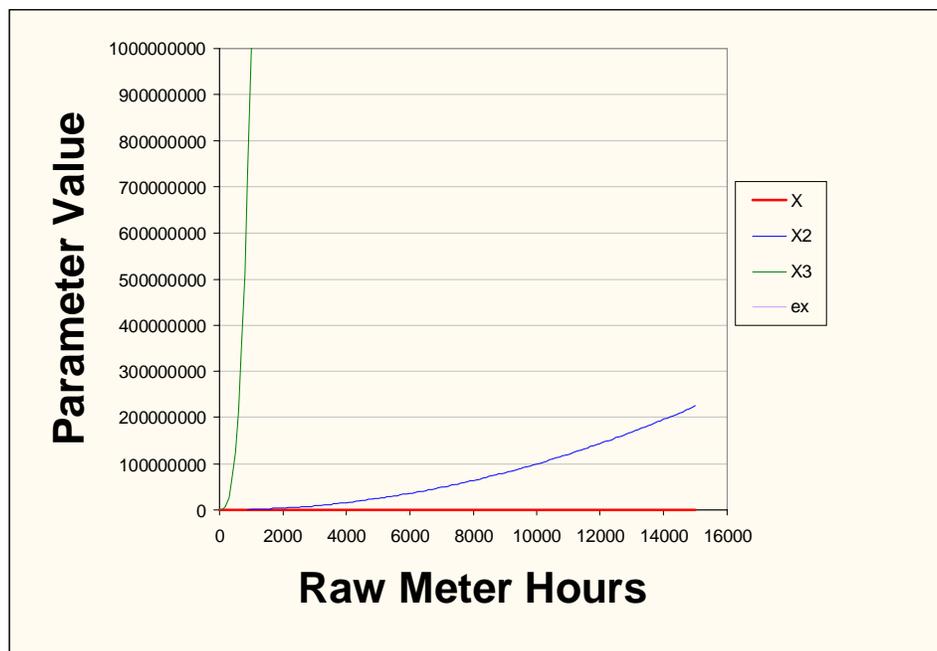
Figure 5-3: Regressor Values: Raw Data
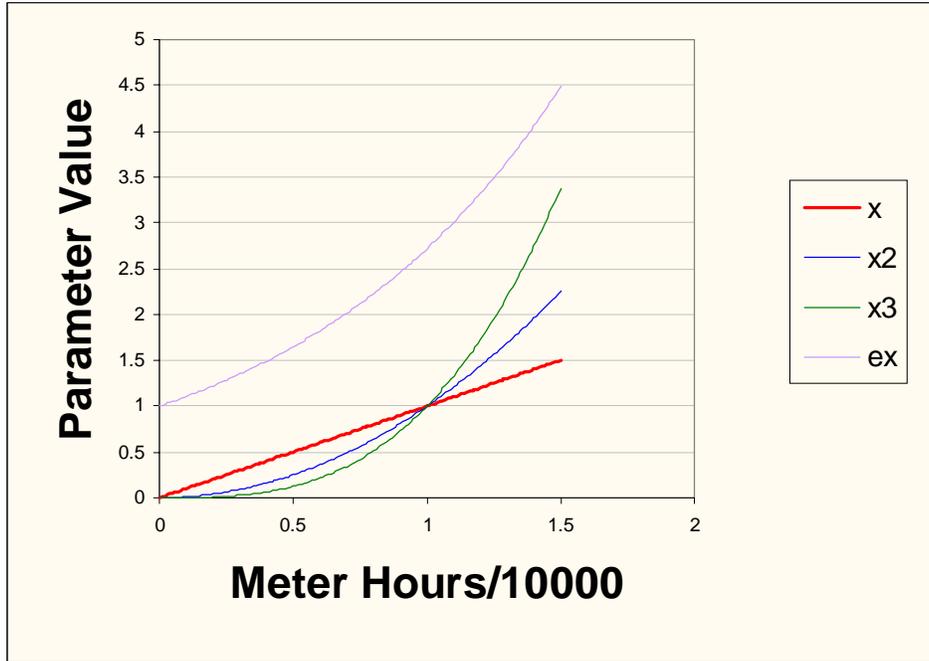
Figure 5-4: Regressor Values--Raw Data/10,000
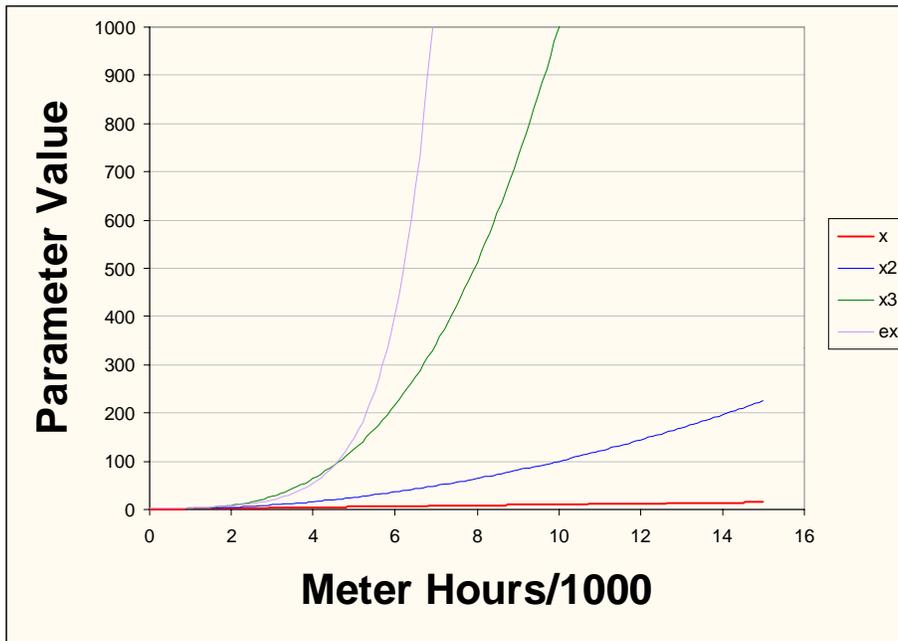
Figure 5-5: Regressor Values--Raw Data/1000

For the three figures that will be discussed, the x-axis covers a range from zero to 16,000 cumulative meter hours. This range is reasonable to look at—most of the machines in this study fall within this range. The y-axis scale has been varied to better focus on what the curves are doing. From Figure 5-3 it can be seen that the values of e^x and x^3 climb very steeply when the raw values of meter hours are used. The e^x curve is so steep that it appears as if it is climbing directly vertical. The x line is so shallow in relation to the other curves that it appears almost horizontal. The relationships between the curves are so greatly varied that they are hard to visualize.

Figure 5-4 depicts what happens to the curves if cumulative meter hours divided by 10,000 is used as the regressor. The relationships between the four curves are well defined. It can be seen that x^2, x^3, and e^x all increase monotonically. Regressions that are run using meter hours divided by 10,000 do not cause the computer to have errors. There is a different problem associated with this amount of scaling, though. From the figure, it can be seen that the nature of the relationship between x, x^2, and x^3 changes dramatically at a value of "1". The change in this relationship is important because it is expected that the optimum life of most machines will fall between the 5,000 to 15,000 cumulative meter hour range. If the nature of the equation changes dramatically in the middle of the range of interest, problems could arise.

Figure 5-5 is a depiction of what happens to the parameters when cumulative meter hours/1000 is used. The relationships between the curves are still reasonably well defined. The problem area depicted in

Figure 5-4 is shifted to an earlier point in machine life. It is not expected that many construction machines will reach their optimum lifespan prior to 1,000 hours of use. The crossover relationship between x^3 and e^x at around 4,500 hours is acknowledged. This crossover is not expected to have as great an impact on the usability of the equations because, as mentioned above, it is expected that most optimum lifespans will fall in the range of 5,000 to 15,000 hours of use.

Cumulative meter hours divided by 1000 was chosen as the best solution to the data scaling problem. The computer programs used did not generate error messages when these data were input. There are changing relationship between the variables, but they occur outside of the range of values that are of primary interest.

5.2.2   Data Splitting

To validate the predictive capabilities of the models developed, a cross-validation procedure will be used when possible. This will require that the data be split into estimation and prediction data sets. Ideally, these two data sets should be of equal size. To split the data into two equal sized sets, the number of observations (machines) "n" should be (Snee, 1977):

$n \geq 2p + 25$ **Equation 5-5**

In this inequality, "p" is the number of parameters in the model. For the data to be analyzed in this study, the maximum value of "p" will be four (x, x^2, x^3, and e^x). To have equal sized prediction and validation data sets there should be a minimum of 34 observations in the full data set. Few of the full data sets to be analyzed will have this many observations. For the data sets that do not have 34 machines, the estimation data set will consist of 17 machines. This is the size the estimation data set would have been if 34 observations had been present. It is important that there be at least 17 machines in the estimation data so there are sufficient degrees of freedom remaining in the regression. The prediction data set will consist of the remaining machines.

The machines to be split off to the prediction data set will be selected using the following process:

1. Random numbers between 0 and 100 will be generated by computer and assigned to each machine.

2. The machines will be rank ordered by their random numbers.

3. The first 17 machines (or one-half the number of machines, whichever is greater) will form the estimation data set.

4. The remaining machines will form the prediction data set.

Once the prediction data set has been formed, a scatterplot of its data pairs will be compared to one of the estimation data set. Although it is important that assignment to the prediction data set be random, it is perhaps equally important that the prediction data set provide an appropriate test of the model developed. With random assignment, it is possible (albeit improbable) for the prediction and estimation data sets to lie at completely opposite ends of the spectrum of cumulative hours of use. In this case, the test on the predictive capabilities of the model would lie in a region that is based entirely on extrapolation. This is not desirable. To have a good cross-validation, the prediction and estimation data sets should be similar, but different (Birch, 1996). The estimation data set should include points that cover the full range of cumulative hours of use. The prediction data set should cover this range as well—this is the "similar" part. The prediction

data set should contain points that appropriately "stress" the model developed by the estimation set. This is the "different" part. If the scatterplots show that the two data sets are not similar but different, the splitting process will be repeated until a favorable data split is achieved. This adds a bit of subjectivity to the process, but this subjectivity is necessary to ensure a cross-validation that has meaning.

Intuitively, the prediction data set becomes a less significant test of predictive capabilities as it gets smaller. If there are less than five machines (approximately 20% of the data), the prediction data set is probably too small. In this case, a measure of predictive capabilities called the PRESS residuals will be used for validation—this will be described in Section 5.3.3.

In summary, if the number of machines in the data set is greater than 22, the data set will be split into an estimation data set and a prediction data set for cross-validation of the model. If there are less than 22 machines in the data set, a different procedure using the PRESS statistic will be used to validate the model and data splitting is not necessary.

5.2.3   Variance Characterization

A variance characterization will be accomplished on each equipment fleet to be analyzed. This characterization will be done using data pairs from each machine in the fleet interpolated to 500 hour intervals. This is the same set of data pairs described in Chapter 4 that was used to formulate the set of average data pairs that will be one of our four regression options.

Something that is expected of all the data sets is the presence of heterogeneous variance. The reason for this is that heavy equipment tends not to break down that often during the early stages of economic life. As economic life progresses machines that are subjected to harsher operating conditions, sub-par operators, etc. should tend to have higher CCIs than similar machines that are well-taken care of. The spread between the CCI of a very good machine and the CCI of a very poor machine should become greater as hours on those machines accumulate.

Normal regression assumes a homogeneous variance of the response variable throughout the range of regressor variables. If the variance is heterogeneous, adjustments to the model may be necessary. These adjustments can be accounted for by using *weighted regression.*

To decide whether or not the variance is homogeneous, the hourly data set will be processed using the PROC MEANS procedure of SAS. This procedure will yield sample means and variances for the response variable with respect to the regressor variable. Another data set will then be formed pairing these sample variances with their respective values of regressor variables. Sample variances that are based on less than nine observations should be eliminated from the data set (Myers, 1990). *If the majority of the sample variances are based on fewer than nine observations, weighted regression will not be used.* According to Myers, using weights that are not correct can be worse than not weighting at all.

Once the new data set composed of variances and meter hour values is constructed, it will be analyzed using the PROC REG procedure of SAS. This procedure will perform a simple linear regression of the variance versus the regressor. The model will be of the form:

$$y = \beta_0 + \beta_1 x$$ **Equation 5-6**

Where $y = \sigma^2$.

The following hypothesis test will then be performed:

$$H_0: \ \beta_1 = 0, \ H_1: \text{not } H_0$$

This test will be performed at a p-value of 0.05. If H_0 is accepted, no adjustments for variance are required. If H_0 is rejected, the variance is not homogeneous and adjustments may be made to the analysis to account for this.

The goal of weighted regression is to minimize the effects of heterogeneous variance. In normal least squares regression, the "squares" that we are trying to make "least" are given by:

$$SS_{RESIDUAL} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$ **Equation 5-7**

Where:

> $SS_{residual} = $ the residual sum of squares

y_i = observed value of CCI

$\hat{y}_i$ = predicted value of CCI

In weighted regression, this equation becomes:

$$SS_{RESIDUAL} = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$ **Equation 5-8**

Where:

$$w_i = \frac{1}{\sigma_i^2}$$ **Equation 5-9**

This gives greater importance to the data points associated with lower variance and less importance to those of high variance. In practical terms, this means that the points that are most subject to the least variance are receiving greater "weight" in the regression.

To accomplish weighting in SAS, we must first define σ^2 in terms of the regressor variable. This will be done by substituting the function of x given by equation 5-6 into equation 5-9. The weights are thus defined as a function of cumulative meter hours. Regression is then accomplished by adding the WEIGHT statement to PROC REG in SAS.

Although none of the data sets analyzed were large enough to reliably use weighted regression, a sensitivity analysis of the results to weighting will be performed and described in Chapter 7.

5.3   ANALYSES

This section will explain the specific statistical analyses to be performed on the data. The quest will be not only for an adequate model, but also for a parsimonious model (Tukey, 1961). With so many models under consideration, more than one will probably yield an adequate solution. The differences in performance between some models may be negligible. In cases like these, it is sometimes wisest to choose the simplest model—the one with the fewest terms that describes what is happening to the researcher's satisfaction. Extra terms should not be included in the

model for small gains in performance. Errors in the value of "x" will be compounded in a more complex models since each of the parameters is a function of "x". Because of this compounding of errors, a simple model that adequately fits the data may be a better choice than a more complex model that is marginally better.

The steps in analyzing the data will be:

1. Preliminary analysis—will provide general idea of which models are best

2. Intermediate analysis—will pick the best models from the group using a more detailed study

3. Final analysis (if necessary)—selects the one best model and data set for this study

4. Model validation—the performance of the model will be judged

5. Influential points—the data will be examined to ensure that no one machine has had an undue influence on the regression

6. Comparisons—general comparisons between companies, types of machines, and sizes of machines will be accomplished

5.3.1   Preliminary Analysis

The preliminary analysis will be performed on all models to be entertained for each fleet of equipment in the study. Since there are so many models involved (17 fleets, 3 data sets for each fleet, 19 models for each data set, totaling 969 regression models), a filtering process is necessary to bring the number of models considered down to a reasonable number. Models that obviously provide either poor fit or poor predictive capabilities will be eliminated from consideration at this level.

A macro developed in SAS Interactive Matrix Language (IML) will be used to perform filtering at this level. The NOINT macro (Noble, 1997), was modified to fit the purposes of this study. This macro and its modifications can be found in Appendix B. This macro will be used to obtain

measures of effectiveness on the first 15 models (all requiring regression through the origin). The output of this model provides parameter estimates for each model considered and rank-orders each model by each of five measures of performance. These five measures of performance are R^2, adjusted R^2, Mean Square Error, C_p, and R^2_{PRESS}.

Not all five of the measures of performance will be used in the model selection process. The measures of performance to be evaluated are adjusted R^2 and R^2_{PRESS}. Raw R^2 will not be used because it assesses no penalty for models that have more regressors than others do. By nature, regressions that have additional regressors tend to explain more of the response behavior than those that have fewer regressors do. Mean square error will not be used because it is inversely proportional to adjusted R^2, the two measures provide essentially the same information except adjusted R^2 is somewhat standardized and can be used to compare different models more easily. C_p will be looked at for the linear models, but will not serve as a primary determinant of model performance. This is because the use of the C_p statistic requires an accurate estimate of model variance to provide a truly meaningful measure of performance. Additionally, the C_p statistic cannot be used to compare the non-linear models to the linear models or vice-versa. Detailed explanations of these measures of performance can be found in Myers (1990).

Adjusted R^2 provides a measure of model fit, i.e. how well does the curve fit the data pairs. The higher the value is, the better the fit is. R^2_{PRESS} provides a quick measure of model predictive capabilities. The higher the value is, the better the predictive capabilities are. Both of these statistics are somewhat standardized—they can be used to assess performance differences between different models.

To compare the many types of data sets and many types of models, a non-parametric technique called the Kruskall-Wallis test will be used. A good explanation of this test can be found in Ott (1993, 792-795). Non-parametric tests were chosen because at this level a parametric test would not have as much meaning. The relative rankings of the various models are more important than the actual values of their measures of performance.

First, an assessment of which of the four data set types provide the best measures of performance will be made. Relative differences will be noted. Following this, preliminary assessments will be

performed on the linear and non-linear models. The best model (or models) in terms of performance balanced with the parsimony principle will be selected for each group (linear and non-linear). The statistics from the best linear and non-linear models will be combined into one smaller group. Non-parametric tests will be performed on this filtered grouping to determine if there is any significant difference in the performance of the models. The best model(s) will make the transition to the intermediate analysis.

5.3.2   Intermediate and Final Analyses

The intermediate analysis will be performed using the PROC REG procedure of SAS. The specific codes employed are given in Appendix C. The purpose of the intermediate analysis is to further filter the list of models obtained in the preliminary analysis. The intermediate analysis may or may not determine a clear-cut winner. If no obvious winner is apparent, the model that best predicts realistic equipment lives will be the one chosen in the final analysis (this process will be described in Chapter 7).

The most important insight that is gained using the intermediate analysis over the preliminary analysis is the significance of model parameters. Although a model that contains all possible regression terms may provide the best fit and predictive capabilities, the parameters themselves may not contribute greatly to the characterization of the response variable. Significance of the model parameters will be ascertained using the following decision criteria:

$$\text{p-value} \leq 0.20\text{--acceptable} ; \quad \text{p-value} > 0.20\text{--unacceptable}$$

This test will be performed on all parameters of the models that make it to this stage of the proceedings. Failures to meet the decision criteria will be noted. Models that consistently have parameters that are not significant will be eliminated.

Additionally, the residual plots will be analyzed. This will give a good feel as to the nature of the variance of the errors. Ideally, the errors will be normally distributed about a mean value of zero. If there is any non-homogeneity of variance, it may be recognizable from the residual plots.

Confidence intervals will be analyzed for the coefficients of the regression. Assessments will be made as to what range of L* and T* values could be expected from a given model.

It must be stressed that model selection is an art as well as a science. Quantitative measures of performance are not always the sole determinant of which model is the best. Qualitative measures, such as parsimony, play as great a role in selecting a model that will give desirable performance characteristics.

5.3.3   Model Validation

Two types of model validations will be employed in this study. The validation technique employed depends upon the number of machines in the data set to be analyzed.

CROSS-VALIDATION: For data sets that had 22 or more observations, data splitting and cross-validation are used to determine the prediction capabilities of the model. The way this will be done is as follows:

1. Using the procedures outlined above, determine the regression model using PROC REG in SAS using only the estimation data set.

2. Insert the "x" values for the prediction data set into the estimation data set. Instead of entering the actual "y" values, enter "." instead. This tells SAS to predict "y" values for the prediction data set, but not to use the "x" values to alter the equation. Run PROC REG again to get the predicted y values for the prediction data set.

3. Using these values, compute the correlation coefficient between the actual "y" values and the predicted "y" values. Use the following equation:

$$r(y, \hat{y}) = \frac{S_{y\hat{y}}}{\sqrt{S_{yy}} * \sqrt{S_{\hat{y}\hat{y}}}}$$ **Equation 5-10**

4. The population correlation is ρ. The number of machines in the prediction data set is n_2. Test the hypothesis:

$H_0: \rho = 0$ vs. $H_1: \rho > 0$

using a p-value of .20. Use the test statistic:

$$t = r\sqrt{\frac{n_2 - 2}{1 - r^2}} \Rightarrow t_{n_2 - 2}$$ **Equation 5-11**

5. If H_0 is rejected, the cross-validation was successful. The model should be regenerated using *all* of machines after a successful cross-validation. This adds more samples to the model that is developed and, theoretically, the model should be a truer representation of the population. If the cross validation was not successful, try fitting the model with fewer regressors and repeat the procedure.

PRESS validation: Validation using the PRESS statistic is more subjective than the analytical test described above for cross-validation. If the PRESS procedure is used, there were not enough machines in the data set to warrant data splitting. If the R^2_{press} of the model being evaluated compares favorably with the R^2_{press} of a reference fleet that underwent a successful cross-validation, the inference will be made that the models have similar predictive capabilities. If the R^2_{press} of the model in question is worse than that of the reference fleet, its predictive capabilities are could be worse than that of the reference fleet. If the R^2_{press} is better, the predictive capabilities may be better than those of the reference fleet.

5.3.4   Influential Points

Identifying influential observations is a science unto itself. There are many different statistics that can be compared to get a picture of which data points are outliers, which are points of high leverage, and which are points of high influence. The SAS computer program will generate these statistics with the code indicated in Appedix C. These statistics include $R_{student}$, h_{ii} (hat diagonals), DFFITS, and DFBETAS.

All of these statistics will be generated and evaluated. $R_{student}$ is a good measure of whether or not a point is an outlier. The h_{ii}'s can be used to determine if a point has high leverage.   DFFITS and DFBETAS statistics provide measures of a data point's impact on the fit and coefficients of the

model. The "DF" prefix stands for difference. Each of these statistics measures the difference that results if the point of observation is removed from the data set. This is similar to the logic used in calculating the PRESS statistic in that they are all single point elimination schemes. The "S" suffix stands for standardized. The difference between what would happen with and without the point of interest is divided by the appropriate standard error to yield a standardized statistic. DFFITS will identify those points that have a large influence on the "FIT" of the regression. DFBETAS will identify those points that have a large influence on the coefficients (Betas) of the regression equation.

For all of these statistics rules-of-thumb exist for determining what is significant. These rules should be used in context. In other words, it is important to look at relative differences in these statistics. The question "Do these observations stick out as odd when the group is viewed as whole?" must be answered. Rules-of-thumb are not substitutes for good judgement.

If a point is indeed suspected of causing undue influence on the regression, that point will be examined in detail. This examination will involve taking a closer look at the machine in question. Was it purchased used? Was it involved in an accident? Was it abused? Data points will not be eliminated to develop tidy equations. They will be investigated fully before any decision to disregard them is made.

An interesting thing about regression through the origin is that, by definition, the data points that are located the furthest away from the origin will have greater impact on the regression than those that are located very close to the origin. In practical terms this means that some of the points that are identified as highly influential are so because of the design of the experiment, not because the machine in question is a lemon.

5.3.5   Comparisons

Comparisons will be performed to ascertain if each or any of the following factors have a discernable impact on the end performance and results of the regression equations:

- type of equipment

- size of equipment

- company operating the equipment

Performance comparisons will be drawn using the non-parametric techniques described in Section 5.3.1. The measures of performance that will be evaluated are adjusted R^2, and R^2_{PRESS}. Generalized conclusions concerning the performance of regression models concerning the above three areas will be drawn. These performance assessments will be of the form: "It seems that all but one of the companies involved in the study have regression equations that yield acceptable and similar measures of performance."

The second type of comparison, that of parameter values, will be a little more involved. The mechanism for doing this will be the cross-validation procedure described in Section 5.3.3. The following is a list of cross-validations that will be performed:

1. Machines of the same class and group that are of different companies will be used as prediction data sets for each other's equations.

2. Machines of the same class but of different groups within the same company will be used as prediction data sets for each other's equations.

3. Machines of unlike class and group within the same company will be used as prediction data sets for each other's equations.

4. All machines of each particular class and group will be recombined to form large data sets. These data sets will be split using the procedures described in Section 5.2.2. The intermediate analysis will be performed on these equations to determine measures of performance after which the cross-validation procedures will be performed.

5. All machines from each company will be recombined to form large data sets. These will be subjected to the same analyses as those in step 4.

The general conclusions drawn from these comparisons will be of the type: "It seems that company of ownership may be a factor when deriving regression equations to characterize the growth of maintenance and repair costs."

5.4   SUMMARY

This chapter has presented and in-depth discussion of the statistical analyses to be performed. The mechanics of these analyses were briefly presented, some more manipulations of the data were discussed, and the actual analyses that will be accomplished were documented. Combined with the material presented in Chapter 4, this concludes *Defining the Work*. The stage is now set for the analyses to take place.

Part III, *The Work*, is comprised of Chapters 6, 7, and 8. Chapter 6 will discuss how the data were prepared for analysis. Chapter 7 will describe the model selection process. Chapter 8 will present the results of regresssions and comparisons. It will also compare the data-based cumulative cost equations to other forecasting methods that were described in Chapter 2.

CHAPTER 6: DATA PREPARATION

This chapter chronicles the gathering and preparation of the data used to support this dissertation. Specific characteristics of the data were provided in Chapter 4. The recommendations from that chapter are implemented here. This chapter is the first in Part III of the dissertation—"The Work". A basic flowchart of the work to be accomplished is provided in Figure 6-1.
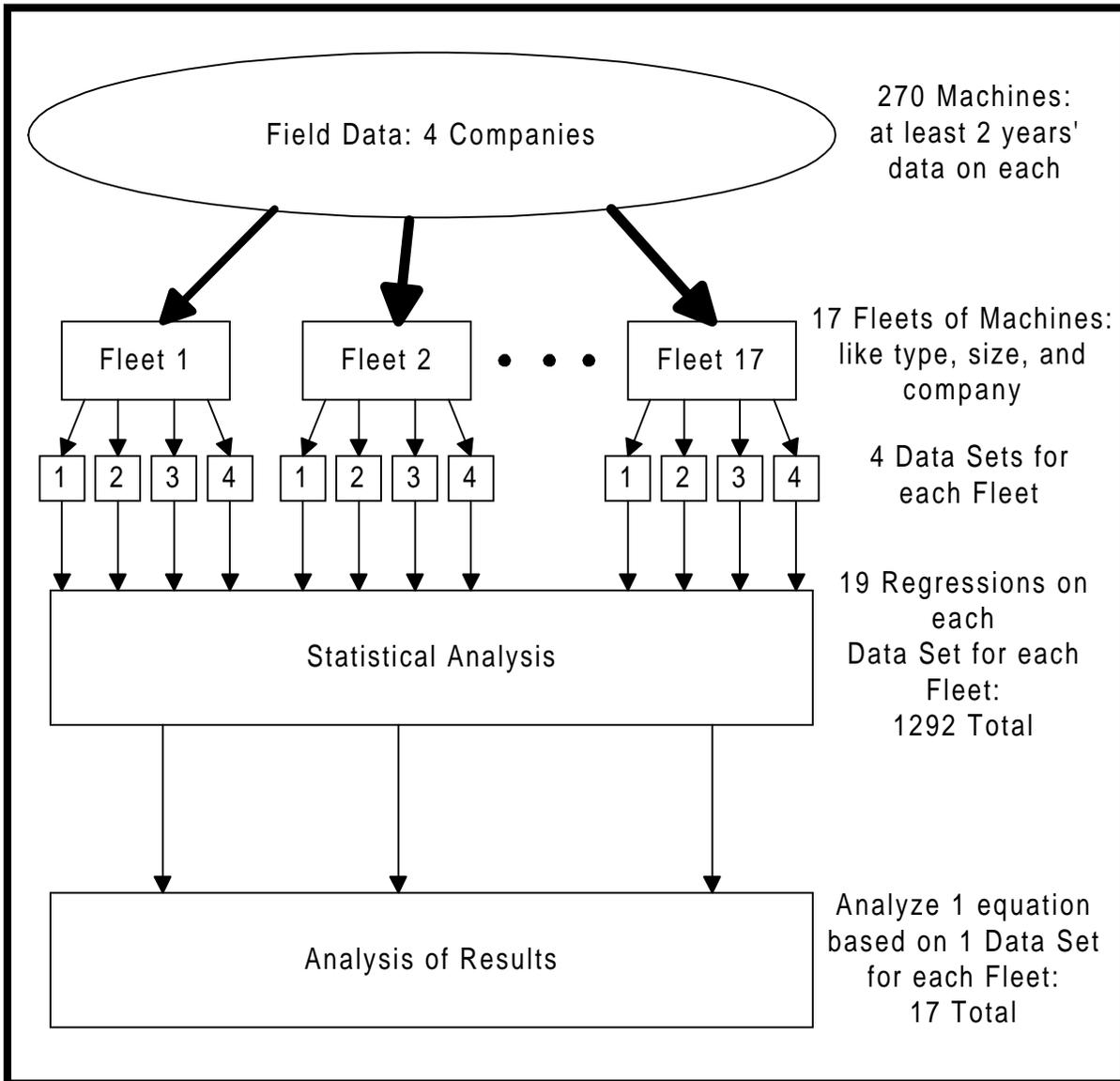
Figure 6-1: Part III Flowchart

This chapter explains how a multitude of data on 270 machines from four companies was processed and filtered to form 68 data sets—four for each of seventeen fleets that were of like type, size, and company. This data is then input into the statistical analysis (Chapter 7) which flows into the Analysis of Results (Chapter 8).

The following issues will be discussed in this chapter:

- Data extraction

- Manual corrections

- Inflation database

- Oil sampling databases

- Creation of SAS® data sets

- Final product

A flowchart representing the data preparation process is given in Figure 6-2. It may be helpful to refer back to this chart throughout the chapter to understand the context of each step of the preparation process.
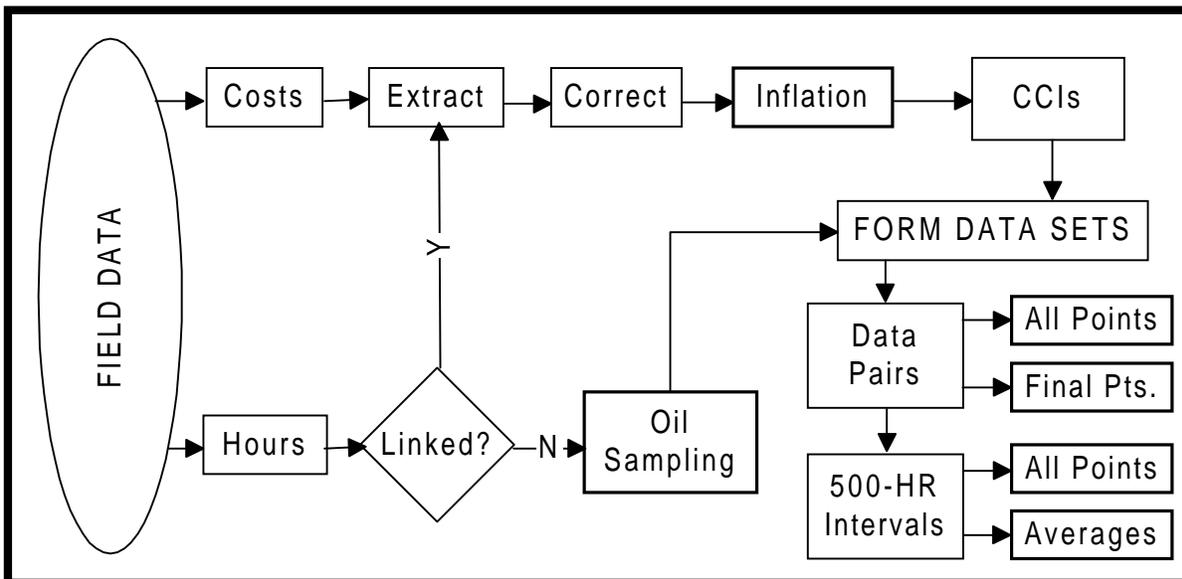
Figure 6-2: Data Preparation Flowchart

6.1   DATA EXTRACTION

Before any formatting could be accomplished, the data had to be obtained. Each company provided their data in a different format. The three main formats of data obtained were:

- PC-formatted

- Mainframe formatted

- Manually obtained

The *PC-formatted* data were by far the easiest to manipulate and assimilate. These were files that came in a format that could be opened and used directly by standard PC-based data manipulation programs. Some of data came in spreadsheet format (Excel or Quattro). These required the least amount of work because they were already in the format needed for preliminary manipulations. Other data came in database format (Access or Paradox). These data were extracted from the database files into Excel format. In some cases, this proved to be less total preliminary work than that required for the data that came in spreadsheet format because queries could be generated to extract the data in exactly the spreadsheet format desired.

The *Mainframe formatted* data posed a different challenge. Not all companies involved in this study do their data manipulation on PCs. They run all reports and generate all printouts from their mainframe computers using applications programmed specifically for their company. The way the mainframe data were transferred to the PC was through the generation of ASCII print files (.prn)—the mainframe computer was told to print its reports to files instead of printers. These files could then be opened on a PC—but the data were not *parsed*. They could not be directly used by spreadsheet programs—all of the data on a given line would be imported to one column of a spreadsheet. Although there are some converters available that allow for the parsing of ASCII data within the spreadsheets, the data must have originally been in a neat tabular format for these converters to work. This was not always the case.

Table 6-1: Example Raw Equipment Data

EQUIP	COST CODE	PAY ITEM	COST TYPE	Beg DATE		END DATE	# PERIOD	G/L AMOUNT
9273 -		DOZER	KOM		(Continued)			
	42 ENGINE		RELATED	REPAR	(Continued)			
	2 INPUT		W/E	05/04/96 TO		06/01/96	PR00-0000	22.69
	2 INPUT		W/E	05/04/96 TO		06/01/96	PR00-0000	16.08
	2 INPUT		W/E	05/04/96 TO		06/01/96	PR00-0000	16.84
	2 INPUT		W/E	09/07/96 TO		09/28/96	PR00-0000	2.43
	2 INPUT		W/E	09/07/96 TO		09/28/96	PR00-0000	1
	2 INPUT		W/E	09/07/96 TO		09/28/96	PR00-0000	2.84
	2 INPUT		W/E	10/05/96 TO		11/02/96	PR00-0000	2
	2 INPUT		W/E	10/05/96 TO		11/02/96	PR00-0000	4.86
	2 INPUT		W/E	10/05/96 TO		11/02/96	PR00-0000	5.72
	3	44472				#######	35490	
	TOTAL	FOR	CODE		42 ENGINE	RELATED	772.53	
	43 HYDRAULICS							
	1	020197-FIELD	PR15	PR15-0237	35431	#######		
	1 INPUT		W/E	01/06/96 TO		02/03/96	PR00-0000	16.8
	2	020197-FIELD	PR15	PR15-0238	35431	#######		
	2	020197-FIELD	PR15	PR15-0239	35431	#######		
	2	020197-FIELD	PR15	PR15-0240	35431	#######		
	2 INPUT		W/E	01/06/96 TO		02/03/96	PR00-0000	3.83
	2 INPUT		W/E	01/06/96 TO		02/03/96	PR00-0000	2.38
	2 INPUT		W/E	01/06/96 TO		02/03/96	PR00-0000	1.72
	TOTAL	FOR	CODE	43	HYDRAULIC		104.74	

As can be seen from Table 6-1, the data were not necessarily in a format that was easily collated. In this example, two costs codes for a dozer are shown: code 42 (engine) and code 43 (hydraulics) Each line in the report does not have the same format. At evenly spaced intervals, the headers are shown (normally the page breaks for printed output).   Following that, the equipment number and type are shown. After that, all applicable cost codes associated with that piece of equipment during the time frame of the report are shown. The cost codes are further broken down into line items. Some of the line items contain cost data, some do not. A range of dates for the expenses are shown instead of specific dates for the expenses. The entry after the

last line item is a subtotal for that cost code. A grand total for all costs codes is listed at the end of each piece's portion of the report. The subtotals and grand totals were not useful because they could not be attributed to a specific month.

The reports also contained cost data that was not a part of this study. Usually, more than one report had to be used to get all of the data on any one machine. The process of extracting the useful data was greatly simplified through the use of a computer program by the DataWatch® Corporation called Monarch®. Templates can be built in Monarch® to recognize line formats. The cost data can be extracted, associated with a specific machine, and subtotaled by cost code for each month. The unwanted cost data can then be filtered out by cost code and an Excel® spreadsheet containing only the desired data can be exported.

The final format of data used for this research was that which was *manually obtained*. These data were the hardest to get. In some cases, folders containing records and receipts for each machine were gone through one page at a time. These data took the longest to extract. The data were recorded by hand and then entered into an Excel® spreadsheet at a later date. Although great care was taken to ensure numbers and dates were accurately recorded and entered, the potential exists that a small portion of the data could have been missed in the scans of the folders, incorrectly written down, or incorrectly entered into the spreadsheets. Although there could also have been transcription errors in the electronically obtained data, the chances of an error in the chain are increased with manually obtained data. Manually obtained data were the least desirable, but in some cases were the only ones available.

6.2 MANUAL CORRECTIONS

After getting all the data into spreadsheet format, the data that was important to the study had to be filtered and collated. Cost accounts relating to maintenance and repair (including capitalized rebuilds) were extracted and summed for each month. In some cases, the output from the company was in terms of cumulative costs, in some cases it was in terms of incremental monthly costs. The same held true for the hours of use data. Some were in the format of cumulative hours; some were in the format of incremental hours. Although cumulative cost and cumulative hours of use are the data needed for generating the data sets, at this stage the incremental costs

and hours were more important. Data that were in their cumulative forms were converted to their incremental forms for this portion of the preparation.

As mentioned in Chapter 4, there were a number of problems with the data that had to be recognized and corrected by hand. The easiest way to do this was by forming two matrices for each of the fleets analyzed. One matrix consisted of machine numbers, incremental cost, and dates (see Table 6-2). The other matrix was composed of machine numbers, incremental hours, and dates (see Table 6-3).

Table 6-2: Incremental Costs

Equip. #	Month						
	Feb	Mar	Apr	May	Jun	Jul	Aug
00102	49.00	30.00	111.00	10120.00	55.00	122.00	153.00
00207	164.00	98.00	120.00	149.00	128.00	108.00	130.00
00208	45.00	116.00	(110.00)	143.00	191.00	5438.00	20.00
00210	52.00	(25.00)	92.00	24.00	46.00	122.00	114.00
00213	156.00	220.00	88.00	93.00	(150.00)	36.00	92.00
00303	0.00	0.00	32.00	72.00	84.00	148.00	164.00
00321	0.00	0.00	0.00	12.00	161.00	116.00	125.00
00342	0.00	0.00	0.00	0.00	0.00	104.00	165.00

For both costs and hours, there were two important scans to do. The first, and most obvious was to scan for "red". The spreadsheet was configured so negative numbers would show up in red. Negative numbers for either incremental cost or incremental hours should not be possible. As mentioned in Chapter 4, the negative numbers for cost usually signified an improper charge had been placed against the machine in a previous month. The negative number is the company's way of correcting that improper charge. In these cases, the correction described in Chapter 4 (get rid of the negative number and reduce the expense for the prior month or months) was applied. Negative numbers for hours did not occur as often. All occurrences of negative hours highlighted a meter change for the given machine. When a new meter is installed, the cumulative hours on that meter are zero. When the incremental hours are calculated using existing cumulative hours, a negative number can result (some companies account for this in their databases—others do not).

Table 6-3: Incremental Hours

Equip. #	Month						
	Feb	Mar	Apr	May	Jun	Jul	Aug
00102	49	30	111	121	55	122	153
00207	164	98	120	1200	128	108	130
00208	45	116	110	143	191	14	20
00210	52	77	92	24	46	(6149)	114
00213	156	220	88	93	44	36	92
00303	0	0	32	72	84	122	164
00321	0	0	0	12	161	116	125
00342	0	0	0	0	0	104	165

For the meter changes, if it could not be determined how many hours the machine had actually worked the month of the meter change, the billable hours were used for that month.

The second scan that had to be done on the table was for unusually large numbers. When exceptionally large repair costs were encountered, the reason for the large number was investigated. For some companies, this was simply a matter of looking at the data in its raw form to see where the costs came from. For companies that provided raw combined costs instead of costs by individual cost code, the equipment managers or equipment receipt files had to be consulted to determine why the large expense incurred. If the large expense was not related to maintenance and repair or if it was just a mistake, it was subtracted from the incremental costs for that month. This is the also the way in which corrections for wrecked or abused machines were made.

6.3   INFLATION DATABASE

The database Microsoft Access® was used to apply the inflation correction factors discussed and detailed in Appendix A. The data were put in a standardized format so that the data from all of the companies could reside and be manipulated within the same database. The database made the process of associating inflation indices for given months with the costs for those months easier than the process would have been had it been done exclusively in a spreadsheet. Some of the formatting had to be consistent across all tables in order for the database to work.

The date fields had to be standardized among the companies. The format selected was: yy-mm where yy consists of the final two digits of the year and mm consists of the month's number (1-12). A second data field that needed to be standardized was that of equipment number. This was more of a table check than a data manipulation issue. It had to be ensured that the formatting of this data field was consistent for the all company-specific tables that contained it. For instance, if the field was labeled as text field with 8 characters in on table, it could not be labeled a double precision numeric field with 6 characters in another table.

Five tables were used for the companies that collected cumulative hours for each month; four tables were used for those companies that did not collect cumulative hours on a monthly basis. Two of these tables were common to all companies. These were the tables of inflation indices, their format is depicted in Table 6-4. Table 6-4a depicts the indices used for adjusting the purchase price of equipment, Table 6-4b shows the indices used for adjusting repair costs. The method for obtaining the indices is described in Appendix A.

Table 6-4: Cost Indices

a   **b**

PurchaseDate	PIndex
87-01	1
87-02	1.000818
87-03	1.00491
87-04	0.995908
…	…
97-09	1.344517
97-10	1.350245

Month	Rindex
87-01	1
87-02	1.002069
87-03	1.006343
87-04	1.00382
…	…
97-09	1.391823
97-10	1.396592

There were also three possible company-specific tables. The first of these tables was used for every company. It contained data that was specific to each machine that did not change with the passage of time or with usage. There was one line entry for each machine in the company that

was to be analyzed. An example of this table is depicted in Table 6-5. The data included in this table are the equipment number, type, class, purchase month, and list price.

Table 6-5: Equipment Static Data

EQNUM	EQNAME	EQCLASS	PMONTH	PP
225	Artic, Volvo A35	ar	93-06	402000
226	Artic, Volvo A35	ar	93-06	402000
307	- DOZER, CAT D-6HXL	d1	94-06	197420
308	- DOZER, D5H XL	d1	92-06	177860
324	- DOZER, CAT D-6H	d1	87-06	165790
356	- DOZER, CAT D-6H	d1	87-01	165790
728	- DOZER, CAT D-8N	d3	89-06	311550
746	- DOZER, CAT D-8	d3	93-06	353490
747	- DOZER, CAT D-8N W/ RIPP	d3	93-06	353490
802	- DOZER, CAT D-8N W/RIPPE	d3	94-06	363930

The next two tables were fairly similar to each other. They are depicted in Table 6-6. Each table contains one line for each machine for each month that data are available. The first of the two tables contains the equipment number, the month, and the incremental monthly cost. For the first month's data on each machine the cumulative monthly cost is used instead of the incremental monthly cost. For example, the $154,916.40 monthly expenditure for machine number 225 in December of 1995 is the cumulative cost of maintenance and repairs up to that point in the machine's life. The $18,174.02 that is shown for the same machine in January of 1996 is the monthly cost for that month. The hour data that are depicted in Table 6-6b are cumulative hour data. The cost and hour data were kept separate at this juncture because they were already in a separated format (Section 6.2) for manual corrections and it was easier to import them into the database separately.   Not all companies had tables that reflected the hours of use—the incorporation of their hourly data will be discussed in Section 6.4.

Table 6-6: Cost and Hour Tables

a   b

EQNUM	MONTH	MCOST
225	95-12	154916.4
225	96-01	18174.02
225	96-02	602
…	…	…
225	97-03	1925.11
225	97-04	864.06
225	97-05	463.68
226	95-12	19453.15
226	96-01	58.01

EQNUM	MONTH	HOURS
225	95-12	18339.25
225	96-01	18471.25
225	96-02	18631.25
…	…	…
225	97-03	21080.75
225	97-04	21104.75
225	97-05	21271.75
226	95-12	5216
226	96-01	5352.5

Once the data have been arranged into tables and imported into the database, the tables can be linked as a query to yield an output of the type desired. Sample output is shown in Table 6-7.

Table 6-7: Output from Inflation Database

EQNUM	EQCLASS	HOURS	PIndex	Rindex	Mindex	PP/Pindex	MCost/Rindex	MCost/Mindex
23789	aw	16654	1.023732	1.337708	1.18072	110245.7	62248.31	70524.83
23789	aw	16732.5	1.023732	1.346301	1.185016	110245.7	840.9633	955.4213
23789	aw	16732.5	1.023732	1.351025	1.187378	110245.7	2984.208	3395.497
23789	aw	16848	1.023732	1.354263	1.188997	110245.7	1104.867	1258.439
23789	aw	17008	1.023732	1.357591	1.190661	110245.7	926.6489	1056.564
23789	aw	17080	1.023732	1.359705	1.191718	110245.7	1376.401	1570.421
23789	aw	17208	1.023732	1.36101	1.192371	110245.7	662.3685	756.0484
23789	aw	17344	1.023732	1.363574	1.193653	110245.7	2206.319	2520.398
23789	aw	17494.5	1.023732	1.365238	1.194485	110245.7	7453.022	8518.442
23789	aw	17704.5	1.023732	1.367712	1.195722	110245.7	1025.326	1172.806
23789	aw	17905.5	1.023732	1.371085	1.197408	110245.7	1623.05	1858.464

The query was run for each company. Most of the fields in Table 6-7 have already been discussed. An important exception is Mindex. Mindex was used in the formulation of the first data pair for each machine. It is the average inflation index for the period of time from machine purchase to the calendar point in time for which the first cumulative repair cost data are available. For some machines, these two points are essentially the same and the Mindex is not needed.

Three calculated fields were: PP/Pindex, RC/Rindex, and RC/Mindex. These calculated fields were the inflation-adjusted cost data used in computing the CCI. This query was run for each company as a whole and output was directed to the spreadsheet. The four table query for companies that did not have cumulative hourly data easily available was similar to this five table query with the exception of the omission of cumulative meter hours. The way in which these meter hours were obtained is discussed in Section 6.4.

6.4   OIL SAMPLING DATABASES

Three of the four companies involved in the study did not explicitly keep track of the cumulative meter hours on their machines. These companies did participate in periodic oil sampling programs, however. Data from the oil sampling databases and equipment receipt files provided the date linkage between cumulative hours of use and cumulative costs. The data obtained from equipment receipt files were taken from preventive maintenance reports and oil sample analysis printouts that were in the machine's individual files. Additional points were also available if there were any repair work orders that gave meter hour readings along with the date the repair was performed. The data obtained directly from oil sampling databases had to be processed and filtered using Monarch®. Before Monarch could be used, the coding process used by the analysis facility that produced the data had to be understood. Raw oil sampling data are depicted in Table 6-8.

The oil sampling reports consisted of strings of data. The vehicle identification portions of the data strings depicted in Table 6-8 have been omitted to save space. Area "A" of the table shows the portion of the data string that signifies the date the sample was taken. The string "951206" signifies that the sample was taken on 6 December, 1995. Area "B" shows the region of the data string that contains the cumulative meter hours for the piece of equipment at the time of the

sampling. In this case, the machine had 2124 hours at the time of the sample. The purpose of this example was to show how difficult it is to extract the data from this file if the exact location of the data is unknown. It is important to note that more than one data string is generated with each round of oil sampling. This is because more than one test is accomplished when the samples are submitted. After the date/cumulative meter hour data pairs were extracted from the data strings, duplicate pairs were eliminated.

Table 6-8: Raw Oil Sampling Data

```
96022253  96021500021520000NU00030429000800170071000200200004000000000000000000000000000000558NNN

96022252  96021500021520000NN00060038000100030011000100170001000000000000000000000000000000878NNN

96022251  960215000215200       YY00030035000100030004000200330001000000000000000000000000000000NNN

95120975  9512060002124         Y007400060001000400080001000400010000000000000000000000000000213NNN

95120974  951206000212402124NN00020076000400300112000100120003000000000000000000000000002368NNN

95120973  951206000212402124NN      65000400080033000100130002000000000000000000000000000000NNN

95120972  9512060002124021        Y   00140001000100060001001200010000000000000000000000000046NNN

95120971  951206000212402204YY00030017000100020004000100180001000000000000000000000000000000NNN

96062170  960614000156500000NN001000110001000800210010000400010002000000000000000000000000000NNN

96062169  960614000156500000NN00010001000100010004000100010001000200000000000000000000000206NNN
```

The date/cumulative meter hour data pairs were then associated with date/cumulative cost data pairs. This is depicted in Table 6-9. It is important to note that cumulative costs were needed for this pairing. This ensured that no incremental costs were eliminated/lost. Associating the hour data with the cost data was a manual process. If a data point from the oil sampling database occurred on or before the 15th of the month, it was assumed to have occurred at the beginning of the month. If it occurred after the 15th of the month, it was assumed to have occurred at the end of the month. Cost data for a particular month are the cumulative costs for the *end* of the month. In Table 6-9, the first data pair from the oil sampling database was not usable to generate a point for use in the analysis. The oil change occurred at the beginning of April, 1996. It should have been associated with cost data from the end of March, 1996. It could not be, so it was not used. It can be seen that every month of cost data did not have an associated oil change to justify a data point.   The last two oil samplings in Table 6-9 are also of interest. Both readings were taken at points in their respective months such that they both should have been associated with the cost data from December 1996. It is not possible to associate two hour readings with one month's

worth of cost data. In this particular case, the hour reading from the 6th of January were used because it was taken on a date that was closer to the 31st of December than the other one. This was the other aspect of ensuring data points were not fabricated. No more than one month was associated with any given oil change and not more than one oil change was associated with any given month.

Table 6-9: Oil Sampling Data Pair Association

Cost Data		Association	Oil Sampling Data	
Cum. Cost	**Month**	**Hours**	**Date**	**Hours**
25000	Apr-96		04/01/96	12,857
25500	May-96	13104	05/22/96	13,104
27457	Jun-96		08/16/96	13,344
30005	Jul-96		11/07/96	13,592
30125	Aug-96	13344	12/22/96	13,878
30125	Sep-96		01/06/97	13,978
30700	Oct-96	13592		
30770	Nov-96			
31015	Dec-96	13978		
32500	Jan-97			
32900	Feb-97			

6.5   SPREADSHEET MANIPULATIONS TO END PRODUCT

The final task in the process of forming the analysis data pairs was accomplished in the spreadsheet program. Although a number of these manipulations could have been done in the database program, some flexibility is lost when attempting to use the database program for data splitting. In a spreadsheet, it is much easier and faster to make observations and necessary adjustments if the prediction and validation sets are not "similar but different" on the first attempt.

The first additional manipulation in the spreadsheet program was to add five additional columns to the output depicted in Table 6-7. These columns are depicted in Table 6-10.

Table 6-10: Additional Columns

EQ#	CRC	CCI	HOURS /1000	CCI-1
23789	70524.83	1.639706	16.654	0.639706
	71365.8	1.647334	16.7325	0.647334
	74350.01	1.674403	16.7325	0.674403
	75454.87	1.684425	16.848	0.684425
	76381.52	1.69283	17.008	0.69283
	77757.92	1.705315	17.08	0.705315
	78420.29	1.711323	17.208	0.711323
	80626.61	1.731336	17.344	0.731336
	88079.63	1.798939	17.4945	0.798939
	89104.96	1.80824	17.7045	0.80824
	90728.01	1.822962	17.9055	0.822962

The first additional column, "eq#", may seem to be a repeat of "EQNUM" from Table 6-7. The subtle difference is that only the first data string associated with each machine has an entry in the "eq#" column. The logic for doing this was "IF (EQNUM(current line) = EQNUM(previous line) THEN = "", ELSE = EQNUM(current line)". The data were sorted by equipment number then date before this was accomplished.

The second column in Table 6-10 was cumulative repair cost. This was calculated using the following logic: "IF (eq#(current line)= "", THEN = crc (previous line) + Mcost/Rindex(current line), ELSE = Mcost/Mindex(current line). This ensured that data with the proper inflation corrections were used.

The CCI was then calculated using Eq ?? from Chapter 1. The fourth and fifth columns, hours/1000 and CCI-1 were the format of the data needed for SAS. At this juncture, the procedures for incorporation of cumulative hours from oil-sampling databases were employed if necessary.

If the data set was sufficiently large, the data-splitting technique described in Section 5.2.2 was used on the data following incorporation of the oil-sampling data (if applicable). The data splitting process was repeated if necessary to come up with suitable estimation and prediction data

sets.   The prediction data were then set aside for future cross-validation. The data were then ready to be broken down into the SAS data sets.

6.5.1   Data Set #1: All but repeated points

The first data set, that of all available data points was formed by eliminating the repeated points from the estimation data set. This was done by adding two columns to the spreadsheet in Table 6-10. The columns were identical to "hours/1000" and "cci-1" except that data pairs that had repeated values of "hours/1000" were eliminated. This is illustrated in Table 6-11. These two columns were then sorted in order of ascending hours/1000 with the blank spaces removed. They were then ready for SAS analysis.

Table 6-11: All But Repeated Points

Unfiltered Data		Data for SAS	
Hours/1000	CCI-1	Hours/1000	CCI-1
16.654	0.639706	16.654	0.639706
16.7325	0.647334	16.7325	0.647334
16.7325	0.674403		
16.848	0.684425	16.848	0.684425
17.008	0.69283	17.008	0.69283
17.08	0.705315	17.08	0.705315
17.208	0.711323	17.208	0.711323
17.344	0.731336	17.344	0.731336
17.4945	0.798939	17.4945	0.798939
17.7045	0.80824	17.7045	0.80824
17.9055	0.822962	17.9055	0.822962

6.5.2   Data Set #2: 500-hour intervals

The second data set formed was that of data pairs at 500 hour intervals.   Three additional columns were added to Table 6-10. This is illustrated in Table 6-12. The column "Rounded Hours" contains the cumulative hours rounded to the next lowest 500 hour interval. The first entry for each machine was marked as a negative number to signify that it would not be used as a data point (the only scenario where the first point could be used would be one in which the cumulative hours at that point fell exactly on a 500 hour interval. Interpolations were performed

between the data pairs associated with the first occurrence of each rounded 500 hour interval and the data pairs that immediately preceded them. As mentioned earlier, the data set depicted is not complete—there were usually more than two interval data pairs associated with each machine. Once again, after all the calculations were done, the data were sorted and the blanks were removed.

Table 6-12: Interval Data Set

Unfiltered Data		Rounded Hours	Data for SAS	
Hours/1000	CCI-1		Hours/1000	CCI-1
16.654	0.639706	-16.5		
16.7325	0.647334	16.5		
16.7325	0.674403	16.5		
16.848	0.684425	16.5		
17.008	0.69283	17	17	0.69241
17.08	0.705315	17		
17.208	0.711323	17		
17.344	0.731336	17		
17.4945	0.798939	17		
17.7045	0.80824	17.5	17.5	0.79918
17.9055	0.822962	17.5		

6.5.3 Data Set #3: Average of 500-hour intervals

The third data set formed consisted of the average values for each 500-hour interval represented in the interval data set. These averages could have been found in a number of different ways. It was found that a good way to do this was to use the pivot table feature in Microsoft Excel. The pivot table yielded the average values in a format that was already sorted with the blanks removed.

6.5.4 Data Set #4: Final data points

The final data set formed consisted of simply the last data pair for each machine. Once again, the data were sorted in ascending order and the blanks were removed. This was the final step of data preparation.

6.6   DESIRED END PRODUCT

With all the data manipulations an filtering described, it is important to now understand what the end product is. The data will be entered into SAS in two columns. The columns are the data pairs that were described in Chapter 4. A depiction of what the analysis data sets will look like is available in Table 6-13.

Table 6-13 : Desired Data Sets

Data for Company "A" Off-Road Trucks							
All Points		**Intervals**		**Ave. of Intervals**		**Final Points**	
Hours/ 1000	CCI-1	Hours/ 1000	CCI-1	Hours/ 1000	CCI-1	Hours/ 1000	CCI-1
7.011	0.624695	0.5	0.006842	0.5	0.003927	0.967	0.007021
7.035	0.635745	0.5	0.002504	1	0.008361	1.083	0.006019
7.203	0.646613	0.5	0.004825	1.5	0.014133	2.172	0.018362
7.504	0.651539	0.5	0.001847	2	0.020078	2.28	0.029484
7.573	0.653004	1	0.011799	2.5	0.031891	3.559	0.043702
7.613	0.653004	1	0.006977	3	0.036653	4.251	0.059629
8.175	0.585759	1	0.009102	3.5	0.066682	6.113	0.200166
8.228	0.585909	1	0.017643	4	0.117545	6.234	0.26857
8.28	0.585953	1	0.00595	4.5	0.186013	7.45	0.242082
8.338	0.587091	1.5	0.01735	5	0.193697	7.613	0.653004
8.388	0.590269	1.5	0.021579	5.5	0.22411	7.89	0.546645
8.452	0.594843	1.5	0.01124	6	0.238062	8.034	0.354451
8.51	0.596139	1.5	0.005462	6.5	0.275317	8.193	0.507619
8.628	0.600877	2	0.031894	7	0.31547	9.532	0.671237
8.715	0.604324	2	0.017498	7.5	0.415183		
8.751	0.607374	2	0.011302	8	0.3771		
9.305	0.645429	2	0.018463	8.5	0.427801		
9.514	0.6463	2	0.015884	9	0.484845		
9.532	0.671237	2.5	0.058791	9.5	0.550709		
…	…	…	…	…	…		

The first column for each of the four data sets consists of the cumulative meter hours divided by 1000. The second column is composed of the Cumulative Cost Indices (CCIs) associated with each of those hour meter values minus one (to facilitate regression through the origin). There are a total of eight columns for each machine. The first two columns are all of the data pairs available

for each machine except the duplicate data pairs. The second set of two columns are interpolated data pairs at 500 hour intervals for each machine. The third set of two columns are the average indices of those interpolated points. The fourth set of columns consists of only the final data pairs for each machine in the fleet. A total of 17 tables in format of Table 6-13 were produced in the course of the data preparation.

6.7   SUMMARY

This chapter was the first of three in Part III of the dissertation—"The Work". . This chapter was devoted to describing the data preparation process. It is the smallest chapter of the three in this part, but it represents the biggest time investment. It was important to be very meticulous at this stage of the research. If the data were unreliable due to improper formatting, conclusions based on them would be unreliable. The statistical analysis is also more streamlined and manageable when the data are all formatted the same way.

Chapter 7 will take the data prepared in this chapter and analyze them thoroughly. The best model and the best data set will be chosen. Model validation for the selected model and data set will be presented. Chapter 7 will also provide a number of comparisons and sensitivity analyses.

CHAPTER 7: ANALYSIS

The purpose of this chapter is to document the selection of the best statistical model for describing the CCI in terms of cumulative hours of use on construction machinery. Chapter 6 explained how the four data sets for each of the 17 fleets of equipment were formed—these data sets will now be used to appropriate regression equations. The selection of the overall best statistical model and the selection of the best of the four data sets will be the end product of this chapter.

In this chapter, the following main areas will be discussed/developed:

- Preliminary analyses

- Intermediate analyses

- Model selection

- Data set selection

- Statistical performance

7.1   PRELIMINARY ANALYSES

The overall purpose of the preliminary analyses was to eliminate some of the 19 models that were under consideration. The preliminary analyses also were designed to give preliminary readings of how well the models performed. There were three aspects of the preliminary analyses in this research that were inter-related but are best discussed separately. They are the analyses pertaining to:

- linear models

- non-linear models

- data set selection
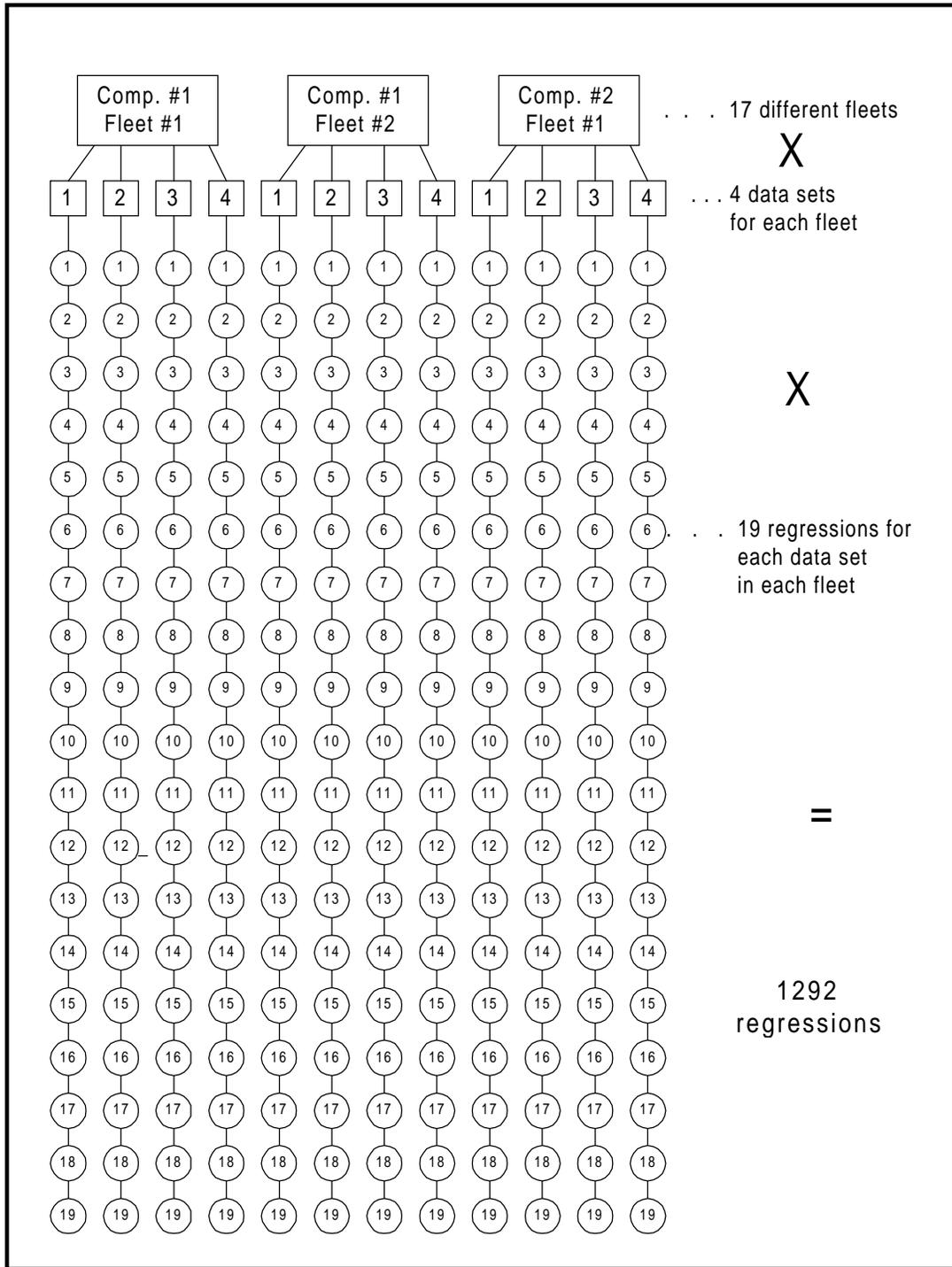
Figure 7-1: Preliminary Regressions

This section will discuss these three analyses and present the results. An overview of the nature of the task can be seen in Figure 7-1. There are a total of 1292 regression models at this stage of the analysis. The 19 different models were originally introduced in Chapter 5. They are presented again here as an additional reference:

20. $y = 1 + \beta_1 x + \varepsilon$

21. $y = 1 + \beta_1 x + \beta_2 x^2 + \varepsilon$

22. $y = 1 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

23. $y = 1 + \beta_4 e^x + \varepsilon$

24. $y = 1 + \beta_1 x + \beta_2 x^2 + \beta_4 e^x + \varepsilon$

25. $y = 1 + \beta_3 x^3 + \varepsilon$

26. $y = 1 + \beta_1 x + \beta_3 x^3 + \varepsilon$

27. $y = 1 + \beta_1 x + \beta_4 e^x + \varepsilon$

28. $y = 1 + \beta_1 x + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

29. $y = 1 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

30. $y = 1 + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

31. $y = 1 + \beta_2 x^2 + \beta_4 e^x + \varepsilon$

32. $y = 1 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

33. $y = 1 + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

34. $y = 1 + \beta_2 x^2 + \varepsilon$

35. $y = 1 + \alpha x^\beta$ transformed to: $\ln(y-1) = \ln(\alpha) + \beta \ln(x)$

36. $y = 1 + x^\beta$ transformed to: $\ln(y-1) = \beta \ln(x)$

37. $y = 1 + e^{\beta(x)}$ transformed to: $\ln(y-1) = \beta x$

38. $y = 1 + \alpha e^{\beta(x)}$ transformed to: $\ln(y-1) = \ln(\alpha) + \beta x$

The first fifteen models are *linear models*—the coefficients are linear even though the regressors (x, x^2, x^3, and e^x) may not be. The last four models are *non-linear models*—at least one of the coefficients is in a non-linear form.

7.1.1   Linear Models

As mentioned in Section 5.1.2, there were 15 linear models considered for expressing CCI in terms of cumulative hours of use. These linear models represent all possible combinations of the regressors: x, x^2, x^3, and e^x except for the model that has no regressors in it. It was expected that some of the models would yield adequate results and that some would not.

The preliminary analysis was performed by subjecting the four data sets from each of the seventeen groups of equipment under consideration to the NOINT macro in SAS PROC IML. Sample output from the NOINT macro is depicted in Table 7-1. Fifteen regressions were performed each time the NOINT macro was used.

Table 7-1: Sample NOINT Output

X	X2	X3	EXP_X	MSE	RSQ	ADJRSQ	CP	RSQPRESS	SETNUM	COMPNUM	EQTYPE
.	0.00499	0.00019	.	0.01099	0.84125	0.84031	0.16	0.83855	1	2	3
0.0096	0.00776	.	.	0.01101	0.84101	0.84007	0.67	0.83868	1	2	3
.	0.00468	0.00023	-6E-07	0.01102	0.84132	0.83991	2	0.8381	1	2	3
-0.0024	0.00564	0.00015	.	0.01102	0.84128	0.83986	2.1	0.83829	1	2	3
-0.0082	0.00754	.	4.6E-07	0.01103	0.84108	0.83967	2.51	0.83819	1	2	3
-0.0003	0.00479	0.00022	-6E-07	0.01105	0.84132	0.83943	4	0.83781	1	2	3
0.01415	.	0.0006	-2E-06	0.01107	0.84062	0.8392	3.49	0.83762	1	2	3
.	0.00642	.	1.6E-06	0.01108	0.84	0.83906	2.8	0.8374	1	2	3
0.01745	.	0.00052	.	0.01111	0.83947	0.83852	3.93	0.83697	1	2	3
.	0.00665	.	.	0.01114	0.83854	0.83806	3.9	0.83677	1	2	3
.	.	0.00084	-6E-06	0.01163	0.83197	0.83097	19.82	0.82898	1	2	3
.	.	0.00073	.	0.01261	0.81728	0.81675	48.9	0.81531	1	2	3
0.04442	.	.	9.3E-06	0.0142	0.79479	0.79358	98.53	0.78956	1	2	3
0.05428	.	.	.	0.0189	0.72619	0.72538	241.79	0.7236	1	2	3
.	.	.	2.8E-05	0.05714	0.17206	0.16962	1415.17	0.13897	1	2	3

The first four columns are the parameter estimates. The next five columns are measures of effectiveness. "SETNUM" is a number from one to four that identifies the data set in use. "COMPNUM" identifies the company and "EQTYPE" identifies the fleet. The NOINT macro was used 68 times—four times (once for each data set) for each of the seventeen fleets. The output from all 68 of these SAS runs was then combined in an Excel spreadsheet. An algorithm

was written in Excel to assign the appropriate model number (1-15) based on the presence of values in the first four columns. The Excel file was then brought back into SAS for Kruskal-Wallis analysis of the rankings of adjusted R^2 and R^2_{press} values. For both of these analyses, a higher mean score signifies a better model.

The rankings for adjusted R^2 for all of the models are shown in Table 7-2. High values signify a model with better performance. The model numbers correspond to those listed in Chapter 5. The p-value of the test was less than 0.0001. Based on these rankings, the best performing single parameter model was x^2 (#15), the best two parameter model was x, x^3 (#7), the best three parameter model was x^2, x^3, e^x (#13), and the full model (#10) performed the better than any of the partial models.

Table 7-2: Linear Adjusted R^2 Rankings

Model #	Mean Score
10	662.669118
13	642.838235
9	632.286765
3	622.830882
5	621.176471
7	601.316176
11	590.941176
2	585.455882
12	572.764706
14	525.455882
8	520.941176
15	502.544118
6	433.132353
1	404.014706
4	161.161765

The rankings according to R^2_{press} are depicted in Table 7-3. Once again, the p-value of the test was less than 0.0001. When ranked by R^2_{press}, the best performing models were x^2 (#15) for single parameter models; x, x^2 (#2) for two parameter models; x, x^2, x^3 (#3) for three parameter models. The full model (#10) performed well, but was not the top performing model.

Table 7-3: Linear R^2_{press} Rankings

Model #	Mean Score
2	632.625
7	625.75
11	619.286765
3	612.727941
10	600.933824
13	572.727941
15	569.823529
9	567.147059
5	560.757353
12	524.610294
6	497.588235
1	494.183824
8	490.316176
14	488.316176
4	184.647059

An aid for visualizing the rankings in the above two tables is provided in Table 7-4. Models that are in the upper left corner of this matrix are the models that performed well in both measures of effectiveness. Models selected for further consideration are depicted with non-filled circles, models eliminated from consideration are depicted with blackened circles. The cutoff line was drawn at roughly a 45 degree angle to separate most of the models selected from the rest of the pack. The models above the cutoff line were: 2, 3, 7, 10, and 13. An exception was made to the cutoff line to include two single parameter models. Model 15 (x^2) was included because it was the best-performing single parameter model. Additionally, the single parameter model 1 (x) was chosen for further study for comparison purposes.

Table 7-4: Comparison Matrix for Linear Models

	Rank	2	7	11	3	10	13	15	9	5	12	6	1	8	14	4
											R²press					
	10					○										
	13						○									
	9								●							
	3				○											
	5									●						
	7		○													
	11			●												
	2	○														
	12										●					
	14														●	
	8													●		
	15							○								
	6											●				
	1												○			
	4															●

(Left axis label: Adjusted R²; Cutoff Line shown as a diagonal through the matrix.)
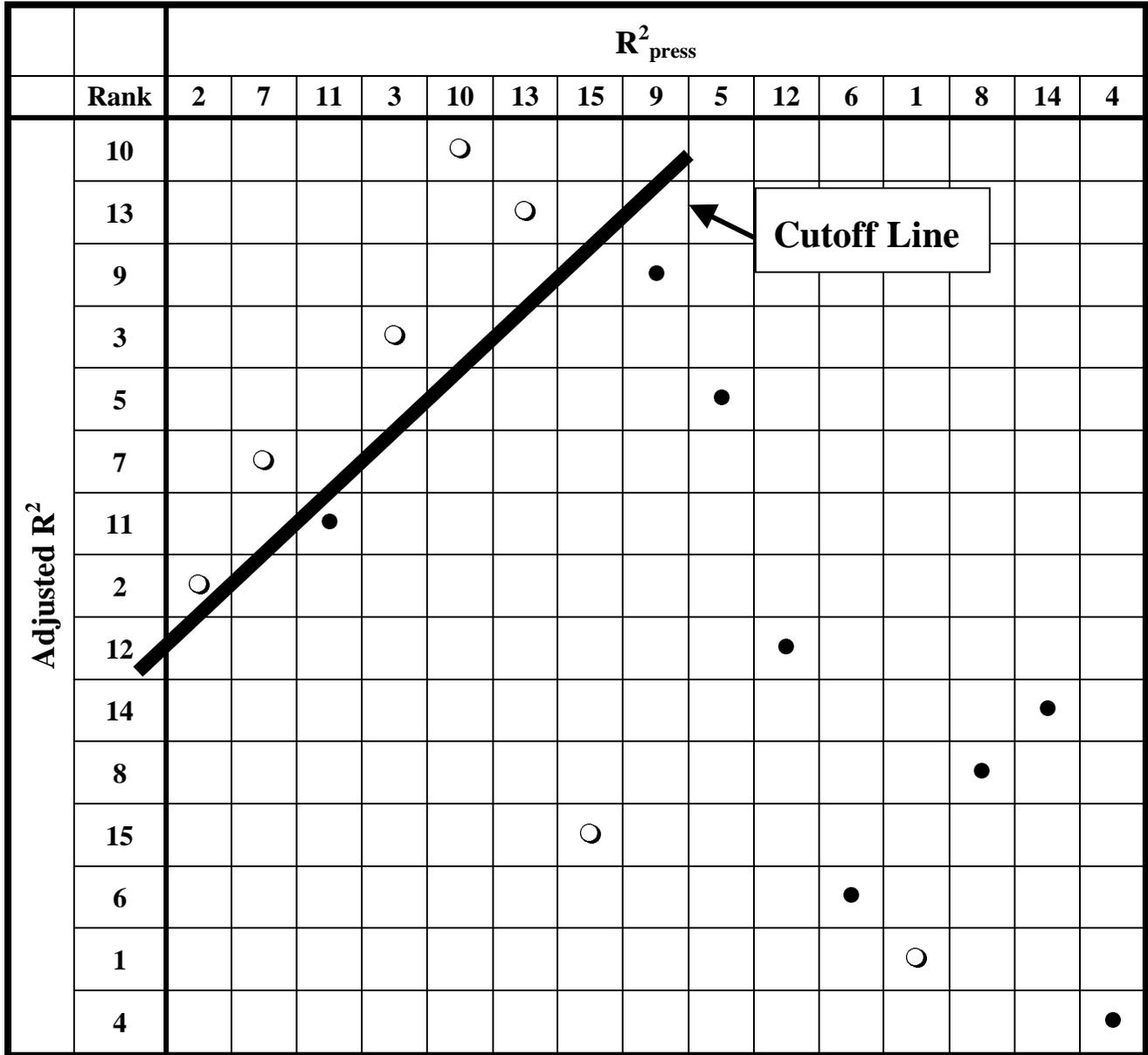
7.1.2   Non-Linear Models

The non-linear models did not lend themselves well to the use of the NOINT macro. This is because some of the non-linear models had an intercept term and some did not (this intercept term is eliminated after all the log transformations are made). So, each of the four non-linear models

were analyzed separately for each of the 68 data sets (17 x 4). The SAS results were collated and filtered in Monarch before being brought back into SAS for non-parametric analysis.

Table 7-5: Non-Linear Adjusted R^2 Rankings

Model #	Mean Score
16	191.6418
18	174.6418
19	91.23881
17	80.47761

Table 7-6: Non-Linear R$^2_{press}$ Rankings

Model #	Mean Score
16	186.4179
18	168.1045
17	93.73134
19	89.74627

Table 7-7: Comparison Matrix for Non-Linear Models

		R^2_{press}			
	Rank	**16**	**18**	**17**	**19**
Adjusted R^2 16	○			
18		●		
19				●
17			●	

The results for the Kruskal-Wallis tests pertaining to adjusted R^2 and R^2_{press} for the non-linear models are shown in Table 7-5 and Table 7-6. These results are combined in Table 7-7.   Model 16 ($y = 1 + \alpha x^{\beta}$) had the best ranking for both of these measures of effectiveness. When all four

non-linear models were analyzed together, the p-values were less than 0.0001 for both measures of effectiveness.

Model 18 ($y = 1 + e^{\beta(x)}$) seemed to perform reasonably well compared to model 16 for both types of R^2. The Wilcoxon rank sum test (same as Kruskal-Wallis except there are only two levels) was performed on the rankings of model 18 vs. model 16. The tests were significant with p-values of 0.0232 for adjusted R^2 and 0.0381 for R^2_{press}. Based on this, model 16 will be the sole non-linear model that will undergo the detailed analysis.

7.1.3   Data Sets

The data sets were evaluated at this point more as a matter of interest than as a filter to cut down on the number of intermediate analyses. It was felt that it would be good to get a preliminary look at how well each data set performed when viewed from a macro level for all possible models.

Each of the seventeen fleets was represented by four different data sets. The four data sets as presented in Chapter 4 were:

- Data set 1: all data pairs except for repeated points

- Data set 2: data pairs interpolated to 500 hour intervals

- Data set 3: average values of data pairs interpolated at 500 hour intervals

- Data set 4: only the final data pair for each machine

The results of the Kruskal-Wallis tests concerning the data set types are given in Table 7-8 and Table 7-9. The p-values for both tests were less than 0.0001.

Table 7-8: Data Set Adjusted R^2 Rankings

Set #	Mean Score
3	680.088235
4	527.564338
2	495.931985
1	474.415441

Table 7-9: Data Set R^2_{press} Rankings

Set #	Mean Score
3	691.387868
1	580.645221
2	560.505515
4	345.461397

Table 7-10: Comparison Matrix of Data Sets

		R^2_{press}			
	Rank	**3**	**1**	**2**	**4**
Adjusted R^2 — **3**		○			
4					○
2				○	
1			○		

These results are presented graphically in Table 7-10. Looking strictly at the numbers, it appears that set number 3 is clearly the best—but the statistical concerns with the different data sets must be considered before making a definitive decision on which data set is the best. If data set 3 is eliminated, the other three data sets have similar performance (with the exception of data set 4 for R^2_{press}.)

7.2   INTERMEDIATE ANALYSIS

The intermediate analysis provides the basis for the further narrowing of model and data set choices. As a recap, the following 8 models were selected for the intermediate analysis during the preliminary analysis stage:

- Model #1: $y = 1 + \beta_1 x + \varepsilon$

- Model #2: $y = 1 + \beta_1 x + \beta_2 x^2 + \varepsilon$

- Model #3: $y = 1 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

- Model #7: $y = 1 + \beta_1 x + \beta_3 x^3 + \varepsilon$

- Model #10: $y = 1 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

- Model #13: $y = 1 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 e^x + \varepsilon$

- Model #15: $y = 1 + \beta_2 x^2 + \varepsilon$

- Model #16: $y = 1 + \alpha x^\beta$ --transformed to-- $\ln(y-1) = \ln(\alpha) + \beta \ln(x)$

Represented on this list are the best one, two, and three parameter linear models for each of the two measures of effectiveness that were considered in the rough analysis. The full linear model and the best non-linear model are also on the list. The only model on the list that is not there due to its performance is model #1. Model #1 is included so that the simplest definition of CCI in terms of cumulative hours of use can be evaluated as it relates to the other models.

The intermediate analysis took part in two stages. Stage one concerned the significance of the parameters in the models. Stage two revisited the measures of performance.

7.2.1   Stage 1: Parameter Significance

Parameter significance is a very important part of model selection. If the parameters involved in a model are not statistically significant, it is doubtful that the model associated with those parameters is the best one to describe the phenomenon under study. The eight models listed above were again analyzed using SAS®. This time instead of using a macro from within PROC IML, the regressions were performed using the PROC REG option—this provided more detailed information on each of the regressions. Figure 7-2 shows the regressions to be accomplished at this stage of the analysis—a total of 544.
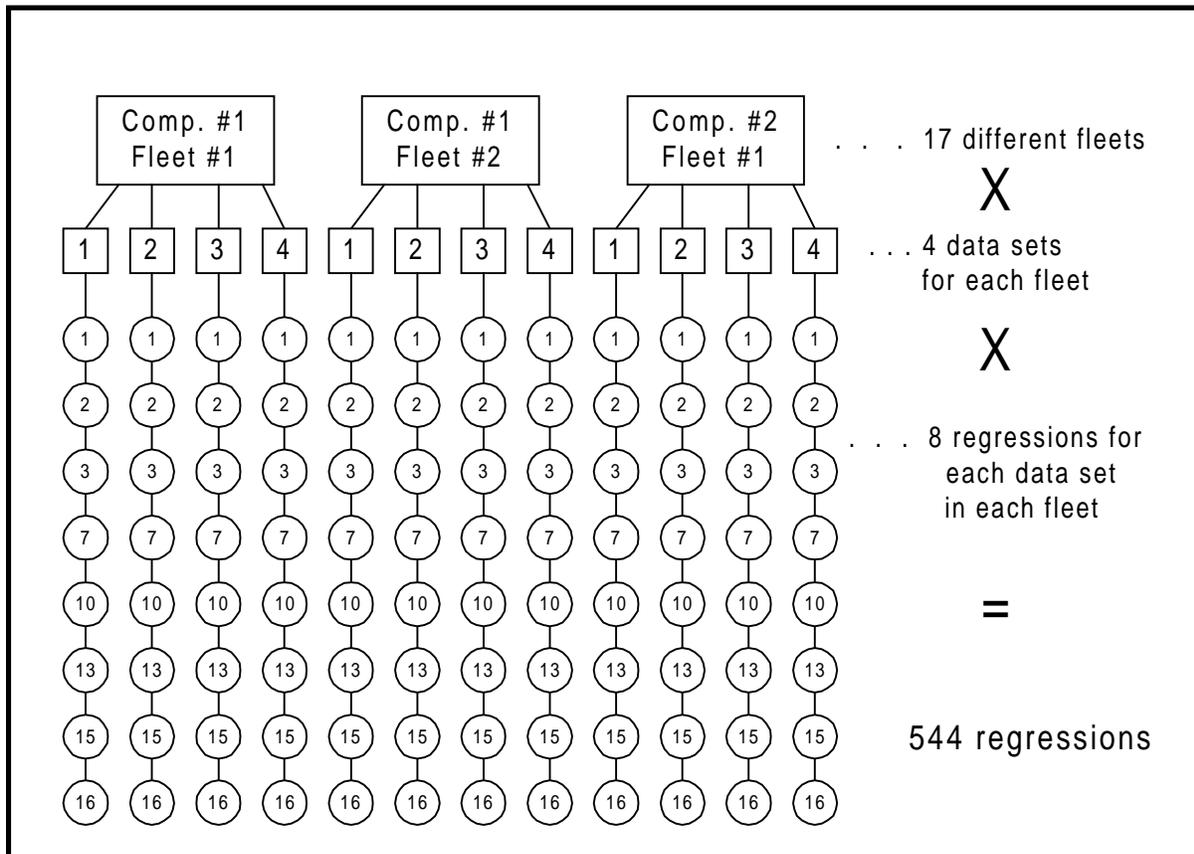
Figure 7-2: Stage One Intermediate Analysis Regressions

The average p-values for each the parameters in each of models listed above are given in Table 7-11. Recall that lower p-values indicate higher parameter significance. Recall that in Chapter 5 the decision criteria of p-value less than or equal to 0.2 was given for acceptance of the parameter as part of the model. On average, models 3, 10, and 13 do not meet this decision criteria. This leaves only the one and two parameter models. This is not entirely surprising. As parameters are added to a model, the significance of the parameters already in the model tends to decrease (p-values go up). Of the two parameter models, model 16—the non-linear model—has the lowest p-values for its parameters. Models 2 and 7 meet the standard and are acceptable, but are not quite as good model 16. The parameter significance for the single parameter model x was slightly better than that of the model x^2—but both were well within the tolerances specified. Summarizing, models 3, 10, and 13 are eliminated from contention at this point. Models 1, 2, 7, 15, and 16 are still under consideration.

Table 7-11: Average p-values For Parameter Significance

	e^x	intercept	ln(x)	x	x^2	x^3
Model 1				0.003033824		
Model 2				0.167873529	0.075336765	
Model 3				0.330389706	0.226152941	0.252830882
Model 7				0.121214925		0.076108955
Model 10	0.277932836			0.346692537	0.330731343	0.246159701
Model 13	0.279404478				0.172791045	0.223204478
Model 15					0.006252239	
Model 16		0.027734328	0.056825373			

Table 7-12: p-values by Data Set

		PARAMETER					
MODEL	**SETNUM**	e^x	**Intercept**	**ln(x)**	**x**	x^2	x^3
1	1				0.0001		
	2				0.000247		
	3				0.000494		
	4				0.011294		
2	1				0.045671	0.042053	
	2				0.121706	0.060482	
	3				0.117112	0.008482	
	4				0.387006	0.190329	
7	1				0.080482		0.027253
	2				0.115329		0.053847
	3				0.128088		0.0158
	4				0.163444		0.21575
15	1					0.000112	
	2					0.001759	
	3					0.004265	
	4					0.019663	
16	1		0.003841176	0.043382			
	2		0.035352941	0.055582			
	3		0.036358824	0.064647			
	4		0.0358625	0.064119			

The performance of the various data sets regarding p-values was also evaluated at this point. Table 7-12 depicts the p-values of each of the five models still under consideration broken down by data set. For the four different data sets, the only one that did not consistently provide p-values that met the criteria was data set number four—the data set consisting of solely the final data pair for each machine. This is due to a smaller number of points available for the regression.

The smaller number of points makes the tests concerning data set 4 less powerful. The other three data sets provided acceptable average p-values for all five models still under consideration. Data set 1 (all but repeated data points) provided the lowest p-values for all models still under consideration. Data set 2 (500-hour intervals) provided slightly better p-values than data set number 3 (averages of 500-hour intervals) in most cases. Data set 4 is eliminated from consideration at this point due to its failure to meet the decision criteria in all cases combined with its performance regarding R^2_{press} (Section 7.1.3).

7.2.2 Stage 2: Measures of Performance

It is important to take a more detailed look at adjusted R^2 and R^2_{press} now that a number of models have been eliminated from consideration. The models left at this point are:

- Model #1: $y = 1 + \beta_1 x + \varepsilon$

- Model #2: $y = 1 + \beta_1 x + \beta_2 x^2 + \varepsilon$

- Model #7: $y = 1 + \beta_1 x + \beta_3 x^3 + \varepsilon$

- Model #15: $y = 1 + \beta_2 x^2 + \varepsilon$

- Model #16: $y = 1 + \alpha x^\beta$ --transformed to-- $\ln(y\text{-}1) = \ln(\alpha) + \beta \ln(x)$

Also, data set number 4 was eliminated. The regressions undertaken for this stage of the analysis are depicted in

Figure 7-3. There are a total of 255 regressions to analyze at this stage. To take a closer look at the actual performance, SAS® was used to calculate the mean values for each of the measures of performance for each model. Then parametric tests were performed to discern the differences between the five models. The SAS output for these tests is given in Figure 7-4 and Figure 7-5. Fisher's Least Significant Difference test was the statistical test used. A good discussion of this test appears in Ott (1993, pp. 807-836).
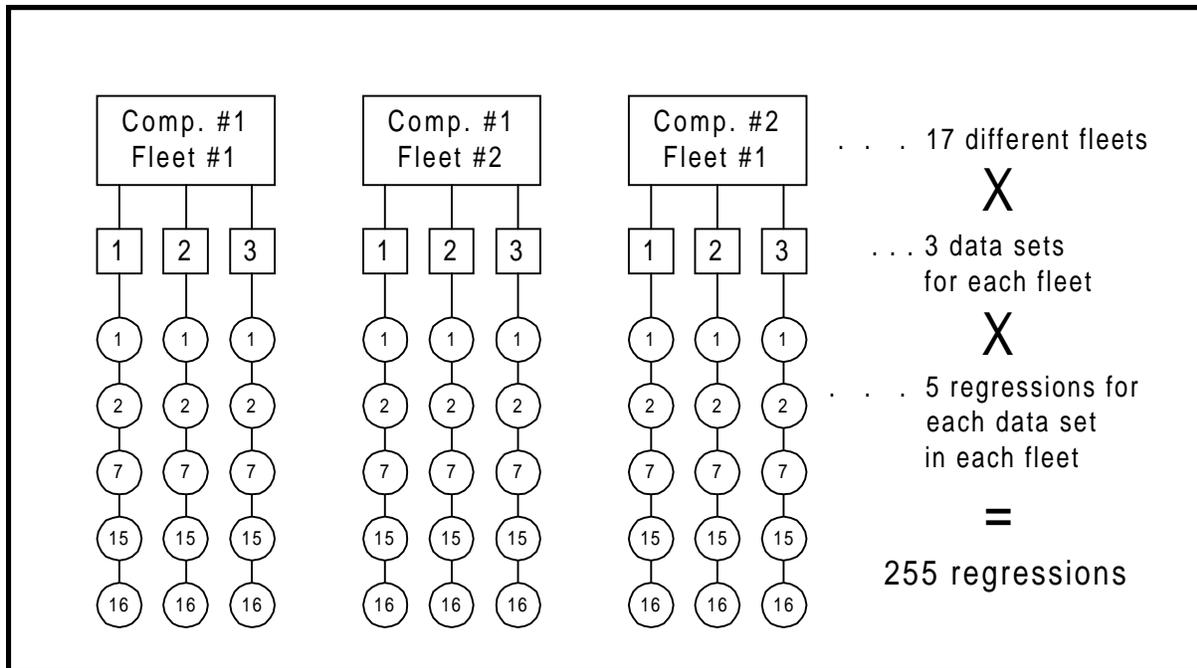
Figure 7-3: Stage Two Intermediate Analysis Regressions

Fisher's Least Significant Difference (LSD) places sample means into groups that can be considered to have similar mean values. For adjusted R^2, there were two groupings. Group "A" contained models 2, 7, 16, and 15 (in that order). Group "B" contained models 15 and 1. The adjusted R^2 values for group "A" were better than those for group "B". Model 1 had significantly worse performance than the two-parameter models. Model 15, although included at the bottom of group "A", did not fit the data nearly as well as the two parameter models. A model that has an adjusted R^2 of less than 0.5 (on average) is probably not as good as one that has an adjusted R^2 of better than 0.75 (on average).

It is important to note that the standard deviation of model 15 was nearly 5 times that of the best performer, model 2. Model 1, the other single parameter model, also had a high standard deviation. The reason that the standard deviation can be greater than the normal range of R^2 (0.00-1.00) is that Myers' definition for adjusted R^2 for models without intercepts allows for negative values (Myers, 1990). In common sense terms, this implies that the one-parameter models can do a decent job in some instances—but do a poor job in others. One parameter does not allow the model sufficient freedom to provide a good fit in all cases. The standard deviations

of the two-parameter models are all very similar—and quite a bit smaller than those of the single parameter models. The tighter standard deviations imply that two parameters provide the models with enough flexibility to fit the data in most cases. The Fisher groupings tell only part of the story. The single parameter models just don't do as good a job at fitting the data as the two-parameter models.

```
         Level of              ------------ARSQ------------

     MDL          N        Mean                SD

     1           51     0.43397137         1.01984163
     2           51     0.76335765         0.26321169
     7           51     0.76254608         0.27343780
     15          51     0.49652647         1.27446186
     16          51     0.76166863         0.30416548


            T tests (LSD) for variable: ARSQ


         Alpha= 0.05   df= 250   MSE= 0.580179
               Critical Value of T= 1.97
         Least Significant Difference= 0.2971


      T Grouping                 Mean      N   MDL


                      A         0.7634    51   2
                      A
                      A         0.7625    51   7
                      A
                      A         0.7617    51   16
                      A
               B      A         0.4965    51   15
               B
               B                0.4340    51   1
```

Figure 7-4: Adjusted R^2 Output

The results concerning R^2_{press} were similar. Once again, the single parameter models had less suitable mean values than the two-parameter models. The single parameter models also had significantly greater variability than the two-parameter models. The "A" group for Fisher's LSD test was identical for the "A" group described above for adjusted R^2. The "B" group included the single parameter models, but also included model 16. Model 16 had a mean R^2_{press} that was less than models 2 and 7. The standard deviation for model 16 was also 50% greater than the standard deviation of the other two two-parameter models.

```
                    General Linear Models Procedure

        Level of         ----------RSQPRESS----------
        MDL        N        Mean              SD

        1          51     0.40481549       1.06850533
        2          51     0.74405745       0.28206933
        7          51     0.73817059       0.29623492
        15         51     0.47509941       1.31228486
        16         51     0.70721221       0.44113279


          T tests (LSD) for variable: RSQPRESS
         Alpha= 0.05   df= 250   MSE= 0.645142
                Critical Value of T= 1.97
          Least Significant Difference= 0.3133


         T Grouping              Mean      N   MDL

                    A          0.7441     51   2
                    A
                    A          0.7382     51   7
                    A
            B       A          0.7072     51   16
            B       A
            B       A          0.4751     51   15
            B
            B                  0.4048     51   1
```

Figure 7-5: R^2_{press} Output

On the basis of performance regarding adjusted R^2 and R^2_{press}, both of the single parameter models are eliminated from consideration at this stage. It is important to note that model 15 (x^2) performed better than model 1 (x). This helps to confirm that the growth of repair costs with accumulated hours is *not* a constant. A model that allows some curvature fits and predicts better than one that allows no curvature.

7.3   MODEL SELECTION

Three acceptable regression models were identified for further investigation in the intermediate analysis. These three models were:

- Model #2: $y = 1 + \beta_1 x + \beta_2 x^2 + \varepsilon$

- Model #7: $y = 1 + \beta_1 x + \beta_3 x^3 + \varepsilon$

- Model #16: $y = 1 + \alpha x^\beta$ --transformed to-- $\ln(y-1) = \ln(\alpha) + \beta \ln(x)$

The most appropriate of these three models will be selected in this section based upon a statistical issues and a comparison of the results obtained using the models. The regressions to be

performed for the initial stages of this analysis are depicted in Figure 7-6—there are 153 for this stage of the analysis.
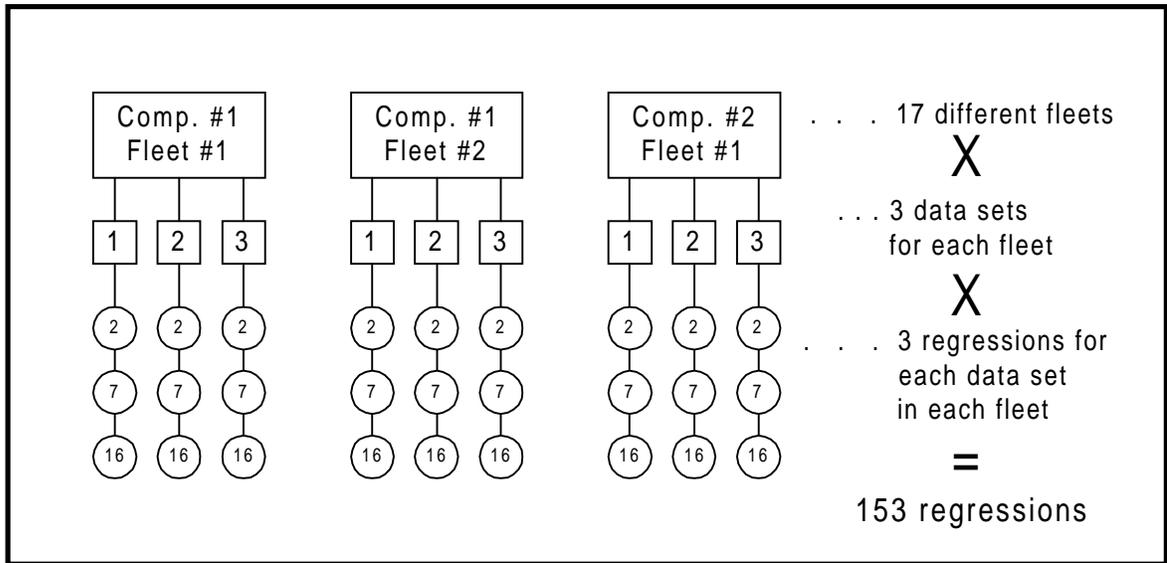
Figure 7-6: Regressions for Final Model Selection

7.3.1   Statistical Issues

A recap of how each of the three models fared concerning parameter significance and measures of performance is in order. Table 7-13 contains the average parameter significance for each of the three models for 17 fleets of equipment with three data sets each.   Table 7-14 contains the average adjusted R^2 values and the average R^2_{press} values for each of the three models. This table also contains the standard deviations for the measures of performance.

Table 7-13: Parameter Significance

MODEL	PARAMETER				
	INTERCEPT	ln(x)	x	x^2	x^3
2			0.0948	0.0370	
7			0.1079		0.0323
16	0.0251	0.0545			

Table 7-14: Measures of Performance

Model	R^2_{adj}		R^2_{press}	
	Mean	s	Mean	s
2	0.7633	0.2632	0.7440	0.2820
7	0.7625	0.2734	0.7381	0.2962
16	0.7616	0.3041	0.7072	0.4411

As can be seen from Table 7-13, the parameter significance for each of the three models improved slightly over what was presented in Section 7.2.1. The reason for this is the elimination of data set number 4 (final data pairs only). Model 2 is slightly better than model 7 for significance of "x" while model number 7 still retains a slight edge in parameter significance for the second parameter. Model number 16 is still better on average than the other two models, but not by the same margin as in Chapter 6. Although the first parameter for model 16 is clearly the most significant in the study, the second parameter is no longer the second best. All p-values were substantially better than the minimum requirement of p-value < 0.20 though, and none of the three models can be ruled out solely on the basis of p-values for parameter significance.

In terms of adjusted R^2 and R^2_{press}, model 2 performed slightly better than models 7 and 16. The differences for R^2_{press} were more than those for adjusted R^2, but not by so much that any one of the three models could be ruled out. However, the standard deviations for both measures of effectiveness were both higher for model 16—around 50% higher for R^2_{press}. But, once again, the measures of effectiveness are similar and acceptable—none of the three models can be ruled out solely on the basis of these measures. It can be said at this point that any one of these three models would probably do an adequate job of describing the CCI in terms of cumulative hours of use. But which one is best?

When working with models that use powers of the same regressor variable to describe the response variable, it is a generally accepted practice that all powers of the regressor variable up to the highest value in the model be included in the model. This is not to say that it is wrong to use a model such as model 7—it just is not as tidy a solution as would be desired. All other things being equal, model 2 is a better choice than model 7. Model 2 conforms with generally accepted

practice. But, all other things are not exactly equal. Model 2 performs better than model 7 for both measures of performance (albeit by a small margin). If the p-values presented in Table 7-13 are averaged for each model the results are:

- Model 2:   0.0659

- Model 7:   0.0701

- Model 16: 0.0398

Again, the difference between model 2 and model 7 is very slight, but model 2 edges out model 7. On this basis, model 7 is eliminated from contention. Model 2 is simpler, and it is ever so slightly better.

It is difficult to choose between model 2 and model 16 on the basis of simplicity or on the basis of straight statistical output presented thus far. Both models are simple to use and clean. Model 16 is better in terms of parameter significance and model 2 is better in terms of measures of performance. The statistical issues presented in Chapter 5 can provide some basis for the selection. The added concern of multiplicative versus additive error terms is a strike against model 16—and makes model 2 a little more attractive.

7.3.2   Preliminary Results

Comparisons of some of the actual results (L*) produced by the two models were accomplished. These results are depicted in Table 7-15. The results are in thousands of hours. For the purposes of this analysis, the results obtained using data set 3 were used (best performance without regard to statistical issues). The L* values in the table that are zero indicate fleets where the regressions produced curves that were not concave (no optimum solution). This will be addressed further in Chapter 8.

A specific problem areas for model 16 was the large dozers. Although some portion of a large L* can be attributed to collateral costs which were not included (this will be discussed further in Chapter 8), some of the L*'s produced using model 16 were exceptionally large. Specifically, the two fleets with lifespans of over 60 years are cause for concern. There was also a good deal more

variability in the results produced by model 16. Using model 2, the three fleets of articulated trucks analyzed had L* values that fell within 6000 hours of each other. The L* values for the same fleets found using model 16 covered a range of over 20,000 hours. The same can be said of mid-size dozers. Model 2 does have the same shortcoming regarding mid-size excavators—but this shortcoming was expected due to differences in the fleets analyzed and will be further addressed in Chapter 8. It is the opinion of the researchers that the L* values produced by model 2 are more consistent with experience than those produced by model 16.

Table 7-15: L* Model 2 vs. Model 16

FLEET	L* Model 2	L* Model 16
Articulated Trucks	17.39	11.54
Articulated Trucks	15.63	31.83
Articulated Trucks	11.98	17.94
Dual-engine Scrapers	36.61	57.88
Large Dozers	32.67	63.79
Large Dozers	44.75	92.86
Large Dozers	0.00	45.56
Large Dozers	0.00	0.00
Mid-size dozers	10.15	6.03
Mid-size dozers	11.13	15.10
Mid-size dozers	18.80	26.33
Mid-size Excavators	22.99	30.65
Small Excavators	51.18	45.03
Small Excavators	11.39	16.22
Track Loaders	22.10	28.22
Wheel Loaders	23.89	22.42
Wheel Loaders	0.00	0.00

Model 2 is superior to model 16 in measures of performance, statistical simplicity, and actual results produced. Model 16 is better than model 2 regarding the significance of model parameters. Based on these considerations, model 2 is chosen as the best model for the purposes of this research. Although models 7 and 16 would probably provide adequate performance, it is felt that model 2 is superior when all things are considered.

7.4   DATA SET SELECTION

The second selection that must be made to complete the analysis is that of which data set is the best. After the selection of model 2 as the best model in Section 7.3, the remaining regressions to be analyzed are reflected in Figure 7-7. The 1292 regressions that the analysis started with have been pared down to 51.
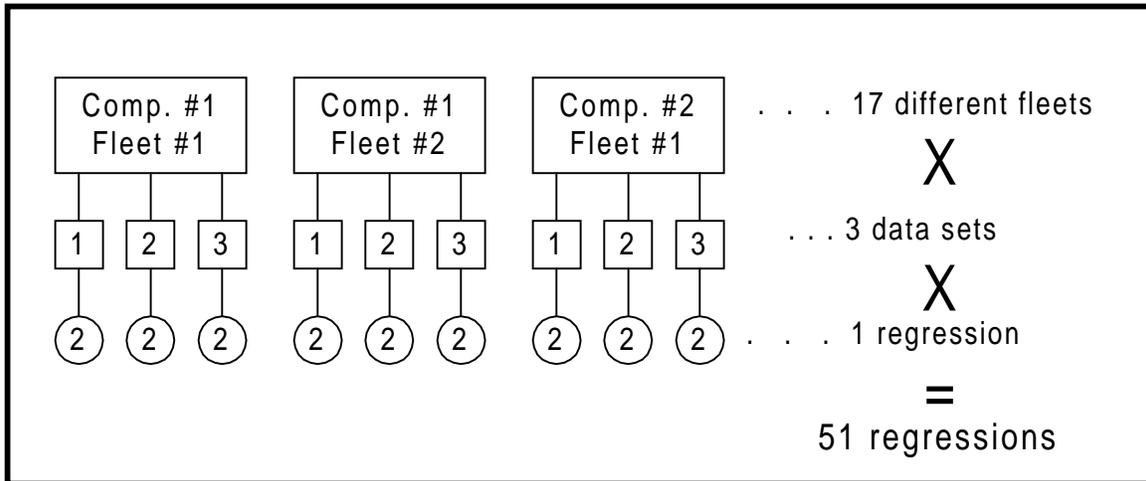
Figure 7-7: Regressions for Data Set Selection

7.4.1   Parameter Significance

Once again, parameter significance is a concern. The parameter significance for the three data sets as they relate to model 2 are depicted in Table 7-16. All three data sets met the criteria of p-value < 0.20 for both parameters. Data set 1 had the best average p-values and the lowest average variability. Data set 3 was second best and data set 2 was the third best. It is important to note that for all three of the data sets, the parameter significance for x^2 was better than that for x. The importance of this will be demonstrated in Chapter 8. Results obtained from this model are more sensitive to the values associated with x^2 than they are to the values associated with x.

It should be mentioned that the results obtained for these p-values are relatively predictable. It makes sense that data sets 1 and 3 did better than data set 2. Data set 1 had a lot of redundancy

(nearly repeated points) in its data pairs. Data set 3 had a certain amount of variability removed when the 500-hour interval values were averaged.

Table 7-16: Parameter Significance for Data Sets for Model 2

Data Set	Data	x	x^2	Total Average
1	Average of PVALUE	0.045670588	0.042052941	0.043861765
	StdDev of PVALUE	0.084678346	0.145816233	0.117426441
2	Average of PVALUE	0.121705882	0.060482353	0.091094118
	StdDev of PVALUE	0.167332664	0.1938043	0.18097597
3	Average of PVALUE	0.117111765	0.008482353	0.062797059
	StdDev of PVALUE	0.232983033	0.02046681	0.17193222

7.4.2   Measures of Performance

Measures of performance should also be compared to select the best data set. Results from Fisher's LSD comparisons of adjusted R^2 and R^2_{press} are presented in Figure 7-8 and in Figure 7-9.

```
              General Linear Models Procedure

 Level of          ------------ARSQ-----------
 SETNUM      N        Mean                SD


 1          17     0.72466000        0.25353625
 2          17     0.72365941        0.26825152
 3          17     0.84175353        0.26545459
      T tests (LSD) for variable: ARSQ
   Alpha= 0.05  df= 48  MSE= 0.068902
          Critical Value of T= 2.01
      Least Significant Difference= 0.181


  T Grouping              Mean      N   SETNUM

          A            0.84175     17   3
          A
          A            0.72466     17   1
          A
          A            0.72366     17   2
```

Figure 7-8: Adjusted R^2 Values for Data Sets

```
                General Linear Models Procedure

  Level of         ----------RSQPRESS----------
  SETNUM      N        Mean                SD


  1          17     0.71629471         0.26056798
  2          17     0.69856529         0.28777808
  3          17     0.81731235         0.29866926
            T tests (LSD) for variable: RSQPRESS


        Alpha= 0.05  df= 48  MSE= 0.079972
              Critical Value of T= 2.01
        Least Significant Difference= 0.195


    T Grouping              Mean      N  SETNUM


            A           0.81731     17  3
            A
            A           0.71629     17  1
            A
            A           0.69857     17  2
```

Figure 7-9: R^2_{press} Values for Data Sets

Although there was only one grouping for both adjusted R^2 and R^2_{press}, data set 3 definitely had better performance than the other two models. Data sets 1 and 2 had values approximately 0.1 below the values for data set 3 for both measures of performance. Data set 1 performed slightly better than data set 2. Once again, this is not a surprise.

7.4.3   Statistical Issues

A short review of the statistical issues concerning each data set is in order. All three data sets seem to produce adequate results. The statistical concerns regarding the data sets will have an impact on which data set is ultimately chosen.

Data set 1 is composed of all data pairs except for those that are repeated. Data set 1 has issues regarding the independence of data pairs, relative dominance, and interval between data pairs. Data set 2 partially addresses all of these issues, but does not completely solve the problems of independence and relative dominance. Data set 3 goes further to eliminate the independence problems and the relative dominance problems—but other problems are introduced. By using average values, the measures of performance are artificially skewed to appear better and any confidence intervals generated are not valid in the same sense as those produced by the other 2 data sets.

Of the three data sets, data set 2 seems to be the best choice statistically. It is not as statistically pure as data set 4 (which was eliminated for failure to produce acceptable p-values for parameter significance), but it does eliminate at least so of the data dependence and relative dominance of data set 1 without creating additional concerns due to the use of average values.

7.4.4 Sensitivity of β's to Data Set

To aid in the decision-making process, an analysis of the sensitivity of the β values to data set was accomplished. As a start to this analysis, Fisher's LSD was performed on the β values for the three different data sets. The results of these comparisons are depicted in Figure 7-10 and in Figure 7-11.

```
        General Linear Models Procedure

Level of        -------------X-------------
SETNUM     N      Mean               SD


1         17    0.00998818        0.02550438
2         17    0.00853806        0.02455123
3         17    0.00775206        0.02483338
        T tests (LSD) for variable: X


     Alpha= 0.01  df= 48  MSE= 0.000623
           Critical Value of T= 2.68
     Least Significant Difference= 0.023


  T Grouping              Mean     N  SETNUM


        A              0.009988   17  1
        A
        A              0.008538   17  2
        A
        A              0.007752   17  3
```

Figure 7-10: Comparison of β Values for x

Fisher's LSD test was performed at the 99% confidence level for these two tests. It can be seen that the differences in the average β values are slight—all three data sets were grouped together. The variances for the β values for each parameter were nearly identical for all three data sets. The data sets are, after all, different permutations of the same data. It is encouraging that the mean values for the β for x^2 were nearly identical. The β values for x^2 have a much greater impact on the results of the regression than those for x. As will be demonstrated in Chapter 8, the β value for x^2 is the sole determinant of L* for a given fleet of equipment.

```
        General Linear Models Procedure

Level of          --------------X2-------------
SETNUM     N        Mean              SD


1          17      0.00259902      0.00335241
2          17      0.00269420      0.00336620
3          17      0.00274414      0.00349645
        T tests (LSD) for variable: X2
   Alpha= 0.01  df= 48  MSE= 0.000012
           Critical Value of T= 2.68
    Least Significant Difference= 0.0031


 T Grouping              Mean      N  SETNUM


         A            0.002744   17  3
         A
         A            0.002694   17  2
         A
         A            0.002599   17  1
```

Figure 7-11: Comparison of β Values for x^2

The β values were also compared by determining the percentage difference for the β values for each of the two parameters for each of the 17 fleets for each data set. Data set 1 was used as the baseline. On average, the β values for x differed by 5% within each fleet. The β values for x^2 differed by only 4%. A 4% difference in β for x^2 equates to approximately a 500 hour difference in the value of L* for a fleet of machines with a baseline L* of 10,000 hours. This is not that great of a difference.

7.4.5   The Selection

There is a good deal more judgment involved in the decision of which data set to select than there was in the selection of the best model. All three of the data sets still in contention provided adequate p-values for parameter significance. Data set number 3 clearly provided the best measures of performance, but the measures of performance for data sets 1 and 2 were adequate. Data set number 2 addresses most of the statistical problems of data set 1 without introducing new ones like data set 3 does.

The bottom line is that all three of the data sets produce nearly the same results. Because of this, data set 2 is chosen as the best data set to use. Data set 1 had too many unresolved statistical

issues. Data set number 3 created too many new statistical issues—its improved performance in adjusted R^2 and R^2_{press} came at a price. Data set 2, although not statistically pure, is the best choice under these circumstances. Figure 7-12 provides a stark contrast to Figure 7-1 which appeared in the beginning of this Chapter. The large number of regressions in the beginning, 1272, has been brought down to 17.
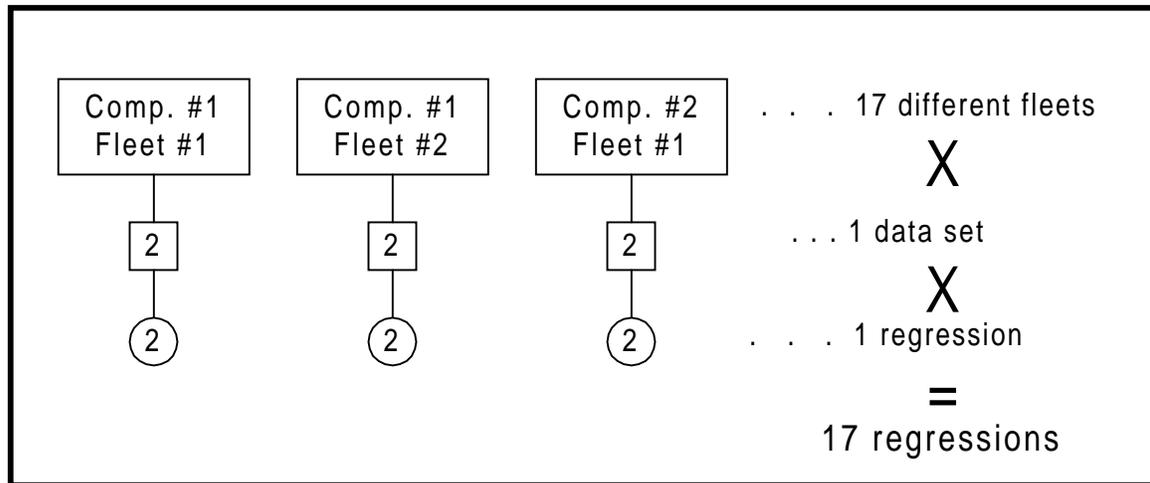
Figure 7-12: Final Model and Data Set Selected

7.5   STATISTICAL PERFORMANCE

With the model and data set selection complete, the model's statistical performance can now be summarized. Areas that will be discussed are measures of performance, model validations, and confidence levels for β's.

7.5.1   Measures of Performance

Seventeen fleets were evaluated. Adjusted R^2 and R^2_{press} values for each of the seventeen fleets are given in Table 7-17. Although the average values for these measures of performance are reasonable, the range of values was quite great. Fleet 4 provided the best fit and prediction with values of over 0.95 for both. Fleet 11 performed horribly with an adjusted R^2 of near zero and an R^2_{press} of less than zero. The reasons for this will be addressed in Chapter 8. It is encouraging that more than half of the fleets had both measures of performance over 0.80. These values could have been made substantially better through the elimination of outlying machines.

However, it is a fact that some machines perform better than average and some perform worse than average. It was felt that machines should not be eliminated from the data sets simply because their repair records were worse or better than others were.

Table 7-17: Measures of Performance for Final Model

fleet	Adj. R^2	R^2_{press}
1	0.52872	0.4051
2	0.87843	0.86282
3	0.88819	0.87788
4	0.9699	0.95012
5	0.85091	0.84178
6	0.84961	0.81754
7	0.6193	0.57946
8	0.93471	0.93192
9	0.80783	0.79892
10	0.94522	0.93949
11	0.0022	-0.0671
12	0.93954	0.93757
13	0.93092	0.92341
14	0.43345	0.40437
15	0.63228	0.61946
16	0.75986	0.74905
17	0.33114	0.30382
average	0.723659	0.698565

7.5.2   Model Validation

Six of the 17 fleets contained enough machines to perform the cross-validation test described in Chapter 5. All six of these fleets passed the cross-validation with p-values of well within the 0.20 limit. Data splitting and the cross-validation process are intended to show how well the models predict values for machines that were not part of the original set of machines for which the equation was developed. Not all fleets had enough machines (more than 17) to allow for data

splitting. To get some sense of how well these fleets perform as far as prediction is concerned the R^2_{press} values for the fleets that successfully cross-validated will be compared to those of the fleets that were not large enough.

R^2_{press} is also a measure of how well a model predicts values for points that are not in the data set—the data are split out one observation at a time instead of a group of machines at once. The average R^2_{press} for the cross-validation fleets was 0.73. The high value was around 0.93 and the low value approximately 0.40. This compares favorably with the average R^2_{press} for all fleets combined, which was around 0.70. From this, it may be inferred that many of the fleets that were not cross-validated would probably have had successful cross-validations had enough machines been present.

7.5.3   Confidence Intervals for β's

Confidence intervals for the β values for each of the 17 fleets were constructed. The levels of the intervals were 95%, 90%, and 80%. After the intervals were constructed, the high and low confidence limits were constructed as percentages of the values of β. These percentages were averaged to come up with the percentages presented in Table 7-18. The averages were constructed for three different levels of adjusted R^2: less than 0.80, 0.80-0.90, and 0.90-1.00. The reason this was done was to see if the confidence intervals decreased with better fitting models. If one were looking for the 80% confidence interval for β_1 (x) for a fleet that had an adjusted R^2 value of 0.78, the interval would be β_1 plus or minus 50%.

Table 7-18: Confidence Intervals for β's

Parameter	adjrsq	95% conf.	90% conf.	80% conf.
	0.90-1.00	127%	106%	82%
x	0.80-0.90	124%	102%	79%
	<0.80	78%	65%	50%
	0.90-1.00	34%	28%	21%
x^2	0.80-0.90	51%	43%	33%
	<0.80	161%	134%	104%

For β_1 (x), the results obtained were almost counter-intuitive. As the quality of fit of the model decreased, the confidence intervals decreased. For β_2 (x^2), the confidence intervals did what was

expected. As the quality of model fit decreased, the size of the confidence interval increased. This makes sense. A better fitting model should yield less uncertainty. Perhaps the reason for the disparity regarding β_1 is the fact that x and x^2 are inter-related. If the uncertainty for one of the parameters goes down, it is possible that uncertainty for the other would go up.

7.5.4   Residual Plots

The plots of the residuals (error terms) versus the regressor values (cumulative hours of use) were studied to see if there was merit to the assumption made in Chapter 4 that the variation in the residuals would be non-constant, increasing with increasing values of the regressor. Viewing the plots validated this assumption. A typical residual plot is shown in Figure 7-13.
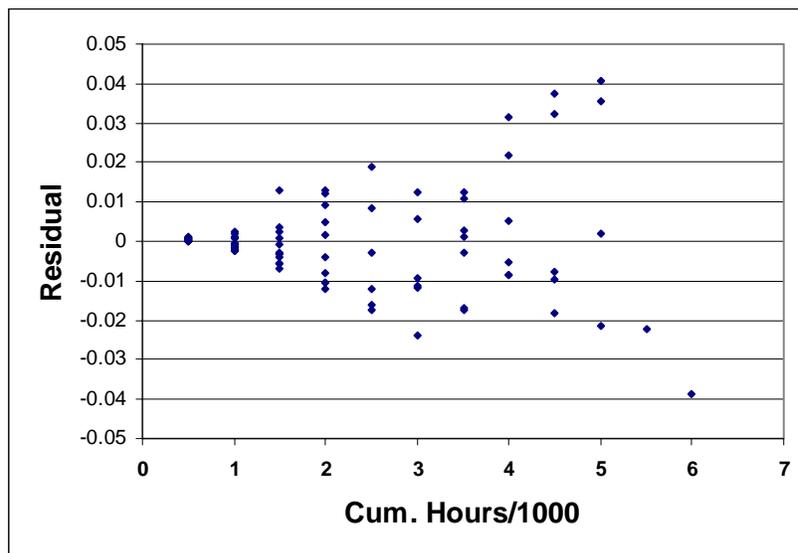
Figure 7-13: Typical Residual Plot

As can be seen in this figure (and the residual plots for most of the other 16 fleets), the residuals seem to be evenly distributed on either side of zero throughout the range of values for the regressor. But, they also show increased dispersion with increasing values of the regressor. This indicates that the variance is, in fact, not constant. Weighted regression would have helped eliminate this problem. Unfortunately, none of the fleets analyzed were large enough to make weighted regression viable.

In to get an idea of what kind of effect weighted regression would have on the results, a weighted regression was performed on the fleet that came the closest to having enough data pairs across the spectrum to adequately describe the variance function. The maximum number of points at any one level was seven. The average number of pairs was 3 per level for the levels that were represented. This is well short of the nine that are recommended (Myers, 1990).

The results indicated that weighting does have a noticeable impact, but not so great an impact as to render the non-weighted results unacceptable. The resulting β coefficients from the weighted and non-weighted regression are shown in Table 7-19.

Table 7-19: Weighted Regression Results

	β_1	β_2	L*	T*
Weighted	0.00881	0.002543	19.82941	0.10967
Non-weighted	0.013697	0.0021	21.82179	0.105349
Difference	55%	17%	10%	4%

From this table, it can be seen that there was a rather large difference (55%) between the weighted and non-weighted β_1 term. The difference for β_2 was only 17%. The differences in measures of performance and parameter significance between the two regressions were negligible. L* and T* values are also listed in this table. The instructions for calculating each of these will be presented in Chapter 8. The difference in L* between the two regressions is around 2, or 2000 hours. In calendar terms, this is around 1 year of operation. The difference in T* is 0.0044—this equates to approximately $0.44 per hour difference in average repair costs per cumulative hour of use for the fleet.

7.6   SUMMARY

In this chapter, the vast amount of data that supports this dissertation was thoroughly analyzed. The models and data sets under consideration were filtered at many different levels. The chapter started with 1272 possible regressions and ended up with seventeen. One model of the nineteen under consideration was chosen as the best. One data set of the four under consideration was chosen as the best.

The final model selected was:

$$CCI = 1 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

This model should be evaluated using the data set composed of data pairs for each machine interpolated at 500-hour intervals.

The final portion of this section of the dissertation will be presented in Chapter 8—Results. While this chapter focused primarily on the statistical aspects, Chapter 8 will attempt to put the results into context.

CHAPTER 8: RESULTS

The purpose of this chapter is give context to the results obtained from the regression model selected. Chapter 7 discussed the analysis in detail, but only began to touch on the presentation of the results. The results are the bottom line of this research. Repair costs associated with construction equipment accumulate and grow with cumulative hours of use. Defining this relationship with equations is one of the main purposes of this dissertation.

In this chapter, the following main areas will be discussed/developed:

- The results and nature of the equations will be discussed

- Sensitivity analyses will be performed on various model components and results

- Comparisons of various fleets will be made

- Comparisons to other forecasting methods will be made

8.1   THE RESULTS

As implied in Chapter 7, the results of this research were quite promising. In most cases, the parameters for the equations were highly significant and the measures of performance were more than adequate. The question that must now be answered is "are the results meaningful?" To address this question, this section will focus on the following areas:

- The equations

- L*

- T*

- L* vs. T*

8.1.1   The Equations

The equations developed for each of the fleets involved in this study were of the form:

$$CCI = 1 + \beta_1 x + \beta_2 x^2 \qquad\qquad\qquad \textbf{Equation 8-1}$$

Where:

CCI = cumulative cost index

x = cumulative hours of use / 1000

β_1, β_2 = coefficients determined by regression

There are two main components that are a part of this equation: the $\beta_1 x$ component and the $\beta_2 x^2$ component. These components are shown in Figure 8-1. It is postulated that the $\beta_1 x$ component should represent the fixed element of repair costs—it provides a baseline measure of how well a company controls essential expenditures on a given fleet of machines. A very low $\beta_1 x$ component could indicate that the company does not spend a lot of money on maintaining and repairing the machine as a part of day-to-day business. A high component could indicate that the company does spend a steady amount of money on the fleet throughout the life of the machines.
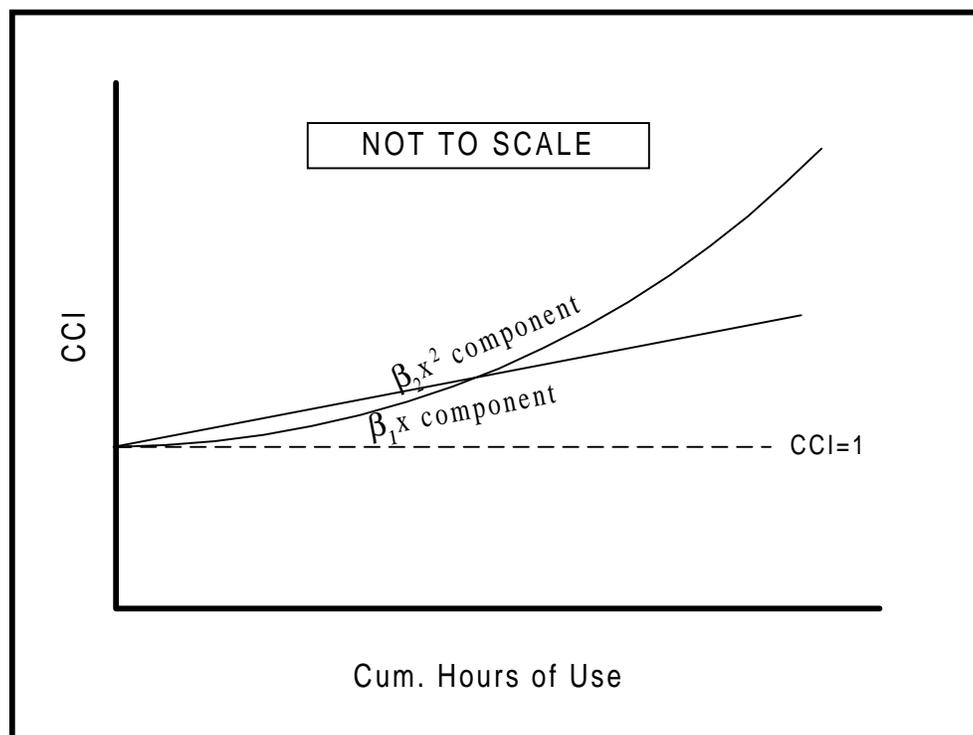
Figure 8-1: The Two Cost Components

The $\beta_2 x^2$ component represents how well the company controls the *growth* of costs. A large component signifies that costs grow rather quickly. A small component indicates that the

company does a good job of keeping cost growth down. The growth of costs is the determinant of L* for the fleet—this will be discussed in greater detail in section 8.1.2.

Table 8-1: Rankings of Values of Cost Components

FLEET	$\beta_1 x$	$\beta_2 x^2$
Wheel Loaders	17	1
Large Dozers	16	2
Large Dozers	15	3
Small Excavators	14	4
Large Dozers	13	5
Dual-engine Scrapers	10	6
Large Dozers	7	7
Mid-size Excavators	8	8
Track Loaders	5	9
Mid-size dozers	12	10
Wheel Loaders	11	11
Articulated Trucks	9	12
Articulated Trucks	6	13
Articulated Trucks	4	14
Small Excavators	1	15
Mid-size dozers	2	16
Mid-size dozers	3	17

An interesting observation is that there is a tendency for fleets with lower $\beta_1 x$ components to have higher $\beta_2 x^2$ components. The values of the coefficients for x and x^2 for each of the fleets were rank ordered from 1 to 17, lowest to highest. The results of these rankings are posted in Table 8-1. Although there are a couple of exceptions, for most of the fleets a high ranking in the x component resulted in a low ranking in the x^2 component. From this it could be inferred that companies that invest in maintenance and repair throughout the lives of their machines experience lower cost growth, and hence longer economic lives for their equipment than those companies that do not invest in the early maintenance and repair of their fleets.

The values of the coefficients were plotted to determine if the relationship between them could be quantified. This plot is shown in Figure 8-2. The regression line that is plotted in the figure highlights the relationship between the two coefficients. The line is a 2nd order polynomial with an R^2 value of 0.785. This is fairly significant. This topic should be revisited in an expanded study of equipment data. One fleet was eliminated from the data to come up with this plot—the first

fleet listed in Table 8-1. This fleet was highly influential on the regression and caused the curve to reach a minima and curve abruptly upward.
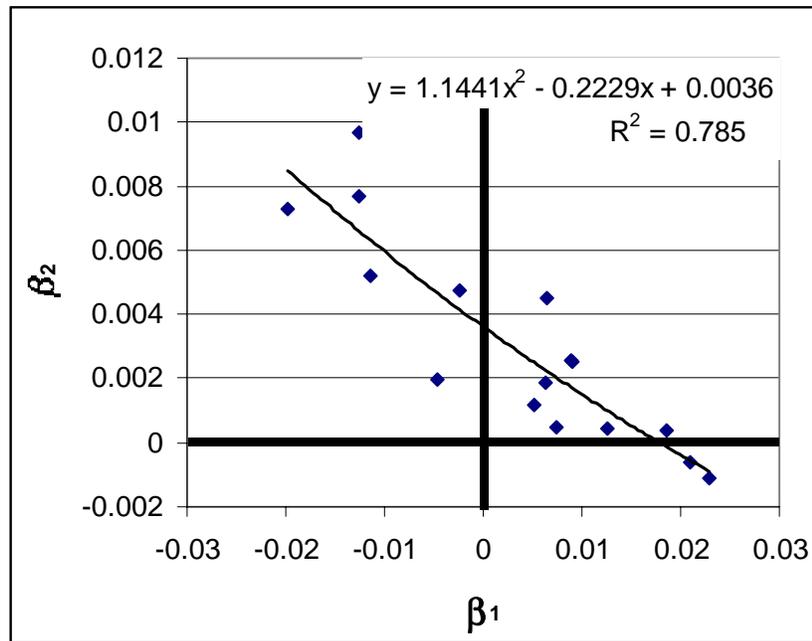

Figure 8-2: Plot of β_1 vs. β_2

In Table 8-2 the coefficient values for each of the fleets are presented. The first important thing to look for is the sign of the β_2 coefficients. If these coefficients are negative, an optimum solution for economic life and repair costs cannot be determined from the equation. This is illustrated in Figure 8-3. Line "A" represents what was postulated at the beginning of this dissertation. The slope of the cumulative cost curve increases with increasing cumulative hours. Because of this, the optimum values L* and T* can be found both geometrically and mathematically (this will be demonstrated in sections 8.1.2 and 0). Line "B" represents a curve with a negative β_2 coefficient. On a curve such as this, the "optimum" is not reached because the tangent cannot be drawn. The machines theoretically have an infinite L*.

Three of the seventeen fleets had negative coefficients for x^2. Two of these fleets were large dozers. The slight negativity of the curve could be explained by management styles of the company involved. The dozers are used in less stressful applications as they accumulate hours— this helps to cut down on the growth of costs. An explanation of how optimization for fleets like these is still feasible will be offered in section 8.1.4.

Table 8-2: β values for the 17 Fleets

Type	β_1	β_2
Articulated Trucks	-0.00246	0.004753
Articulated Trucks	0.006392	0.004496
Articulated Trucks	-0.0115	0.005175
Dual-engine Scrapers	0.007391	0.000491
Large Dozers	0.005168	0.001169
Large Dozers	0.012567	0.000446
Large Dozers	0.020905	-0.00063
Large Dozers	0.022861	-0.00113
Mid-size dozers	-0.01252	0.009651
Mid-size dozers	-0.01256	0.007659
Mid-size dozers	0.009023	0.002529
Mid-size Excavators	0.006304	0.001893
Small Excavators	-0.01978	0.007303
Small Excavators	0.018518	0.000389
Track Loaders	-0.00471	0.001963
Wheel Loaders	0.00881	0.002543
Wheel Loaders	0.09075	-0.0025
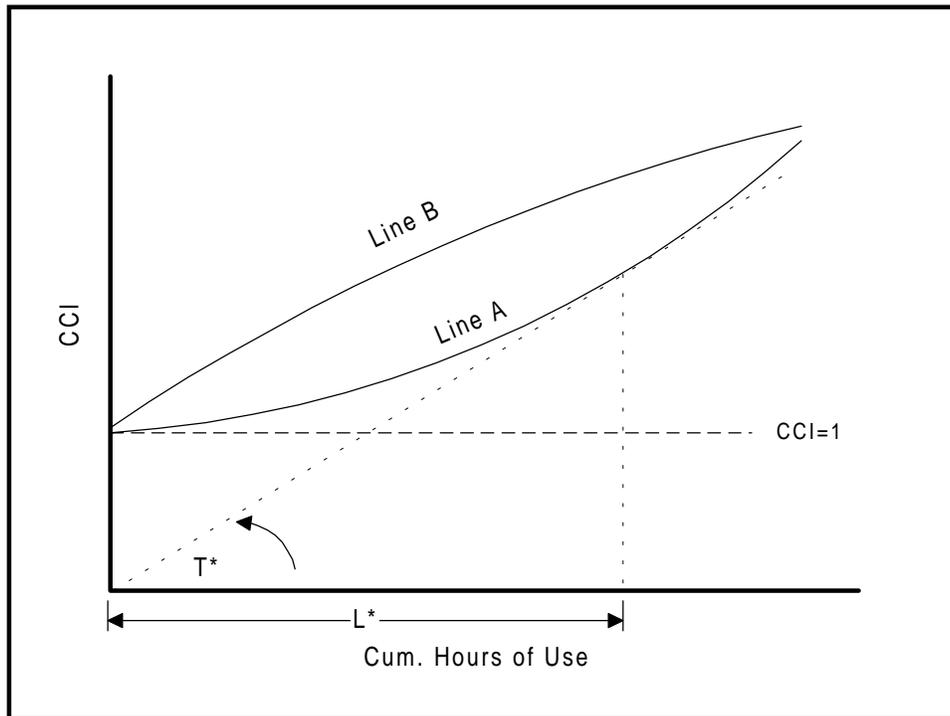
Figure 8-3: Effect of Negative β_2 Term

The problems with the third fleet, wheel loaders, require an explanation of the data used for the regression. The fleet was one of the smallest analyzed at 6 machines. The range of operation for which data were available for these machines was between 13,000 and 24,000 cumulative hours of use. No data were available for the earlier ranges of operations. Typically, the early ranges of operations have low repair costs. When the data pairs with these low costs are placed in the same regression with data pairs of higher hours and higher costs, an upwardly curved line results. Because of these two problems, small sample size and incomplete range, the equation developed for this fleet may not be reliable. Not surprisingly, this fleet also had the worst measures of performance as discussed in Chapter 7. The adjusted R^2 value was 0.0022 and the R^2_{press} value was –0.0671.
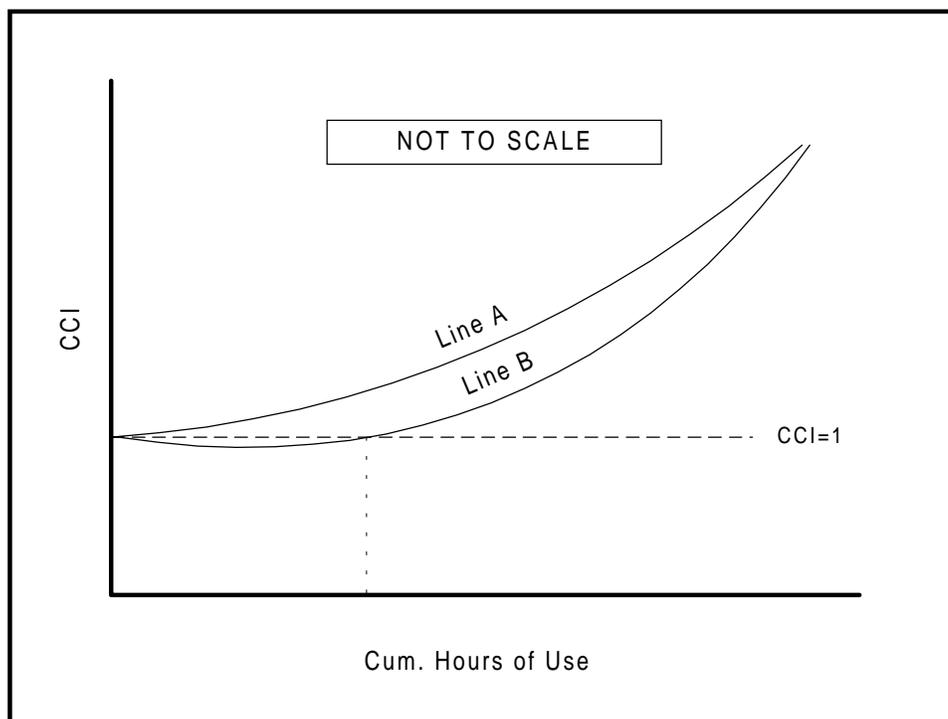
Figure 8-4: Effect of Negative β_1 Term

A second thing that should be looked at in the parameters is the sign of the β_1 term. If this term is negative, there will be a certain range of cumulative hours for which the equation will predict negative repair costs—which is not possible. This is illustrated in Figure 8-4. Line "A" represents an equation with a positive β_1 term; line "B" represents an equation with a negative β_1 term. There were six fleets that had negative values for β_1.

 It can be seen that line "B" dips slightly below the CCI=1 line during the early hours of cumulative use for the fleet. The line then recovers above CCI=1 and predicts positive repair costs. The average number of hours for these six fleets to get to positive repair costs was 1,800 hours. The longest it took for one of these fleets to get to positive repair costs was 2,700 hours. Although it is not ideal to have negative repair costs predicted for any portion of a machine's life, the range of use affected by this problem is small and not critical. Many of the repairs that take place during that range are covered by warranty. As will be demonstrated in section 8.1.2, the β_1 term has no impact on L*. But, most of the fleets with negative β_1 terms had large β_2 terms (see Table 8-1) which do have an impact on L*.

8.1.2   L*

The optimum length of time to operate a fleet of equipment based on optimizing for the lowest average costs is L*. This is depicted in Figure 8-5. It is defined by a tangent line drawn from the origin to the cumulative cost curve.   Graphically, it is very easy to understand L*. Mathematically, the solution is also fairly straightforward.
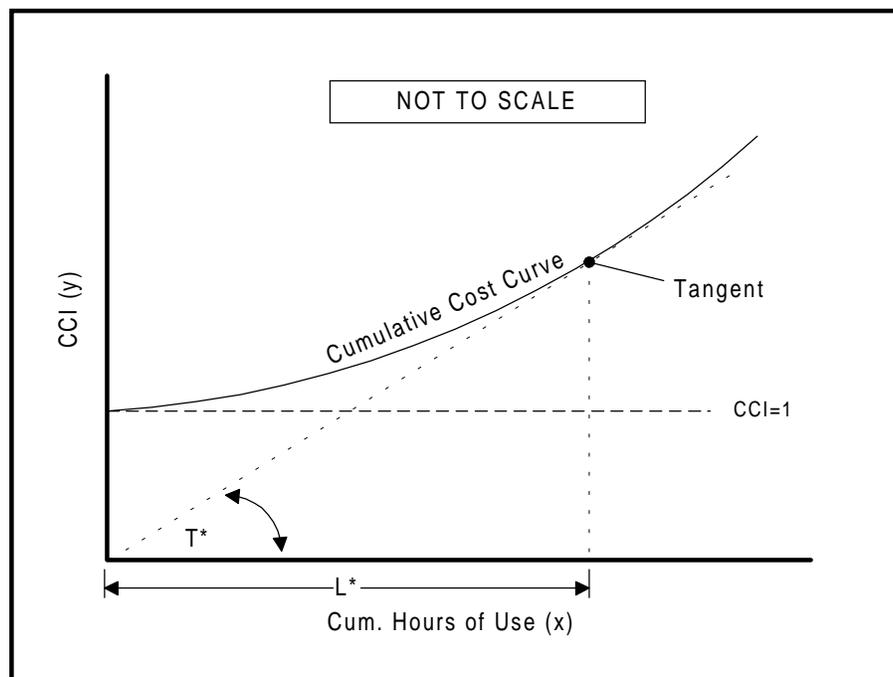

Figure 8-5: L* and T*

Using Figure 8-5, the derivation of L* will be performed:

The equation for the cumulative cost curve is defined by equation 8-1:

$$CCI = 1 + \beta_1 x + \beta_2 x^2$$

The equation for the tangent line is:

$y = mx$ **Equation 8-2**

Where:

m = slope of the tangent line

y = vertical component along the CCI axis

Set these equations equal to each other at the tangent point:

$mx = 1 + \beta_1 x + \beta_2 x^2$ **Equation 8-3**

Differentiate with respect to x:

$m = \beta_1 + 2\beta_2 x$ **Equation 8-4**

Substitute equation 8-4 into equation 8-3:

$(\beta_1 + 2\beta_2 x)x = 1 + \beta_1 x + \beta_2 x^2$ **Equation 8-5**

Simplify equation 8-5:

$\beta_2 x^2 = 1$ **Equation 8-6**

Solve for x:

$$x = \frac{1}{\sqrt{\beta_2}}$$ **Equation 8-7**

L* is the length of time from the purchase of the machine to the tangent point defined in equation 8-7. The solution is simple and clean. L* is solely a function of the growth of costs. L* values for the fleets analyzed are given in Table 8-3. Note that only 14 fleets are represented on this table. The 3 fleets that could not be optimized were removed from the table. The fleets are listed in order of decreasing L*--the units for L* are cumulative hours/1000. The L* values in this table seem reasonable for many of the machines. If the values are in error, they seem to be in error on the high side versus the low side. This is possibly due to the absence of collateral costs. This will

be addressed further in section 8.1.4. Three possible exceptions are the first three in the table. The large dozers and dual-engine scrapers are long-lived machines, but it is doubtful that their optimum overall costs occur at 45,000 hours of use—this topic will be addressed further in section 8.1.4.

Table 8-3: L* and T* for Fleets Analyzed

	Fleet	L*	CCI @ L*	T*
A	Small Excavators	50.70853	2.939021	0.057959
	Large Dozers	47.35137	2.595065	0.054804
	Dual-engine Scrapers	45.13396	2.333585	0.051704
	Large Dozers	29.24527	2.15114	0.073555
	Mid-size Excavators	22.98517	2.144898	0.093317
	Track Loaders	22.57331	1.893589	0.083886
	Mid-size dozers	19.88461	2.179419	0.109603
	Wheel Loaders	19.82941	2.174697	0.10967
	Articulated Trucks	14.91375	2.095329	0.140496
	Articulated Trucks	14.50433	1.964276	0.135427
	Articulated Trucks	13.90096	1.840153	0.132376
B	Small Excavators	11.70171	1.768505	0.151132
	Mid-size dozers	11.42674	1.856434	0.162464
	Mid-size dozers	10.17936	1.872565	0.183957

A general observation is that the fleets of smaller machines usually had smaller L* values than similar large machines. All of the mid-size dozers had shorter L*'s than the larger dozers. This could be due to many causes. Smaller machines sometimes serve as jacks-of-all-trades, doing a wide variety of jobs. Large machines are usually employed in a more static situation. The frames and components on larger machines have more metal in them and should hold up to greater stresses. If a small machine is used in an application for which a large machine should be used, the small machine will probably (in addition to having less productivity) have more breakdowns because the machine is at the upper end of its limits instead of being right in its designed operating range. Yet another reason that small machines reach L* sooner could be the cost of labor. Parts for larger machines cost proportionally more than parts for small machines. But, the labor charges involved with changing the parts are relatively constant. There could be additional labor because the larger parts are heavier and could require expensive equipment to manipulate, but this could be balanced out by the fact that larger machines allow more room for mechanics to work. More room to work can enhance the mechanics' productivity.

An important exception to this rule was one fleet of small excavators (see "A", Table 8-3) that had a much larger L* than the other fleet of small excavators—"B". This fleet (like the fleet of wheel loaders with the negative β_2 term) had a very large gap in coverage of the spectrum of cumulative hours of use. There were only two machines with data pairs below 8,500 hours and no machines with data pairs in the range 8,500 to 18,500 hours. This is a critical range of hours—especially for smaller machines. The lack of data in this range could have produced unreliable results.

8.1.3   CCI and T*

The cumulative cost index (CCI) is the main result of the regression equations developed. The CCI at L* for each of the fourteen fleets which could be optimized is reflected in Table 8-3. Although is appears that machines with high L* values have higher CCI values at L*, the range of values for these CCIs is small compared to the range of values for L*. L* has a mean value of 23.88 with a standard deviation of 13.99. The corresponding CCI values had a mean of 2.13 with a standard deviation of 0.32. This tightness of this distribution indicates that it may be possible to derive an empirical rule for CCI at L*. It seems that the number 2 would be a good starting point for this empirical rule. *When the initial purchase price of the machine has been spent on maintenance and repair (CCI = 2), the machine is very close to L*.* Further research is warranted to validate this rule.

The lowest average cost for a fleet is achieved when L* is reached. This cost is T* (see Figure 8-5). The equation for calculating T* is simply the CCI at L* divided by L* (the slope of the tangent). The equation for this is:

$$T* = \frac{1 + \beta_1 L* + \beta_2 L*^2}{L*} = \beta_1 + 2\sqrt{\beta_2} \qquad \textbf{Equation 8-8}$$

T* values for the fourteen fleets with non-negative β_2 terms are given in Table 8-3. The units for T* as presented are 1/(cumulative hours of use/1000). The dollar portion of units is not present because the CCI is a ratio of dollars to dollars. There appears to be a definite relationship between L* and T*. This will be explored in the section 8.1.4.

8.1.4   L* vs. T* Curve

The L* vs. T* relationship is depicted in Figure 8-6. A regression line of the form $y = ax^b$ was fit to the curve. The line fit with an R^2 value of 0.9755. By definition, T* is related to L* (see equation 8-8). Since L* is in the denominator, T* should decrease with an increase L*. But, there is not a direct inverse proportionality to the relationship because the equation for CCI is in the numerator. The β_1 coefficient has an effect on T* in the numerator. Conceivably, this coefficient could produce T* values that would be randomly scattered around the plot with very little correlation. But, it was demonstrated in section 8.1.1 that there is a reasonably strong relationship between β_1 and β_2. This relationship manifests itself in a very strong relationship between T* and L*.
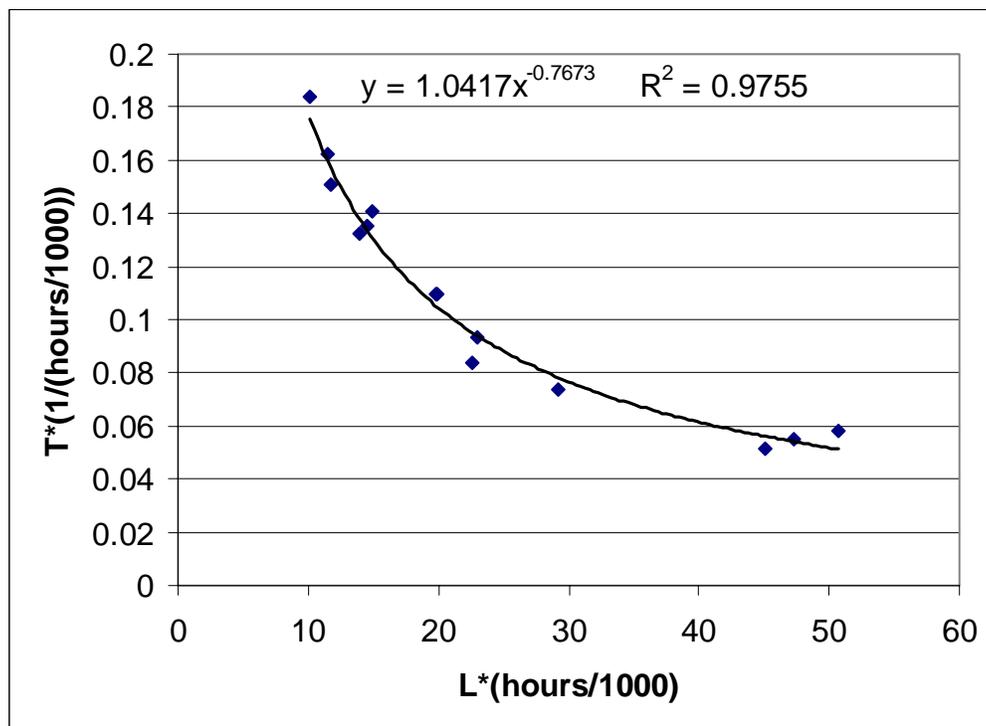
Figure 8-6: L* vs. T* plot

At this point, some observations about the L* vs. T* plot are in order. First, it seems like there is a continuum of L*/T* values along which most fleets will lie. A broader study with more data could further solidify and define this relationship.

Second, it seems like the L* values which were lower reflect the actual nature of when equipment is replaced more accurately. Perhaps the reason for this is the lack of inclusion of collateral costs. Very few construction companies that buy and operate new machines have machines with as much use as 45,000 hours. There must be some reason other than the growth of repair costs for selling these machines before they accumulate 45,000 hours. The answer probably lies in a combination of the collateral costs described in Chapter 4. The costs associated with more frequent and longer breakdowns make it less economically feasible to keep such machines in a production fleet.

Two of the fleets which were eliminated due to negative β_2 terms and one of the fleets which had a very high L* were large dozers. The company that operates these machines starts them out in very stressful applications, such as ripping. As the machines grow older and become less reliable, they are relegated to less stressful applications, such as pushing scrapers. This deflates the growth of repair costs because the nature of the machines' uses changes. It is postulated that if the collateral costs for these machines were tabulated and incorporated into the regression equations, the two fleets eliminated would re-enter the fold and the fleet with the high L* would have a lower L* based on the inclusion of the additional costs.

Another aspect of this observation is that it seems that collateral costs would play a less significant role in the determination of true L* for the fleets that have smaller L* values due solely to repair costs. In these cases, it may be possible to neglect collateral costs when making decisions concerning economic life. It seems that this assumption may hold true for fleets that have L* values of less than 20 (20,000 hours).

8.2   SENSITIVITY ANALYSES

It is important to determine how sensitive the results presented in section 8.1 are to various aspects and variables in the study. The sensitivity analyses to be presented are:

- L* to β terms
- T* to β terms

8.2.1   L* to β's

One of the primary uses of the cumulative repair cost equations within the cumulative cost model will be to determine economic life. If a machine is bought or sold at the wrong time, money can be lost. L* is equivalent to the DMCL as defined in Chapter 3. As β_2 varies, L* will vary. Variations in β_1 have no effect on L*. This sensitivity analysis will look at hypothetical machines that have predicted L* values of 10, 15, 20, 25, and 30. This covers a wide range of the results obtained during this research. The results are depicted in Table 8-4.

Table 8-4: Sensitivity of L* to β_2

L* Initial	New L* due to increase in β_2					
	5%	**10%**	**25%**	**50%**	**75%**	**100%**
10	9.759001	9.534626	8.944272	8.164966	7.559289	7.071068
15	14.6385	14.30194	13.41641	12.24745	11.33893	10.6066
20	19.518	19.06925	17.88854	16.32993	15.11858	14.14214
25	24.3975	23.83656	22.36068	20.41241	18.89822	17.67767
30	29.277	28.60388	26.83282	24.4949	22.67787	21.2132
	New L* due to decrease in β_2					
	5%	**10%**	**25%**	**50%**	**75%**	**100%**
10	10.25978	10.54093	11.54701	14.14214	20	∞
15	15.38968	15.81139	17.32051	21.2132	30	∞
20	20.51957	21.08185	23.09401	28.28427	40	∞
25	25.64946	26.35231	28.86751	35.35534	50	∞
30	30.77935	31.62278	34.64102	42.42641	60	∞

It can be seen that negative changes in the coefficient β_2 can have a much greater impact on L* than positive changes. This is encouraging because the coefficients for growth of cost are probably slightly underestimated due to the absence of collateral costs. Positive and negative changes of 10% or less are not that bad. The maximum that L* is off for changes of 10% is 1.62, or around 1,600 hours of use. Most machines work this much in less than a year. Above 50% change, the results are probably unacceptable. The maximum that L* is off at 50% is 12.42, or approximately 12,500 hours—this could represent 5 or more calendar years for some types of fleets.

8.2.2   T* to β's

It is important to look at the sensitivity of T* because the T* values can be used to help forecast maintenance and repair costs. If the forecasts are off, the company could find itself with excess money that could have been invested elsewhere, or, worse still, with not enough money to pay the bills. Unlike L*, T* is sensitive to both β terms. This sensitivity analysis will look at T* values ranging from 0.05 to 0.15. For analyzing β_1, β_2 was held constant at 0.0016 (L* = 25). For analyzing β_2, β_1 was held constant at 0.01. The selection of these fixed constants was determined so that they corresponded to a realistic point that could be a part of the plot of the coefficients in Figure 8-2. The results of this analysis are depicted in Table 8-5.

Table 8-5: Sensitivity of T* to β Terms

T* Initial	New T* due to increase in β_1					
	5%	10%	25%	50%	75%	100%
0.05	0.0515	0.053	0.0575	0.065	0.0725	0.08
0.10	0.101	0.102	0.105	0.11	0.115	0.12
0.15	0.1535	0.157	0.1675	0.185	0.2025	0.22
	New T* due to decrease in β_1					
	5%	10%	25%	50%	75%	100%
0.05	0.0485	0.047	0.0425	0.035	0.0275	0.02
0.10	0.099	0.098	0.095	0.09	0.085	0.08
0.15	0.1465	0.143	0.1325	0.115	0.0975	0.08
	New T* due to increase in β_2					
	5%	10%	25%	50%	75%	100%
0.05	0.050987	0.051952	0.054721	0.05899	0.062915	0.066569
0.10	0.102222	0.104393	0.110623	0.120227	0.129059	0.137279
0.15	0.153457	0.156833	0.166525	0.181464	0.195203	0.20799
	New T* due to decrease in β_2					
	5%	10%	25%	50%	75%	100%
0.05	0.048987	0.047947	0.044641	0.038284	0.03	0.01
0.10	0.097721	0.095381	0.087942	0.07364	0.055	0.01
0.15	0.146455	0.142816	0.131244	0.108995	0.08	0.01

Putting T* into perspective may help with the understanding of this analysis. A T* of 0.05 corresponds to to an average of 5% of the purchase price of the machine being spent every 1000 hours of operation. For a $100,000 dollar machine, this corresponds to an average cost of $5.00 per hour of operation. With a T* of 0.15, that same machine would have an average cost of

$15.00 per hour. So, a change in T* of 0.01 corresponds to a $1.00 per hour increase in the average cost of a $100,000 machine to get to L*.

Increases in β_1 result in increases of T*. Decreases in the coefficient result in decreases of T*. The same was true for changes in β_2. These increases and decreases become more pronounced as the percentage of change goes over 10%. Below 10% change, the magnitude of the increases and decreases are small. Above 25%, the magnitudes can result in noticeable differences in the T* values (more than $1.00 per hour change for a $100,000 machine).

8.3   COMPARISONS

A number of comparisons will now be made to determine what, if any, conclusions and generalizations can be made concerning how the equations of the different fleets relate to one another. The comparisons will be of both an objective and subjective nature. Only the 14 fleets that had positive β_2 values will be used in the comparisons. Statistical procedures will be used where appropriate. Generalizations will be used when they are needed. The results presented in this section are by no means definitive. They are open to interpretation. The comparisons to be performed fall into three major categories:

- comparisons of different fleets within the same companies
- comparisons of similar fleets across different companies
- Comparisons of fleets of the same category but differing size

The specific comparisons made are shown graphically in Figure 8-7. There are four company comparisons, four similar fleet comparisons, and three size comparisons.

	Company			
	A	B	C	D
Artics 25T	x	x		
Artics 35T		x		
Dozer, Medium	x	x	x	
Dozer, Large				2
Excavator, Small		x	x	
Excavator, Medium		x		
Loader, Wheel		x		
Loader, Track			x	
Scraper, 2-eng.				x
Company	Similar Fleet	Size		

Figure 8-7: Comparisons

8.3.1   Company

There were four different companies represented in the 17 fleets that were analyzed. The companies had differing management styles and differing data collection methods. Each of the companies had more than one fleet in the study, so a comparison within each company is possible.

A reasonable test to compare regressions is direct substitution and validation. Mechanically, the procedure is identical to the cross-validation procedure described in Chapter 5. All fleets except for one within each company will be regressed together and the comparisons will be made. The process will be repeated so that each fleet within a company is cross-validated against the rest of the fleets in the company. This is a total of 14 tests—one for each fleet. All of these cross-validations were 99% significant or better. This implies that the company from which the fleet came is important.

Since the cross-validations were successful, an additional test that was done was a regression on all available machines within each company. The results of these regressions are shown in Table 8-6.

**Table 8-6: Comparison Regressions**

Comparison	L*	T*
company A	14.37696	0.137379
company B	17.96923	0.118976
company C	28.47859	0.081543
company D	70.18624	0.039822
medium dozers	23.8705	0.11172
artics	13.43402	0.145268
small excavators	58.12382	0.055321
large dozers	129.8581	0.028842
excavators: size	12.27387	0.146417
dozers: size	21.29589	0.109068
artics: size	11.82294	0.153669

There was substantial variation in the L* and T* values between the different companies. Some of this could be attributed to the types of machines analyzed, but companies A, B, and C all had relatively the same types of equipment involved in this study. Company D had substantially larger machines than the other three companies had. The L* for company A was approximately 50% of the L* for company C.

For company A, the L* and T* forecasts using the company model worked out well for one fleet. For the other fleet, the forecast L* was 4000 hours off and the forecast T* was off by over $5.00 per hour per $100,000. For company B, L* was off by more than 5000 hours for 4 of the 6 fleets analyzed. T* fared better—3 of the 6 fleets were within $3.00 per $100,000 of their forecast T* values. Due to the wide range of L* values for company C's individual fleets (19.88 to 50.70), none of the three fleets in this company were within 5000 hours of the forecast L* for the company. Two of the three fleets in company C were within $2.00 per $100,000 for their T* values.

The results obtained using company D's fleets were perplexing. The forecast L* for the company was 25,000 hours greater than the largest individual L* of the three fleets analyzed. The forecast T* for the company was more than $1.00 per $100,000 cheaper than the cheapest individual fleet.

Although a relationship among the fleets of a particular company is supported by the successful cross-validations, it is believed that this relationship is not strong enough to warrant using one equation for each company. When equations are fit to the combined data, the results are different than those obtained using individual fleets. It is not possible to say which is right and which is wrong based strictly on numbers. If the fleets are looked at by company on an L*/T* plot (see Figure 8-1), it appears that there is a loose relationship among the fleets of the various companies. It is stronger for some companies than it is for others. It is recommended that equations for individual fleets be used over company-wide equations if a choice exists.
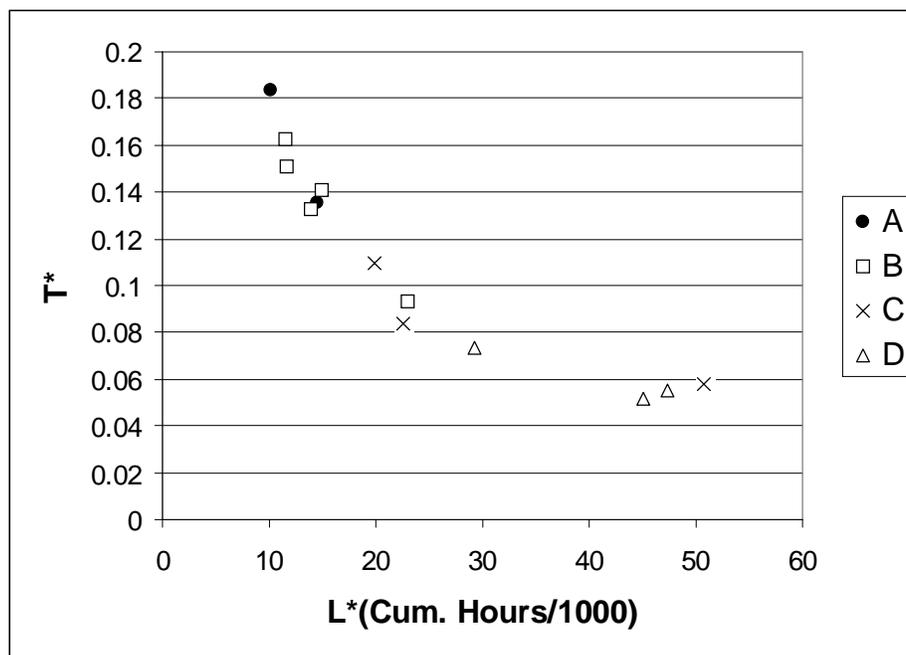
Figure 8-8: L* vs. T* Companies

8.3.2 Machine Type

There were four different types of fleets that had multiple representation in the 14 fleets that had acceptable β_2 values. These were: medium dozers, articulated trucks, small excavators, and large dozers. The two fleets of large dozers came from the same company. One fleet was composed of slightly larger machines than the other, but they were used for essentially the same applications.

Once again, the cross-validation process was used. The results were significant at better than a 90% level for all comparisons (note that this is not as significant as the tests performed on the

companies). Once again, regressions were performed on all the available machines within each fleet type to form one combined regression equation. The results of these regressions are shown in Table 8-6.

The results for the medium dozers, small excavators, and large dozers were of the same nature as those obtained for company D—disappointing. The forecast L* was 4000 hours greater than the largest individual L*. It was more than 10,000 hours greater than the L*'s for two of the three fleets involved. The T* value was within $1.00 per $100,000 for one fleet, but more than $5.00 per $100,000 off for the other two. The L* values for small excavators and large dozers were also excessively high—the T* values were excessively low.

The results obtained for the articulated trucks were more promising. Both fleets were within 1500 hours of the forecast L* and within $1.00 per $100,000 for the forecast T*. But, the errors were once again in the same direction. This time the combined model gave lower forecasts for L* and higher forecast for T*(unlike for the other three equipment types discussed above).

Viewed graphically (Figure 8-9), the disparities among some of the equipment types are apparent. Large dozers and small excavator data points are widely separated. But, the medium dozer data points were not separated by as great a distance as the combined regression implied. The articulated truck fleets are obviously closely grouped. The third fleet of slightly larger articulated trucks is also shown for comparison purposes. There is a very tight grouping between these three fleets.

Once again, equations developed for individual fleets are recommended over equations developed for equipment types. The three poorly performing models outweighed the one that performed quite well. The model for the articulated trucks shows that the idea of standardized regression equations for equipment types should not be discarded. Further research with more data may allow for the reliable formulation of such equations.
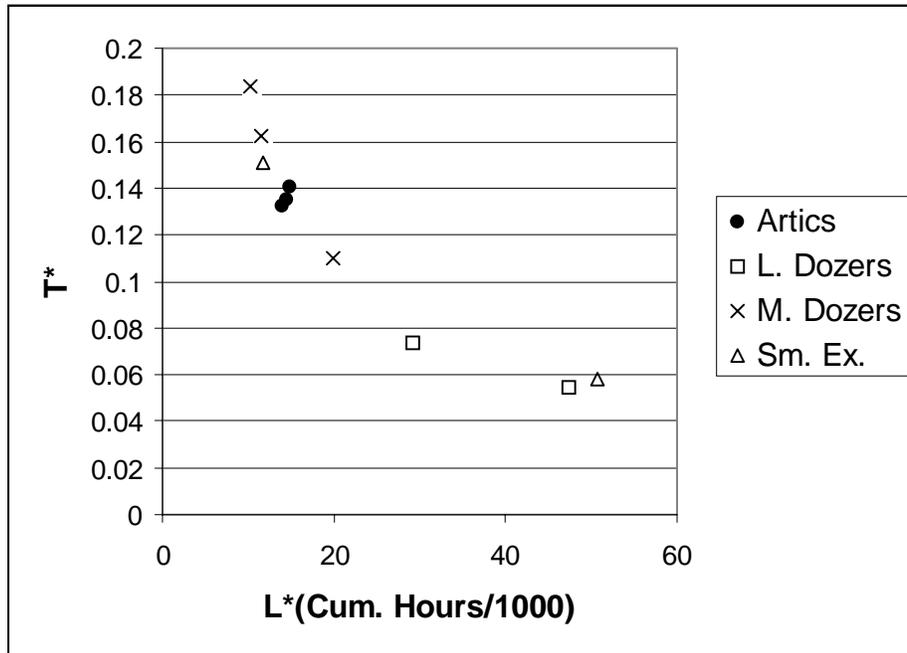
**Figure 8-9: L* vs. T* for Equipment Types**

8.3.3   Machine Size

Three comparisons of machine size were made. They were for dozers, articulated trucks, and excavators. The cross-validation tests for all three of these size comparisons were successful at greater than 95% confidence. Regression equations were formed to evaluate the tangible results of the combination equations. The L* and T* values for these equations are shown in Table 8-6.

For the dozers, the predicted L* for the combined data fell directly in between the L* values for the large and medium fleets. All of the medium dozer fleets had L* values less than the combined value and all of the large dozer fleet had L* values greater than that of the combined fleet. The exact opposite was true of T*--the medium dozer T* values were greater than the combined value and the large dozer T* values were less than the combined value. The exact same relationships were true of the comparison between the small and medium excavators. The combined L* and T* values for the these two size comparisons did not fit any of the individual fleets very well.

For the articulated trucks, the results were a little different. The combined L* and T* values were lower and higher, respectively, than each of their individual fleet counterparts. This leads to an interesting observation. For equipment sizes where the L* and T* differences between the

individual fleets is noticeable, the L* and T* components of the combined equations tended to seek middle ground. When the L* and T* components of the individual fleets were closely related, the components of the combined fleet skewed off in one direction.

When looked at graphically, the differences between the excavators and the dozers are obvious (Figure 8-10). The articulated truck fleets are very close together. A generalization that can be made is that larger equipment tends to have large L* and small T* values.
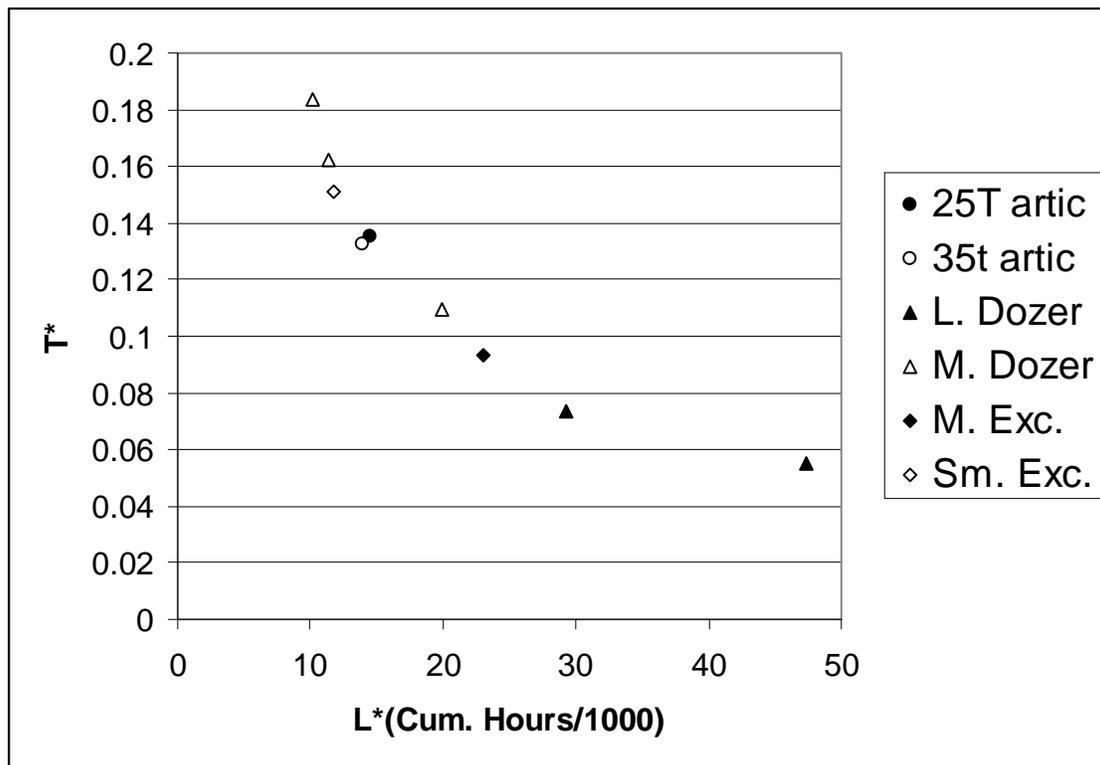
Figure 8-10: L* vs. T* for Size Comparisons

The combined equations for equipment size should not be used for forecasting L* and T*. There are obvious differences between some of the sizes of equipment which are not apparent when using the cross-validation procedure.

The cross-validation test simply determines whether or not it is feasible that a given set of points could be part of a specified regression equation. Since all of the equations point in roughly the same direction, some with more slope than others, all of the data pairs should fall along that path. Figure 8-11 shows the data that are part of this study. They all follow roughly the same trend.

The regression equation composed of all the data combined had an R^2 value of around 0.80. The L* was approximately 30 (30,000 hours) and the T* was 0.0816. This L*--T* pairing falls almost squarely on the L* vs. T* continuum depicted in Figure 8-6. But, an all-encompassing equation does not provide an adequate forecast of when machines will reach the point when average repair costs are optimized. One equation can provide a rule of thumb, but equipment managers need to have more precise forecasts in order to buy, operate, and sell their construction equipment in the manner that will be the most economically advantageous to their companies.
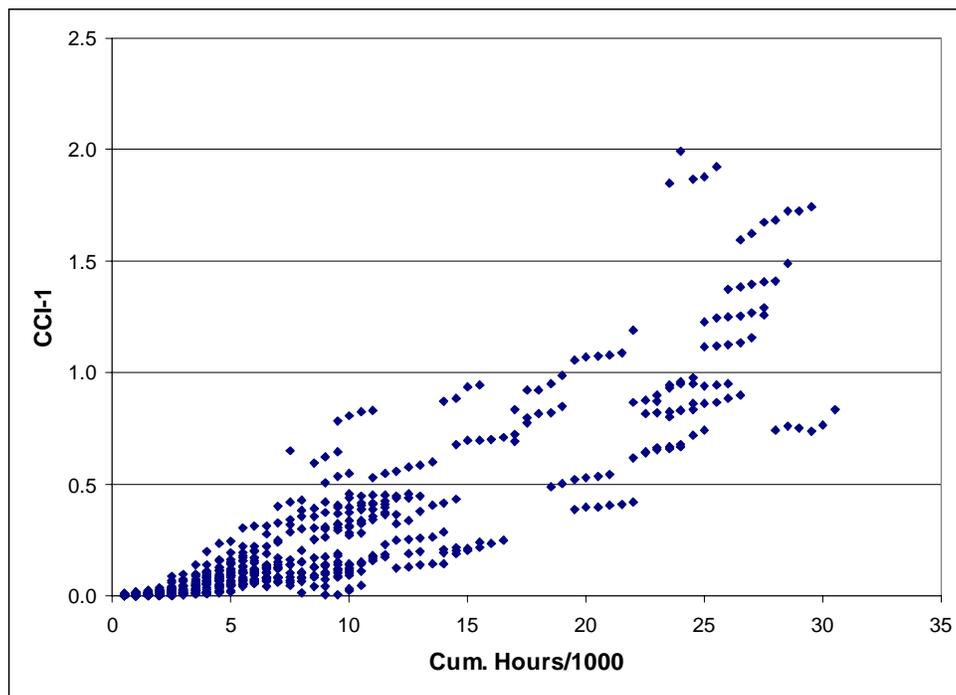
Figure 8-11: Scatterplot of Data Set 2

8.4   PERFORMANCE VS. OTHER METHODS

The final portion of this chapter will investigate the differences between the equations developed in this study and some of the other repair cost forecasting methods described in literature. The three methods that the cumulative repair cost equations will be evaluated against are:

- The Nichols Method

- The Nunnally Method

- Straight Line Methods

8.4.1   Nichols

As discussed in Chapter 2, Nichols (1976) proposed a method of estimating equipment repair costs that made use of a wide variety of different factors. There are factors for type of equipment, total hours of use, years of useful life, temperature, work conditions, maintenance, type of service, operators, experience, equipment quality, and work pressure. Nichols' factors for hours of use separate his method from the others presented in literature as one that directly accounts for different average repair rates for machines that are kept for different cumulative hours.

Table 8-7: Nichols' Factors

Category	Description	Factor
Type of Equipment	End Loader, 4WD	1.0
Total Hours of Use	20,000	3.0
Years of Useful Life	13	1.76
Temperature	Normal	1.0
Work Conditions	Average	1.0
Maintenance	Average	1.0
Type of Service	Contractor	1.0
Operators	average	1.0
Experience	Average	1.0
Equipment Quality	Average	1.0
Work Pressure	Average	1.0

To compare this method with the equations developed, the fleet of wheel loaders with an L* of approximately 20 (20,000 hours) will be used. The average inflation-adjusted purchase price of these machines was approximately $100,000. The machines worked approximately 1500 hours per calendar year—this equates to a calendar lifespan of 13 years. This yields the factors given in Table 8-7.

Multiplying these factors together yields a combined factor of 5.28. This is multiplied by 1/10,000 the purchase price of $100,000 to come up with an average repair cost of $52.80 per hour. Using the regression equation, T* for this fleet was 0.10967 which equates to $10.96 per hour average costs to get to L*. This $10.96 includes the average cost of ownership. This cost must be subtracted out in order have only the average repair cost. To do this, simply divide the purchase price by the number of hours of operation. In this case, $100,000/20,000 hours or $5.00 per hour. This means that the average repair costs are $5.96 per hour. These average repair

costs differ from those of the Nichols method by nearly a factor of 10. These costs are not comparable.

The L* of 20,000 was based on extrapolation for this particular set of data. To compare these two methods in a non-extrapolated region, an assumption will be made that the owner sells the loaders at an average cumulative hours of use of 10,000. This equates to a 6.7-year calendar life at 1500 hours per year. The factors that change for the Nichols' method are "total hours of use" which drops to 1.6 and "years of useful life" which drops to 1.07. The average repair cost based on these numbers is $17.12 per hour. Using the equation for CCI for this fleet, the CCI at 10,000 hours of use is 1.342, which yields an average cost of $13.42 per hour. This drops to $3.42 once the ownership costs are factored out. These numbers are more comparable, but the Nichols method still delivers forecast average repair costs that are too high. Perhaps the reason for this is the age of the Nichols text. The most current edition was published in 1976. The first edition was published in 1955. There have been numerous breakthroughs in equipment quality and reliability since the 1950's. It is felt that the Nichols' method could still provide reasonable figures for repair costs if the factors were updated.

8.4.2   Nunnally

Nunnally's method as presented in Chapter 2 attempts to estimate repair costs as a percentage of purchase price in a manner similar to the way that depreciation is figured (1993). The same fleet of wheel loaders described in the previous section will be used for this comparison.

The first costs that will be compared are the average lifetime costs of the repairs. Using the Nunnally method, the average lifetime repair costs are found by multiplying the purchase price by a repair cost factor. This number is then divided by the number of hours of operation. For wheel loaders, the factor is 0.60. Multiplying this by $100,000 and dividing by 20,000 hours provides a forecast average repair cost of $3.00 per hour. Although a little on the low side, this figure is much closer to the $5.92 per hour figure derived using the cumulative cost equations.

It may be more realistic to compare the repair costs near the point at which the repair costs reach 0.60 times the purchase price, or at CCI =1.6. This will allow the comparison to made at the

point that Nunnally's method is geared towards. Solving for x in the CCI equation, a CCI of 1.6 occurs at approximately 13,750 hours. Using Nunnally's method, the average repair costs are $4.36 per hour. Using the CCI equation, the average repair costs are also $4.36. Since the average repair costs are equal at this point, it provides a good basis for a comparison of how the two methods arrive at this point.

To do this, Nunnally's equation was adapted to provide a CCI instead of hourly repair costs. This equation is as follows:

$$CCI = \left\{ \frac{\text{Year Digit}}{\sum \text{Year's Digits}} \times \text{Lifetime Repair Cost Multiplier} \right\} + \text{Previous Year's CCI} \qquad \textbf{Equation 8-9}$$

The results were calculated at 1500 hour intervals (approximately one year's operation). They are shown in Table 8-8. Nunnally's method provides CCI forecasts that are incredibly close to the values obtained using the cumulative cost equation. The values are so close that one can barely differentiate between the two data streams when they are plotted. When a regression line was fit to Nunnally's points, the line lied nearly on top of the cumulative cost curve.

The difference between the cumulative cost curve and Nunnally's curve is that the cumulative cost curve is based on actual data. Nunnally's curve was made to fit the cumulative cost curve by providing it with a point that was common to both equations. Nunnally provides no methodology to come up with this point. There is no description of how to find the optimum values for either life or cost. The Nunnally equation *does* provide a very accurate facsimile of the cumulative cost curve if it is given information related to the optimum cumulative hours of use.

8.4.3 Straight-line

A number of straight-line methods were described in Chapter 2. For comparison purposes, only one of these will be looked at—percentage of straight line depreciation. Peurifoy et. al. (1996) recommend using an annual repair cost that is based on a percentage of straight-line depreciation that is determined from historical records. The same wheel loader fleet will be used for this comparison.

Table 8-8: CCI Values For Performance Comparison

Hours/1000	CC Curve	Nunnally	Straight-Line
0	1	1	1
1.5	1.018937	1.013333333	1.087598
3	1.049317	1.04	1.175196
4.5	1.091141	1.08	1.262793
6	1.144408	1.133333333	1.350391
7.5	1.209119	1.2	1.437989
9	1.285273	1.28	1.525587
10.5	1.372871	1.373333333	1.613184
12	1.471912	1.48	1.700782
13.5	1.582397	1.6	1.78838
15	1.704325	1.733333333	1.875978
16.5	1.837697	1.88	1.963575
18	1.982512	2.04	2.051173
19.5	2.138771	2.213333333	2.138771
21	2.306473	2.4	2.238771

The L* value and the CCI value that corresponds with this fleet will be used to calculate the straight line depreciation percentage. Assuming that the loaders have no residual value when they are disposed of, the baseline depreciable value is $100,000. This makes the annual depreciation $5128. The depreciation in terms of CCI is given by the equation:

$$CCI = 1 + \frac{\text{year's digit} \times (CCI @ L* - 1)}{\text{total number of years}}$$
Equation 8-10

This can be converted to a percentage of depreciation by the following equation:

$$Percentage = (CCI @ L* - 1) \times 100$$
Equation 8-11

In this case, repair costs are approximately 115% of annual depreciation. The CCI line for staight-line depreciation is shown is shown in Figure 8-12. Although the two lines end up at the same point, they arrive at that point in fairly different fashions. The straight-line method overestimates the CCI until the lines intersect. This overestimation of the CCI is due to an

overestimation of repair costs early in the lives of the machines. After approximately 9,500 hours (where the two lines have equal slopes), the straight line method *underestimates* repair costs. These variances from the actual way that the repair expenditures occur could have an impact on the cashflow planning for the company concerned
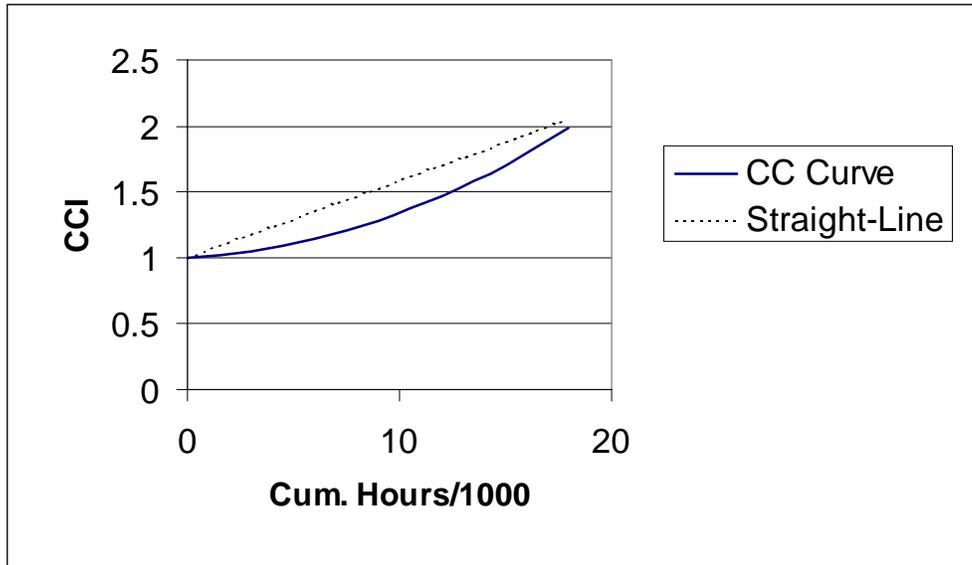
Figure 8-12: Cumulative Cost Curve vs. Straight Line

8.5   SUMMARY

In this chapter, the results of this study were discussed in detail. In the first section, the numbers obtained were evaluated as to how realistically they portray reality. In most cases, the equations did provide models that made sense. Equations for the derivation of L* and T* were presented. It was suggested that it may be possible to come up with an average CCI at which most machines have reached their optimum average cost. A very strong relationship between L* and T* was demonstrated. It was theorized that collateral costs may not have that great an impact on fleets with low L* values—but collateral costs may play an important role in the determination of L* for those fleets with higher L* values. These fleets are usually heavier, production-oriented machines.

Sensitivity analyses were performed to discern the effect that changes in β values have on L* and T*. The regression equations of the different fleets were compared to each other on the basis of

company, equipment type, and equipment size. Although the results obtained were not conclusive, it seems that the company that owns the fleet has an impact on the equations. It was also shown that it may be possible to derive industry-standard equations for some types of equipment with a more comprehensive study. It also was suggested that heavier, larger machines of the same type have longer L* values and smaller T* values than their smaller counterparts.

Finally, the results obtained were compared to results that would have been obtained through the application of three other methods in literature. Although the Nichols method proved to have possibly outdated factors, the Nunnally method provided excellent replication of the cumulative cost curve once it was provided with a target CCI. The straight-line method was shown to overestimate the repair costs for newer machines and underestimate the repair costs for older machines.

With the results fully discussed, Part III of this dissertation, The Work, comes to a close. This part covered the methodologies for preparing the data, the analysis of the data, and the results obtain from that analysis. Part IV of the dissertation looks to the future with suggestions for the use and implementation of the models derived.

CHAPTER 9: INTEGRATION

The first three parts of this dissertation presented the problem, defined the work, and described the analysis and results that followed. This part focuses on the future. Methods for using and defining the cumulative cost model will be described for those who will use it. The dissertation will be summarized and areas for future research will be proposed. This chapter concentrates on some items that will, hopefully, bring the CCM and the cumulative repair cost equations into mainstream usage.

This chapter will flow as the dissertation did. The first two sections will describe bringing the theoretical cumulative cost model into usable spreadsheets. The second two sections will focus more on the details of properly defining the β terms. Specifically, the topics will include:

- A spreadsheet solution to the rebuild decision

- A preliminary analysis of the NEL

- A usable methodology whereby companies can develop their own equations

- A proposed framework for the development of industry benchmarks

9.1   AN EXAMPLE: THE REBUILD DECISION

Chapter 3 of this dissertation provided rough explanations of how to use the CCM as an aid in making economic decisions concerning the buying, operation, and selling of construction equipment. This section will focus on one of those decisions that can be supported with the equations developed in this research.

Equipment management decisions described in Chapter 3 that do not relate to the buying and selling of equipment can be organized in a continuum to better understand their nature. These decisions have certain attributes that distinguish them from each other. The decision—attribute continuum is depicted in Table 9-1.

Table 9-1: Decision Continuum

Attribute	Decision											
	Maintain			**Repair**			**Major Repair**			**Rebuild**		
Regular	Y	Y	Y	N	N	N	N	N	N	N	N	N
Frequent	Y	Y	Y	y	y	y	N	N	N	N	N	N
Costly	N	N	N	n	n	n	y	Y	Y	Y	Y	Y
Failure	N	N	N	Y	Y	Y	Y	Y	Y	N	N	N

The four major decision types depicted are: maintain, repair, major repair, and rebuild. Major repair was not defined as a decision type in Chapter 3, but there are subtle differences between repairs and major repairs that will be discussed here. The four attributes depicted are regularity, frequency, cost, and failure requirement. The letter "Y" signifies that the decision type posseses the attribute in question. The letter "N" denotes the opposite. Lowercase "Y"s and "N"s signify the degree to which the decision type possesses (or does not possess) that attribute is less than other decision types.

Maintain decisions occur on a regular basis, normally scheduled at certain intervals of cumulative hours. They occur frequently and are relatively inexpensive. They do not occur as a result of failure of a machine—in fact they are undertaken to *prevent* equipment failures. *Repair* decisions occur frequently, but not regularly—repairs are generally unscheduled because they are a reaction to some type of failure on the machine. Repairs can be more expensive than routine maintenance, but are still relatively inexpensive compared to the remaining two decision types.

Major Repairs occur infrequently. They are costly repairs that take place due to a major failure of some component of the piece of equipment. Although some major repairs are very costly, some are not quite as expensive, thus the lowercase "y" for part of the cost continuum. *Rebuilds*

also occur infrequently. They are generally expensive. Rebuilds do not occur as a result of failure. If failure has occurred, the "rebuild" is in actuality a major repair.

The rebuild decision is interesting and pertinent. It is a bit more involved than the initial purchase decision which can be made simply by comparing the T* values of the various alternatives. There are many factors that must be considered.

The rebuild decision as presented in Chapter 3 was based on comparisons of the NEL of the machine being evaluated and NEL of its rebuild. The NEL was not defined in this research. For the purposes of a rebuild, evaluation of the GEL should yield results similar to those that would be obtained by evaluating the NELs. Both of the GELs will be displaced vertically a similar distance from their respective NELs. There will be some error in the solution due to the fact that the solution is based on angles, not vertical displacement. The magnitude of this error should be small if the machine is not sold prematurely because as machines age their GELs approach their NELs—which means that the angular difference between a T* to the GEL and a T* to the NEL will be small.

There are three important questions that relate to the rebuild decision:

- When?

- How much?

- What is gained?

It is postulated that a rebuilt machine possesses the same cumulative cost curve (GEL) as it did before the rebuild. The curve is simply shifted vertically and horizontally on the cumulative hours/CCI plane. This is illustrated in Figure 9-1. The "When?" is the machine age at which the rebuild is evaluated. In the figure, this age occurs at cumulative hours/1000 of 8. This determines the horizontal reference point for the shifted GEL. The "How much" is the percentage of the purchase price that the rebuild will cost. In the figure, that percentage is illustrated as a vertical difference between the two GELs at the evaluation age.
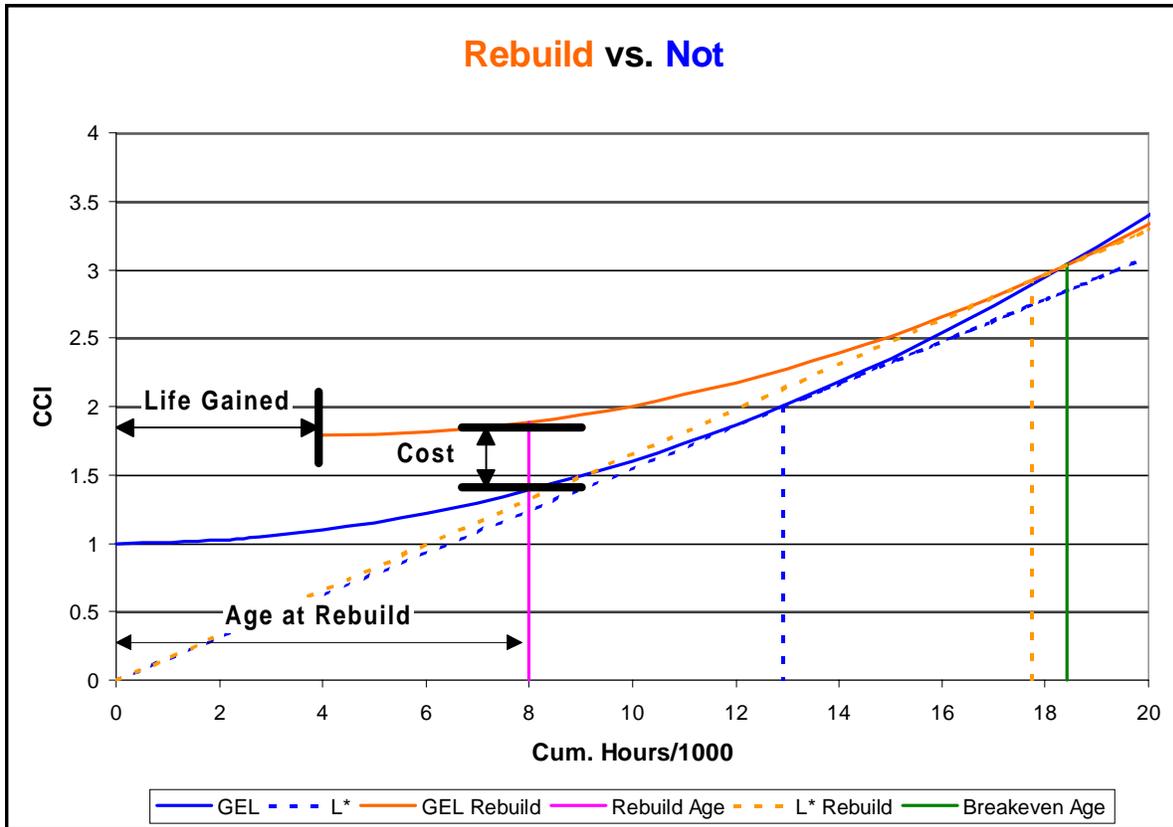
Figure 9-1: Three Aspects of Rebuild

The "What is gained?" is the amount of life that is purchased when the rebuild is accomplished. At the end of a rebuild, the machine should behave like a younger machine. The person performing the rebuild should be able to give an estimate of the form "This 8000 hour machine will perform like a 4000 hour machine after the rebuild is accomplished." In this case, 8000 minus 4000 equals a gain in life of 4000 hours. This shifts the starting point of the GEL for the rebuilt machine 4000 hours to the right.

The GEL for the machine before the rebuild is curve that starts at age = 0 and CCI = 1. The dotted lines signify the T* and L* for this machine. The GEL for the machine after the rebuild starts at a point above and to the right of the non-rebuilt GEL. Once again, the T* and L* are shown by dotted lines. The single non-dashed vertical line indicates the point at which the GEL for the rebuilt machine intersects the GEL for the non-rebuilt machine. The average cost per period for the rebuilt machine is cheaper than that of the non-rebuilt machine after this point. This does not, however, mean that the rebuild is the best option to take. This is determined by

comparing the T* of the rebuilt machine to the T* of the machine before the rebuild. If $T^*_{rebuild} <$ T*, then the rebuild *may* be the best option. $T^*_{rebuild}$ must also be less than $T^*_{challenger}$ (if purchasing a new machine is feasible).

A software tool was developed to compare $T^*_{rebuild}$ with T*. The tool was developed in Microsoft® Excel®. The file can be accessed by clicking on the button below if this dissertation is being read electronically from Virginia Polytechnic Institute and State University (Virginia Tech). Alternatively, the file can be obtained by contacting the Virginia Tech University Libraries.

LINK TO REBUILD.XLS

A screen view of the spreadsheet is depicted in Figure 9-2. The user inputs are:

- x coefficient (β_1)
- x^2 coefficient (β_2)
- Age at rebuild
- Cost of rebuild
- "Age" after rebuild

The coefficients are input exactly as they are calculated. The age items are input as hours/1000. The cost is input as a fraction of purchase price. A \$40,000 rebuild on a \$100,000 machine would be input as "0.4". As the user inputs these figures, the following tasks are automatically accomplished by the spreadsheet:

- Both original and rebuild GELs are computed and plotted
- Both original and rebuild L* and T* are computed, plotted, and displayed
- The "breakeven" line is computed and plotted
- T* values are compared and user is informed of outcome

The calculations for $L^*_{rebuild}$ and $T^*_{rebuild}$ are a little less straight-forward than the those of the machine before the rebuild. The formulas for these calculations are provided in Appendix F. The "breakeven" line is a vertical line drawn from the x-axis to the intersection of the two GELs.. The

user is given a recommendation in the form of a "Rebuild" or a "Don't Rebuild" as the first entry in the results box.
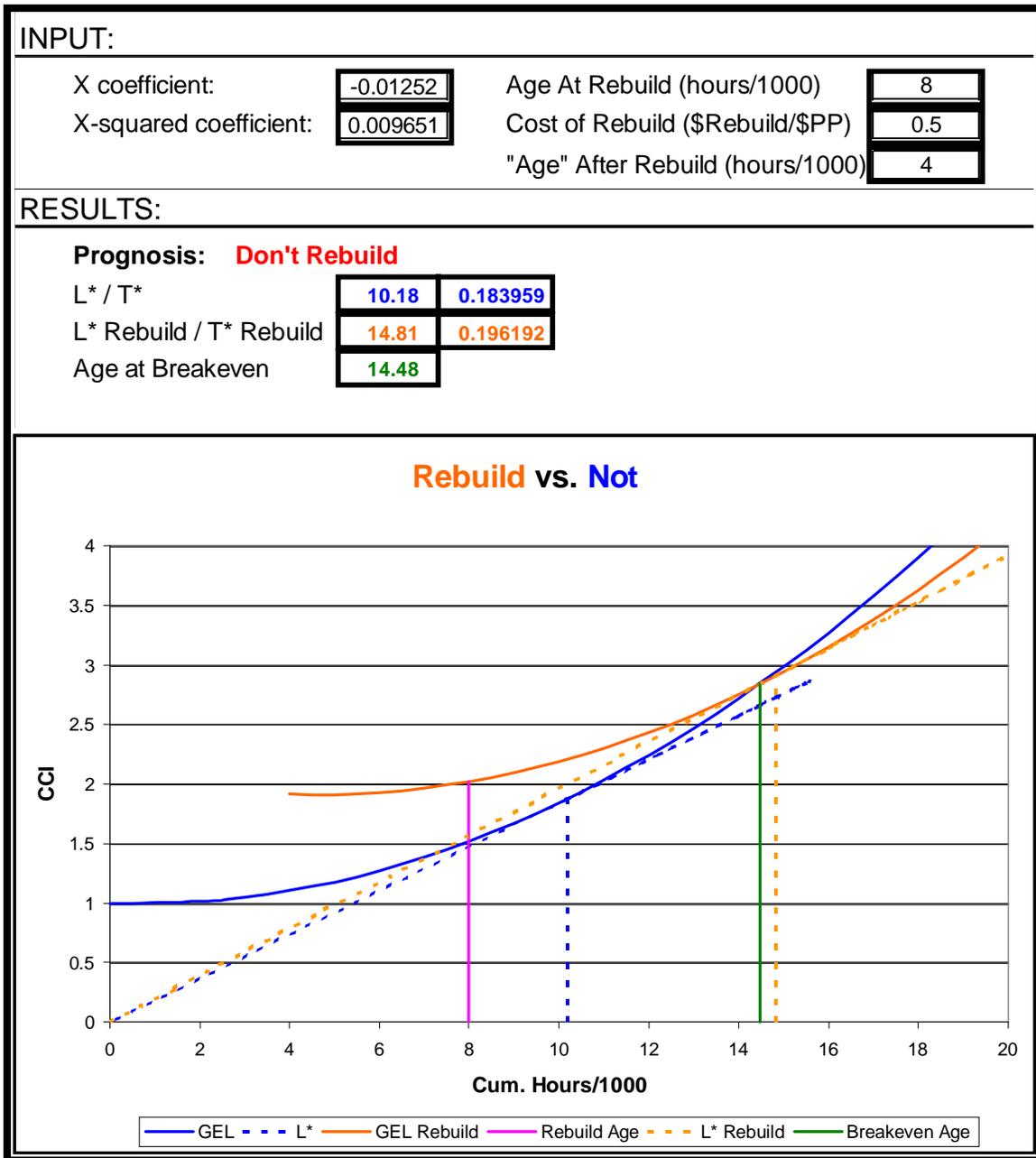
Figure 9-2: Rebuild Spreadsheet

An example will help with the understanding of this spreadsheet. For the example, the fleet of mid-size dozers with the smallest L* value will be used. The coefficients are entered into the spread sheet first. Then, the user inputs the other three variables. These inputs should be based

on a valid rebuild estimate or based on the equipment manager's experience with similar rebuilds.For the first case, the equipment manager assumes a rebuild age of 8,000 hours with a cost of 50% of the purchase price. The machine will seem like a 4,000-hour machine after the rebuild.
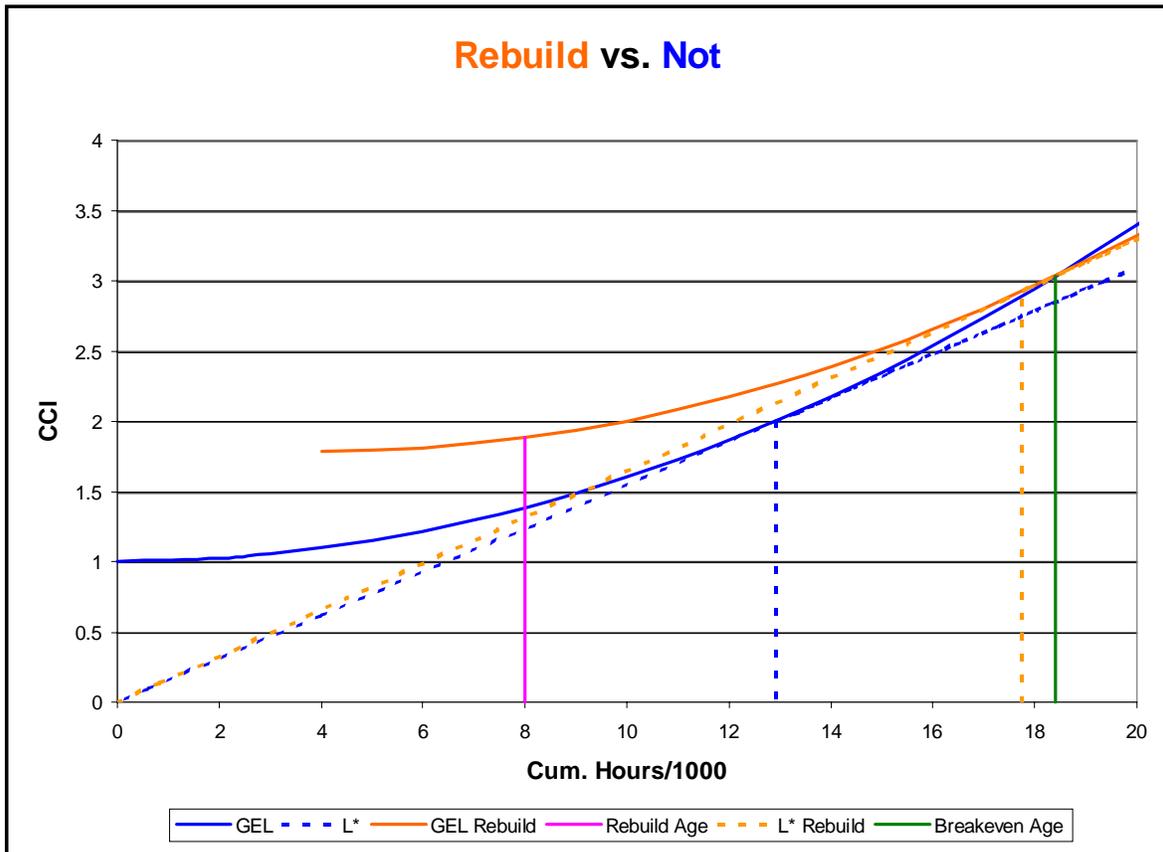
Figure 9-3: Case #1 Rebuild

As can be seen in Figure 9-3, the T* of the rebuilt machine is greater than that of machine before the rebuild, so it is probably best not to accomplish the rebuild with the given parameters. In order for the rebuild to be chosen, the cost must decrease or the age after the rebuild must decrease. The equipment manager could also increase the age at rebuild. In any case, the GEL for the rebuilt machine must be shifted either down or to the right in order to flatten the angle on the T* line.

Assume that by performing some of the rebuild in-house, the cost of the rebuild can be brought down to 35% of the initial purchase price. The new cost is input, the curve is shifted down, but

the prognosis is still "**Don't Rebuild**" because although the T* values are close, the machine before the rebuild still has a slight edge.

If the rebuild can be delayed until 9000 hours, the T* of the rebuilt machine is slightly less than that of the machine before the rebuild. The prognosis is "Rebuild". This is shown in Figure 9-4.
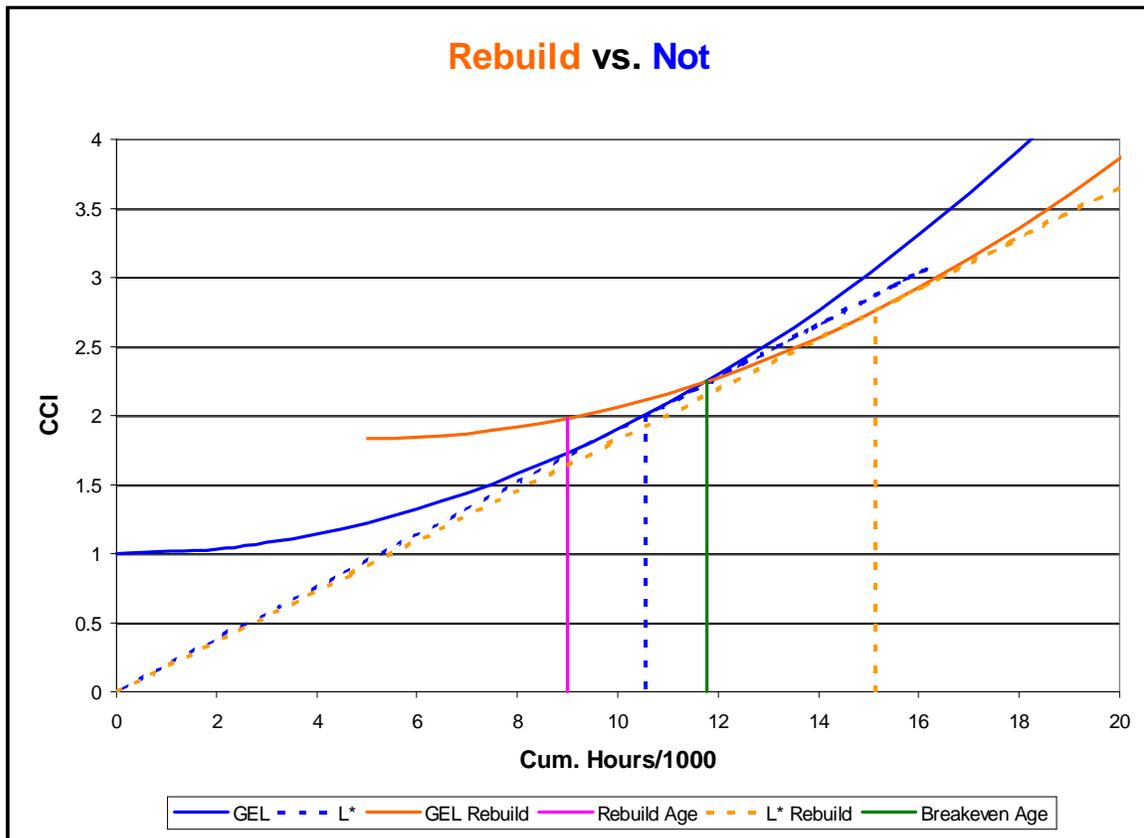
Figure 9-4: Case #2 Rebuild

The fact that $T^*_{rebuild}$ is less than T* is graphically evident in Figure 9-4. The user should also rely on the numerical values for T* that are calculated and check the prognosis reading. The value of this spreadsheet tool is that the equipment manager can attempt any number of combinations of the three parameters and see the results both graphically and numerically. Before making a final decision on the rebuild, the equipment manager must remember to compare $T^*_{rebuild}$ to T* of any challengers that may provide better economy than the original machine.

9.2   PRELIMINARY STUDY OF THE NEL

Up to this point, this dissertation has focused almost exclusively on the development and interpretation of the GEL. Although the GEL provides an approximation of the NEL as machines age, as pointed out in Chapter 3 the NEL should be the true basis for economic decisions when possible.

To get an idea of the differences between the NEL and GEL, a cursory study of residual values was accomplished on articulated trucks. Actual selling prices and trade-in values were obtained for 14 articulated trucks in two companies. These prices/values were compared to the purchase prices of the machines to obtain an expression for the residual value in terms of cumulative hours of use. This was accomplished using regression analysis.

A starting point for the analysis was obtained during a conversation with the academic coordinator for the Association of Construction Equipment Managers (Vorster, 1998). A rule of thumb that has been used by equipment managers for obtaining residual values is given by:

$$\text{Residual Value} = \frac{1}{\sqrt{\dfrac{\sum hours}{1000}}} \times \text{Purchase Price} \qquad\qquad \text{Equation 9-1}$$

The residual value is equal to the reciprocal of the square root of cumulative hours/1000. For a 4000 hour machine, the residual value is 0.5 times the purchase price. The data were fit to the model:

$$\text{Residual Value} = \beta_1 \frac{1}{\sqrt{x}} + \varepsilon \qquad\qquad\qquad\qquad \textbf{Equation 9-2}$$

Where:

β_1 = coefficient

x = cumulative hours of use

ε = error term

The data fit this model with an adjusted R^2 value of over 0.99. The coefficient value was 1.03 which had a p-value of less than 0.0001. This particular data fit the residual value rule-of-thumb quite nicely. Based on this analysis, equation 9-1 was used to compute the residual values for this exercise. The equation used to generate the NEL is as follows:

$$CCI = 1 + \beta_1 x + \beta_2 x^2 - \frac{1}{\sqrt{x}}$$ Equation 9-3

An Excel® spreadsheet was developed to plot the GEL, NEL, their tangents, and their optimum lives. This spreadsheet can be accessed by clicking on the button below or by contacting the Virginia Tech libraries.

<div style="border:2px solid blue; text-align:center;">LINK TO NEL.XLS</div>

This spreadsheet (Figure 9-5), like the rebuild spreadsheet, requires the user to input the coefficients for x and x^2. As these coefficients are input, the spreadsheet calculates the GEL and NEL lines. Tangents to these lines are drawn and vertical lines from the tangent points are drawn to delineate the points at which L* is reached.

The tangent to the GEL was found as described in Chapter 8. The tangent to the NEL was found through an iterative process. The slope of the tangent line was defined in terms of equation 9-3. The first derivative of the resulting equation was taken to define the point of minimum slope by the following equation:

$$\beta_2 x^2 - 1.5x^{-0.5} - 1 = 0$$ **Equation 9-4**

This equation was solved iteratively for x to yield L* for the NEL. A series of iterative solutions were performed for varying values of β_2 to formulate a regression equation for the solution of L* for the NEL. This equation is:

$$L^*_{nel} = 0.3548 \, \beta_2^{-0.6209}$$ **Equation 9-5**

INPUT:

X coefficient:	0.008
X-squared coefficient:	0.0045

RESULTS:

L*/ T* Based on GEL:	14.9137	0.1421
L*/ T* Based on NEL:	10.1705	0.1212
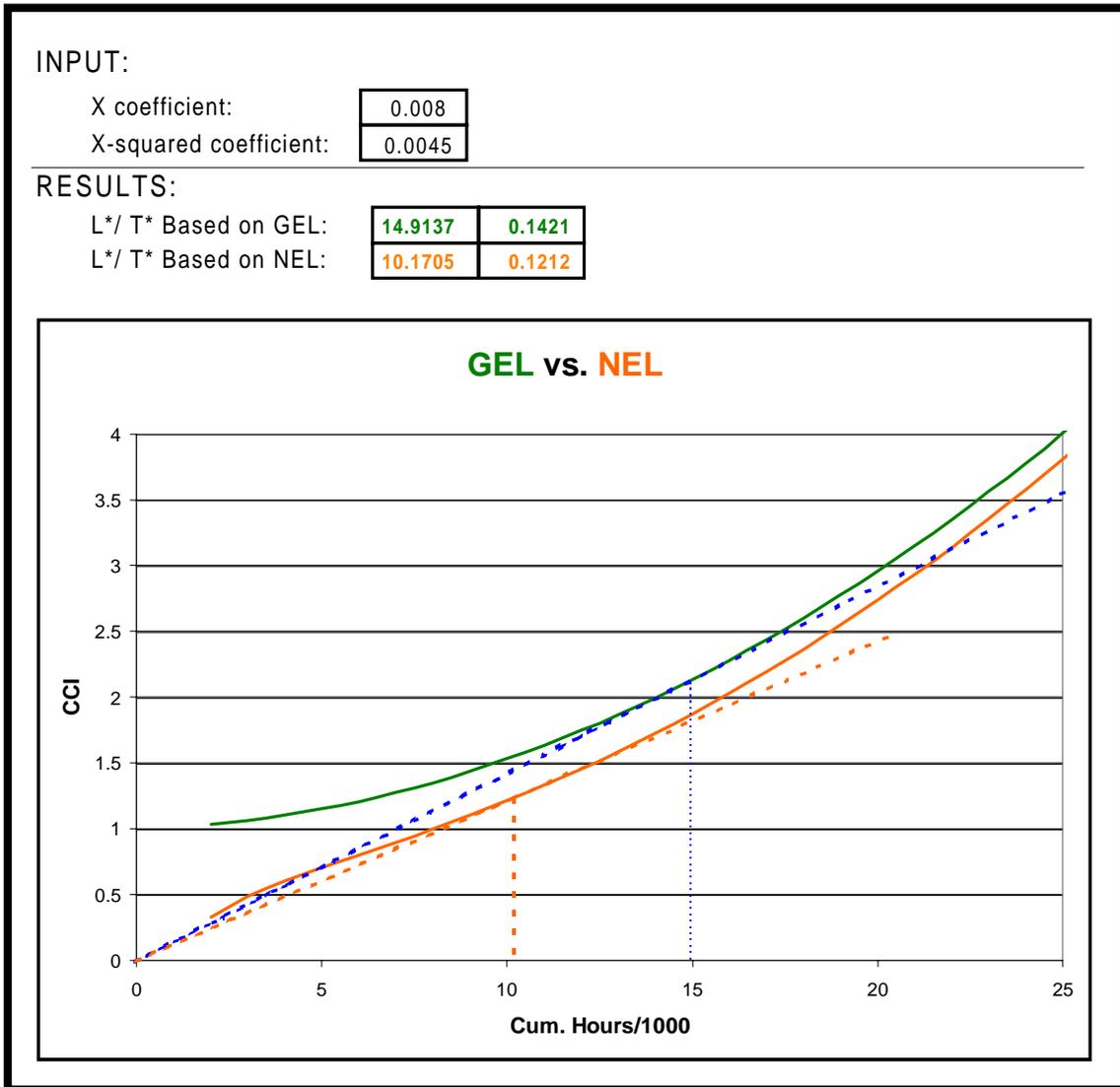
Figure 9-5: NEL Grapher

This equation is only valid if equation 9-2 is valid. As an example, one of articulated truck fleets is presented in Figure 9-5. The coefficient values are input by the user. The spreadsheet automatically produces the graph and the results for L* and T*. Both L* and T* are lower for the NEL. The L* is lower because the residual value grows smaller with accumulated hours which forces the NEL to gradually converge with the GEL. The T* is lower because the average cost per period is reduced when the loss in residual value is spread out over the life of the machine. The differences between the values were significant. L^*_{gel} was nearly 50% larger than

L^*_{nel}. For this particular fleet, that equates to around two calendar years of operation—a figure that cannot be ignored.

Using the GEL instead of the NEL will have differing impacts depending upon the decision being made or the information being retrieved. To forecast repair costs, the GEL as calculated in this dissertation is the best line to use. The average repair costs per hour at a specific cumulative hours of use can be found by taking the first derivative of the cumulative repair cost equation at the point in question. Depending upon which type of decision is being made, using the GEL may or may not be a good decision. The pros and cons relating to the different decisions are described below:

- *Purchase:* NEL should be used. A machine that holds its residual value well but costs more up front might lose out to a cheaper machine that loses its value quickly if the GELs are the basis for the decision.

- *Maintain:* Use of the GEL should provide the proper decision since alternative maintenance strategies do not relate directly to residual values. The L^* and T^* values computed may be slightly higher than what will actually be experienced, but they will be higher by the same relative amounts for each strategy analyzed.

- *Repair:* Must have *both* the GEL and the NEL to evaluate repair limits. The NEL may be obtained through historic data and the current residual value of the machine under evaluation. For forecasting repair costs, the GEL should be used.

- *Rebuild:* The GEL should provide the proper decision for reasons discussed in section 9.1. Once again, actual L^* and T^* values may be lower than those which are forecast. The errors become compounded if there is a suitable challenger involved.

- *Replacement:* If the machines/production teams lose their residual value at similar rates, the GEL may provide the proper decision but high L^* and T^* values. If the loss of value rates differ, the NEL should be used.

- *Retire:* The NEL should be used. If the GEL is used, it may direct the user to keep the machine longer than the optimum economic life. This was demonstrated in this section.

9.3   FIELD IMPLEMENTATION

Implementation of the ideas presented in this dissertation by construction equipment companies is a goal that this research team hopes to achieve. The cumulative repair cost equations are not exceedingly difficult to derive using available data already at the equipment manager's disposal. The equations can give the equipment manager a better forecasting tool for repair costs throughout the life of the fleet and at individual points of interest. It was already demonstrated that the use of the repair cost equations can help in making rebuild decisions. By using a rule-of-thumb or historical data, the NEL can be approximated and other economic decisions can be made. Also, through implementation the model's strengths and weaknesses can be evaluated.

This section will discuss the following:

- Data collection

- Data analysis

- Use of the equations

9.3.1   Data Collection

As was demonstrated in this dissertation, many companies already have at their disposal all that they need to derive cumulative repair cost equations. In some cases, the data are not that easy to come by or manipulate—but they are there just the same. Many techniques for handling difficult situations relating to the data were discussed in Chapter 4. This section will not serve as a review of those techniques. They are available to the user if they are needed. What this section will do is describe a data collection methodology that can work to support the derivation of cumulative repair cost equations. The resulting database will provide a good way that a company can utilize existing data to come up with the equations.

The data structures described are not intended to serve as a substitute for a complete equipment management database. The databases are only described insofar as they support the derivation of cumulative repair cost equations and the implementation of the CCM. An excellent description of a complete equipment management database design can be found in Chapter 10 of *Computer Applications in Construction* (Paulson, 1995).

There are a certain amount of static, non-changing data associated with each machine. These data are given in Table 9-2.

Table 9-2: Static Data

Machine #	Type	Size	Purchase Date	Purchase Price

The formatting of the machine number is at the discretion of the company—it serves simply to identify the machine as unique. The type of machine shows the general classification (dozer, articulated truck, etc.). The size indicates the machine's size category within its type. This can be done by bucket size, horsepower, weight, etc. The purchase date and purchase price are necessary for the formulation of the CCI. The data required for this table should be relatively easy to acquire if they are not already in the company's accounting database.

The second table required is that of maintenance and repair data. The table is given in Table 9-3. The machine number is as described above—it provides the linkage between the two tables. The account relates to the type of repair. Only four accounts are needed—but more accounts could help with other aspects of equipment management. The necessary accounts are Tires & Tracks, Ground Engaging Implements, All other Maintenance and Repair, and Abuse. Costs can be broken down into more than two categories, or they can be combined into one category. It is usually good practice to track parts and labor separately. The date is simply the date upon which the repair took place. The meter hours are the cumulative meter hours as read by the mechanic when the repair or maintenance action was performed (if the company tracks this information).

Table 9-3: Repair Table

Machine #	Account	Parts Cost	Labor Cost	Date	Meter Hours

The meter hours field is where this table may differ slightly between companies. Most companies already have databases that provide everything except for meter hours. The solution that would work the best for the cumulative repair cost equations is to have the mechanics record the meter hours at the completion of every work order. The meter hours will then be input into the computer by the same person that keys in the work orders. The calendar date/cumulative hours marriage will be complete and a data table and some manipulation will be eliminated. If the company cannot implement a process whereby the meter hours are recorded with maintenance and repair actions, another table is needed. This table is depicted in Table 9-4. This table can be supplied with data from a variety of different sources. The oil sampling database (if used by the company) can provide a quick way to get this data, but is not the ideal situation. Sometimes oil samples are not recorded or oil changes are accomplished late. In these cases, some of the 500-hour interval data pairs can be lost. A better way to do this (which has been implemented with some success in the field) is through direct recording of the cumulative meter hours for each machine on a periodic basis. Some companies do this every time that the equipment is refueled. Others require either the mechanics or the job superintendents to provide the meter hours on all machines in their charges on a regular basis (weekly readings work well.) The cumulative meter hours should be recorded on at least a monthly basis.

Table 9-4: Date/Hours Table

Machine #	Month	Hours

The final data table required to develop equations is the inflation table (Table 9-5). It is felt that the Consumer's Price Index (CPI) provides an adequate measure of inflation as it affects all sectors of our society—it is also very easy to obtain. The CPI will be used to adjust the costs

incurred to *current* dollars. If the user so chooses, the inflation table can be ignored. The user does this with the knowledge that the L* and T* values obtained using non-adjusted data will differ from the actual costs incurred. It was shown in Appendix A that the effects of inflation are not negligible. The effects are especially apparent if comparing old machines to newer machines. Unadjusted data will have a cumulative repair cost line that is above that of the adjusted data. This could make it look like older machines have a higher T* than they actually do. Data for the CPI can be obtained over the internet from the Bureau of Labor and Standards website using the instructions provided in Appendix A.

Table 9-5: Inflation Index Table

Date	CPI

The inflation indices are the last of the data necessary to construct cumulative repair cost equations. The next section, data analysis, will explain how to use these database tables to derive the cumulative repair cost equations.

9.3.2   Data Analysis

The usable methodology for forming cumulative repair cost equations differs in many ways from the experimental methodology used in this dissertation. The usable methodology is much simpler. It is designed for implementation using only two PC-based software programs—a spreadsheet and a relational database. Microsoft® Excel® and Access® were the programs for which this methodology was tailored. Other competitive packages should be able to provide similar results. The general steps for accomplishing this analysis are flowcharted in Figure 9-6.

It is important to note that skilled programmers could combine some or all of these steps into one operation. The purpose of breaking the analysis into five steps was for ease of understanding. The user should feel free to streamline the process when they are capable.

The first step in the analysis is to generate a summary report of monthly repair expenditures for each machine in the fleet to be analyzed. The fields in this summary report are shown in This is done in a database program. The report is filtered so the summaries are generated only on the fleet of interest. The repair table must be linked with the static data table to perform this filtering. The monthly expenditures should include all maintenance and repair cost accounts with the exception of Tires and Tracks, Ground-Engaging Tools, and Abuse. The monthly expenditures should be in their *incremental* form. This is necessary for the application of inflation indices. If the expenditures are generated in their *cumulative* form, additional manipulations must be accomplished to get them to the incremental form. Purchase dates and purchase prices are included with the report since the static data table was linked for the filtering.
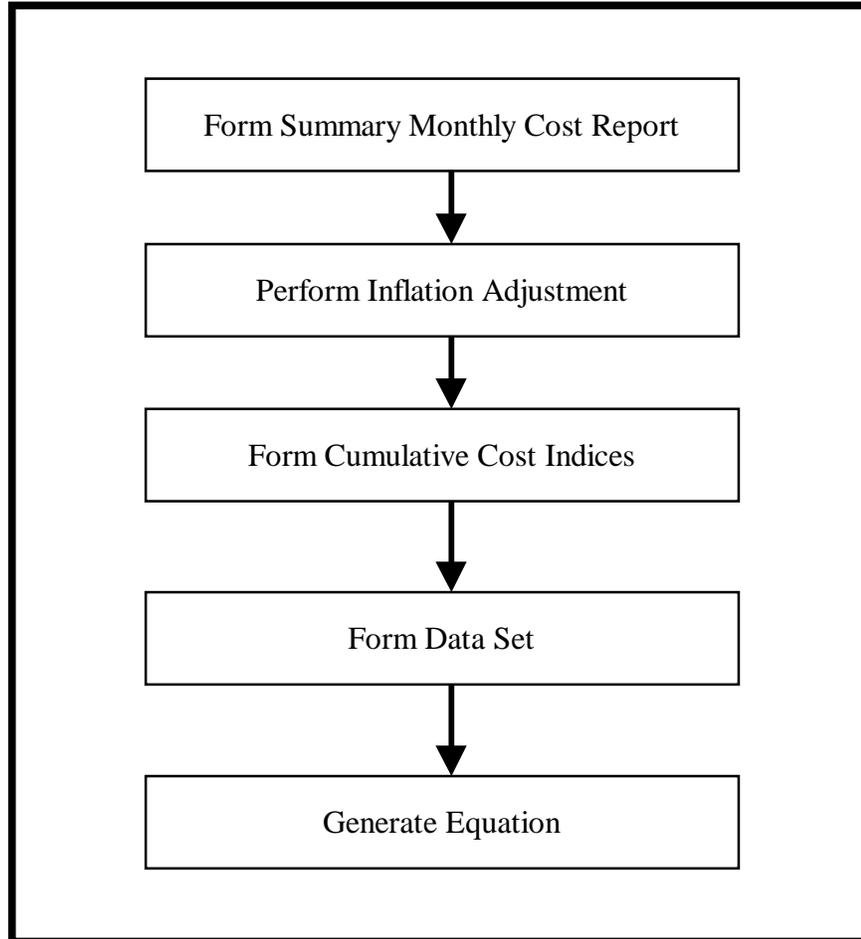
Figure 9-6: Analysis Flowchart

The date/hours table can also be linked into this report. When the hours are linked, they should be divided by 1000 to get them ready for the analysis.

Table 9-6: Summary Report

Machine #	Purchase Date	Purchase Price	Cumulative Hours/1000	Repair Month	Incremental Repair Costs

The second step of the process is to perform the inflation adjustment. In the experimental methodology, all costs were indexed to 1987. The reason for this was that no machines were older than 1987 models and the data were received at differing time—it was more efficient to use an index month in the past than one in the present. For field users, however, the results will be more useful if they are expressed in terms of current monetary units rather than units of some time in the past. The equation that should be used to index the costs for inflation is as follows:

Indexed cost = (cost * index of current month)/ (index of month incurred)   **Equation 9-6**

This equation will increase the values of previous expenditures in order to express them in current day dollars. Apply this equation to all monthly incremental costs and to the purchase prices. The output table should be nearly identical to Table 9-6 with the exception that the purchase price and incremental repair costs will be indexed for inflation.

The third step in the process is to form the cumulative cost index for each monthly entry. This can be done in either the database or the spreadsheet program. This is, however, a good point to transition to the spreadsheet program. To form the cumulative cost index, the indexed monthly repair costs must be converted from their incremental to their cumulative form. This can easily be accomplished in the spreadsheet. The index is then calculated for each month using equation 4-1.

The fourth step is the final formation of the analysis data set. The user should now have a list of cumulative hours and CCIs for each machine on a monthly basis. Each machine should follow the next with no spaces between machines. The monthly CCI basis must be converted to one data pair for every 500 cumulative hours. This is done through the process of interpolation. An

automated method of doing this is depicted in Table 9-7. The code in column D creates a number called "Floor" which is the cumulative hours/1000 rounded down to the nearest 0.5. Column E, "Interval", eliminates repeated values of column D and replaces them with blanks. Column F, "CCI", interpolates between the CCI values of the current line and the previous line.

Table 9-7: Excel® Codes for Interpolation

	A	B	C	D	E	F
1	Mach. #	Hours /1000	CCI	Floor	Interval	CCI
2	mach.#	hours/ 1000	CCI	=IF(A2=A1,FLOOR(B2, 0.5),(0-FLOOR (B2,0.5)))	=IF(D2=0,"",IF(D2< 0,"",IF((D2+D1)=0," ", IF(D2=D1,"",D2))))	=IF(E2="","",C2- (B2-E2)*(C2- C1)/(B2-B1))

The final step in the analysis, the actual formation of the equations is fairly straight-forward.

1. Select columns E and F from the spreadsheet depicted in Table 9-7.

2. Copy the columns and paste them in columns G and H using the EDIT—PASTE SPECIAL—VALUES. This command pastes the values (not the formulas) of the cells in columns E and F into columns G and H.

3. Next, select columns G and H and use the command DATA—SORT—BY(Column G)—ASCENDING. This will eliminate the empty cells and provide a neat list for graphing.

4. Select only the cells that contain data pairs in columns G and H this time. Now, use INSERT—CHART—XY(Scatter)—SUBTYPE (points only). Click the NEXT button three times to scroll through various screens, then click FINISH. A scatterplot graph of the cumulative hours vs. CCI should be on the screen.

5. Select the data series (the points) on the chart. When the series is highlighted, press the right mouse button. Select ADD TRENDLINE. A pop-up box will appear with various trendline types. Select POLYNOMIAL, 2nd ORDER. Click the OPTIONS tab within the ADD TRENDLINE dialog box. Select SET INTERCEPT = 1 (the default value is zero). Select DISPLAY EQUATION ON CHART and DISPLAY R-SQUARED VALUE ON CHART. Click the OK button.

6. The regression line, equation, and R^2 values will now be displayed on the chart. Copy down the equation for future reference.

The analysis is now complete. The displayed R^2 value is not the same adjusted R^2 used during the experimental analysis. It will, however, provide some measure of the fit of the curve. The displayed R^2 value will be higher than the actual adjusted R^2.

9.3.3   Use of Equations

With the analysis complete, the user can apply the equations in two different ways. The equations can be used as part of the CCM or they can be used as forecasting tools on their own merits. The use of the equations within the CCM was discussed in sections 9.1 and 9.2. The user can also calculate L* and T* using the equations described in Chapter 8—this is also related to the CCM.

The use of the equations as forecasting tools has been alluded to, but not discussed in detail. Two applications of how to forecast average costs using the equations will be discussed. The first example is finding the average repair cost in dollars per hour for machines that are of a specific age within an analyzed fleet. This is done by evaluating the first derivative of the CCI equation at the point of interest. The first derivative is given by the equation:

CCI/1000 hours = $\beta_1 + 2\beta_2 x$ **Equation 9-7**

The β components are taken from the cumulative repair cost equation. The "x" value should be expressed in hours/1000. The resultant is a number with the units of $/$/1000 hours. To convert this to a repair cost per hour, multiply by 1000, then multiply by the purchase price. This cost can be used to adjust internal rental rates based on the average age of the fleet. Alternatively, it can

be used as a yardstick against which machines of similar age can be judged. If the machine has lower average repair costs, it is performing better than average. The converse is true if it has higher than average repair costs.

A second type of forecast that can be performed with the equations is a period forecast. If the equipment manager would like to get an idea of how much it will cost to operate an average machine of age "x" a number of hours equivalent to "z". To do this, evaluate the CCI equation for "x" and for "x + z". Subtract CCI_x from CCI_{x+z}. The difference is the average period cost in terms of CCI. To convert this to dollars use the following formula:

$$Cost_z = \frac{CCI_{x+z} - CCI_x}{1000z} \times \text{Purchase Price}$$

Equation 9-8

9.4   INDUSTRY BENCHMARKING

The first three sections of this chapter have dealt with ideas that the cumulative repair cost equations can be used for now. This final section looks towards the future. In Chapter 8, it was pointed out that it may be possible to develop equations that are representative of a general type and size grouping of equipment if sufficient data were available. This section provides a roadmap for obtaining, analyzing, and evaluating such data.

Although it would eventually be desirable to develop industry-wide benchmarks for every type and size of equipment, the concept must first be proven on a small scale—one general category and class of equipment. Due to the similarities between the fleets of articulated trucks evaluated during this dissertation, it is recommended that the proof on concept be focused on 25-ton articulated haul units. If the project proves successful, other categories and classes of equipment can be evaluated.

Even a small-scale project would require the backing of an organization that possesses greater resources than any single university. It is recommended that a non-academic champion be selected to help assure the project's success. An organization that has a wealth of equipment management experience and resources is the Association of Construction Equipment Managers (ACEM). The group counts among its members some of best equipment management specialists

in the industry. These equipment managers represent a wide cross-section of construction companies—from multi-national conglomerates to small, regional firms. These companies have tremendous data resources at their disposal.

The data should be collected over a number of years. There should be sufficient data to develop a curve that covers the full range of accumulated hours over which these machines would be expected to operate—ideally full coverage would be available from 0-20,000 hours. An even wider range of data would be better, but data up to 20,000 hours should provide a very good idea of how these machines accumulate repair costs. This may require tracking the sales of machines so that more than one company's data is involved. In addition to maintenance and repair costs, selling prices, hours, and condition of used equipment should also be collected for a parallel study of residual values. The data should be in a standardized format. Each company should collect and report the data in the same fashion, if at all possible. The drive for standardization would be made easier if all the participating companies could agree on a standard—the ACEM would provide a logical forum for this.

After the data are collected, they should be analyzed using the same model and data set type that were selected during this dissertation. However, weighted regression should be used to eliminate any influence of non-standard variance. This was not possible in the limited study conducted for this dissertation. One-half of the data should be set aside for model validation.

After the equation is developed, it should be given the widest dissemination possible to determine its validity (provided that the equation developed has suitable p-values and measures of performance). Feedback and validation should be actively sought from companies that own and operate this type of truck but were not a part of the study. This feedback and the equation's performance in the field should be evaluated.

If the feedback on the equation is positive and it provides satisfactory field performance, the study should be expanded to incorporate all categories and classes of equipment. This should be done gradually, if necessary. It will take years and could prove to be expensive. When the final industry-wide benchmarks for all equipment are published, the work will not be over. Equipment manufacturers are constantly improving the performance/reliability/economy of their products.

New products and improvements to old products should be evaluated to determine their impact on existing equations.

If the feedback or performance of the equation is poor, the study would not have been in vain. It would have the served the purpose of promoting company-specific equations for each fleet of equipment and will have provide an opportunity for equipment managers to discuss operations and compare ideas with other equipment managers.

9.5 SUMMARY

This chapter discussed a wide range of ideas concerning the use and furtherance of the cumulative repair cost equations presented in this dissertation. It provided some logical uses for the fruits of the labor contained herein.

Two aspects of incorporating the cumulative repair cost equations into the cumulative cost model were presented with accompanying software tools. The rebuild decision is an economic decision that can probably be made without the use of the NEL. The methodology for doing so was presented. A preliminary study of the NEL as it relates to the GEL was presented. Actual resale data were used to provide partial validation for a generic rule-of-thumb for the estimation of residual value.

The chapter then refocused to the cumulative repair cost equations themselves. A methodology whereby construction companies can develop and use their own, company-specific equations was presented. Finally, a proposed expanded study of one category and class of equipment was outlined and discussed. The study is based on the somewhat promising results obtained when comparing articulated trucks. The ultimate purpose of the study would be to provide industry wide benchmarks on all types and sizes of construction equipment.

This concludes this dissertation's contribution to the body of knowledge. The final chapter will summarize and revisit all that has been accomplished.

CHAPTER 10: CONCLUSION & RECOMMENDATIONS

A vast amount of material has been presented within the pages of this document. This chapter serves the purpose of attempting to tie it all together. This will be done by providing an overview of the dissertation, discussing the dissertation's contributions to the body of knowledge, identifying the applications and benefits of this research, and presenting some avenues for future research.

10.1 DISSERTATION OVERVIEW

A recap of what has been covered should help when placing the contributions into perspective. The dissertation was organized into four main parts. Figure 1-3 depicts these four parts as they relate to each other and the chapters of the dissertation.

10.1.1 Part I: Understanding the Challenge

Part I provided the frame of reference and context for the dissertation.

In Chapter 1, the topic and research was introduced. The hypotheses were put forth. The objectives, scope, limitations, and assumptions were presented. An outline of the dissertation was provided.

Chapter 2 provided valuable background information to aid in the understanding of economic modeling and the forecasting process. The chapter first investigated economic replacement theory. The two optimization theories, cost minimization and profit maximization were described and contrasted. Repair limit theory was also discussed. The works of Taylor, Hotelling, Preinreich, Terborgh, Douglas, and Collier & Jacques were discussed as they relate to economic modeling. The uses and types of economic forecasts were presented. Numerous methods of forecasting maintenance and repair costs on heavy equipment were described.

Chapter 3 was a detailed discussion of the cumulative cost model (CCM). The CCM was introduced by Vorster (1980). It combines the useful functions of economic replacement theory and repair limit theory in one model. The model can provide both numeric and graphical solutions to a number of equipment management problems. The decisions supported by the CCM include, but are not limited to: purchase, maintain, repair, capital rebuild, like-for-like replacement, production capacity replacement, and retire. The use of the model in making these decisions was discussed in detail. Decision rules were identified for each type of decision.

10.1.2 Part II: Defining The Work

Part II addressed the work to be accomplished by providing further details on the nature of the data and the analysis definition aspects of this dissertation. Chapter 4 gave an in-depth look at the data available and its idiosyncrasies. Chapter 5 followed with a detailed description of the test methodology.

The data used in this study were not perfect or ideal as pointed out in Chapter 4. They were field data obtained from real companies. There were structural and statistical issues concerning this data. The structural issues of field data, differing machines, machine age, differing times, data collection periods, cost, data pairing, and confidentiality were discussed. The bottom line with the structural issues was that different companies do things differently. To compare results on a like basis, the data needed to be placed into the same format for every fleet analyzed. This would allow for the formulation of CCI values that were consistent with other companies. Statistical issues discussed included: data independence, non-constant variance, relative dominance, repeated points, and varying intervals. Solutions to the structural and statistical issues were proposed. The four different data sets used in this dissertation were introduced.

Chapter 5 commenced with a discussion of the types of regression to be performed. This study was limited to linear regression models and non-linear models that could be transformed into linear models. For the linear models, regression through the origin was used. This forces the GEL to pass through the point (0,1) on the age/CCI axis system. A total of 19 different models were identified for consideration. Four were non-linear transformed, the rest were linear. The data were scaled to allow for a better relationship between the raw components of the models (x,

x^2, etc.). For fleets with more than 34 machines, data splitting was proposed as a validation technique. The analysis had to be broken down into phases since there were so many models and data sets under investigation. The preliminary phase used non-parametric techniques. The latter phases used parametric evaluations. SAS® was introduced as the primary research tool to be used for the data analysis.

Part I	Understanding the Challenge	1. Introduction 2. Literature Review 3. The Cumulative Cost Model
Part II	Defining The Work	4. The Data 5. Statistical Methodology
Part III	The Work	6. Data Gathering 7. Analysis 8. Results
Part IV	The Benefits	9. Integration 10. Conclusion & Recommendations
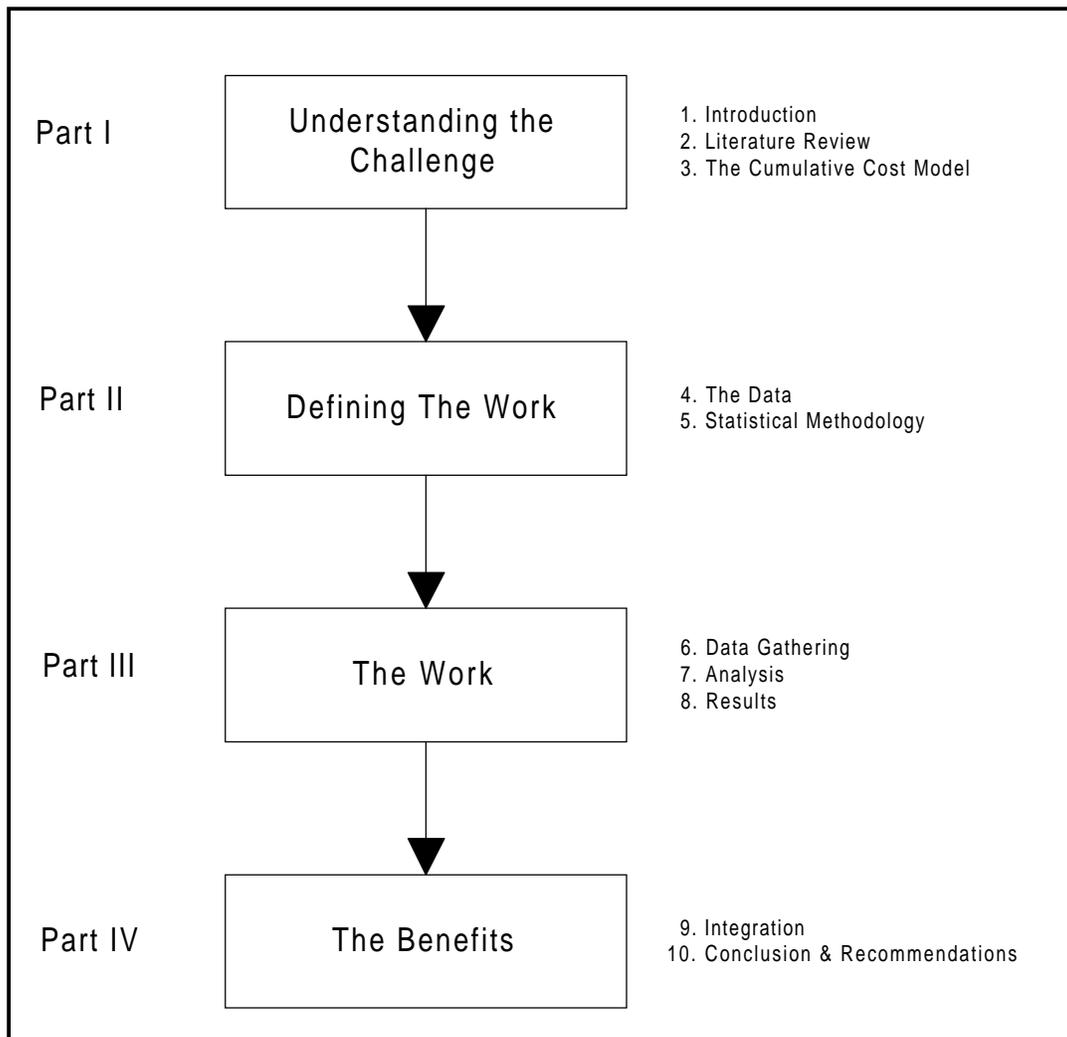
Figure 10-1: The Organization of the Dissertation

10.1.3 Part III: The Work

This part of the dissertation was where most of what was actually done was described. The complicated process of preparing the data for analysis was covered in Chapter 6. These prepared data were analyzed statistically in Chapter 7. Assessments were made about the usefulness of the results obtained during Chapter 8.

Preparing the data for analysis required a greater time investment than the analyses themselves. Chapter 6 described how multitudes of data on 270 different machines were extracted from the company databases. A number of manual corrections had to be made on the data after they were obtained. Obvious errors (such as negative repair costs for a given time period) had to be corrected. The data then had to be corrected for the effects of inflation. This was done using indices available from the Bureau of Labor and Standards. Since many of the companies involved did not explicitly track machine age in cumulative hours of use (the regressor variable), a method was devised to associate cumulative hours of use from oil-sampling database with the cumulative costs in the accounting databases. After this was accomplished, the four data sets for each of the 17 fleets were prepared.

In Chapter 7, the process of analyzing these data within the framework of the methodology defined in Chapter 5 was described. Eleven of the nineteen models under consideration were eliminated during the preliminary analysis using non-parametric techniques. The eight remaining models contained the best one, two, three, and four-parameter linear models in terms of both measures of performance (adjusted R^2 and R^2_{press}). The best transformed non-linear model was also included. Three of those eight models and one of the four data sets were eliminated upon examination of the average p-values for inclusion of parameters—their average p-values were greater than the 0.20 specified in Chapter 5. The models eliminated at this stage were those with more than two terms. The second stage of the intermediate analysis involved comparisons of the measures of performance for the different models. The two single-parameter models were eliminated due to measurably worse performance. The three remaining two-parameter models were compared on the basis of measures of performance, parameter significance, statistical issues, and preliminary results. The linear model that contained terms of x and x^2 was the model selected. The data set that contained data pairs interpolated at 500-hour intervals was selected as

the most appropriate. Cross-validations for this model were successful. Confidence intervals for the β terms were calculated. A preliminary study of the effects of weighted regression was accomplished.

In Chapter 8, the validity of the results obtained using the model selected in Chapter 7 was examined. A study of the equations revealed that there is an inverse relationship between the β_1 and β_2 coefficients. This could indicate that companies that invest in continuous maintenance and repair over the life of the machine should have smaller β_2 components. It was shown that optimum life, L^*, is solely a function of the β_2 term. T^* is a function of both parameters. It was proposed that there might be an empirical relationship between CCI and L^*. For the fleets in this study, two was the average CCI at the point that the fleets reached L^*. There is a curvilinear relationship between L^* and T^*. Large machines in heavy production roles tended to have larger L^* values and thus, lower T^* values (in terms of CCI). Smaller machines in multi-purpose roles tended to have smaller L^* values and larger T^* values. It was observed that the smaller L^* values were more in line with conventional thinking on when to replace machines. It was proposed that some of the inaccuracies in L^* and T^* for the larger equipment could be accounted for if collateral costs were included. Sensitivity analyses were performed to see how L^* and T^* vary with changes in parameter values. Comparisons of all fleets in the same company, similar fleets in different companies and similar types of fleets with differing sizes were performed. The statistical tests did not support any definitive conclusions about any of these comparisons. Some observations were made, but further testing should be done to support any conclusions. The performance of the regression equations in relation to three repair cost forecasting methods proscribed in literature was presented.

10.1.4 Part IV: The Benefits

The final part of the dissertation focused on the uses and contributions of the work performed. Chapter 9 provided a linkage between the equations developed and the CCM. It also provided information on how companies can derive their own equations. Chapter 10 recapped all that was accomplished.

The first tool presented in Chapter 9 was an application to assist users in making a capital rebuild decision. It was pointed out that there are three dimensions to such a decision: when the rebuild will be accomplished, how much it will cost, and how much machine life will have been gained when the rebuild is complete. These three dimensions were accounted for in a spreadsheet application that computes the GELs for the machine prior to the rebuild and after the rebuild based on user-supplied information. L* and T* values for both GEL curves are calculated and a decision can be made. The second tool described had the function of plotting the NEL in relation to the GEL, permitting the user to make comparisons. The calculations to compute the NEL were based on an empirical rule for residual value that was validated using actual resale and trade-in values for articulated haul units. The results indicate that there is a significant difference between the L* and T* values computed on the basis of the two different curves (NEL and GEL). After the two tools were presented, detailed instruction on how companies can develop their own equations were provided. This process can be accomplished within the capabilities of spreadsheet and database programs readily available for personal computers. Equations were presented that will allow the user to forecast average hourly repair costs and average period repair costs for fleets of a specific age. Finally, a framework for the development of industry benchmark equations was proposed.

Chapter 10 was the conclusion. The dissertation was summarized, the contributions were noted, and ideas for future research were presented.

10.2 CONTRIBUTIONS

This dissertation has provided important contributions to the body of knowledge concerning construction equipment economics. The contributions will be discussed briefly in terms of the hypotheses presented at the beginning of this dissertation. A more detailed review of the specific contributions will follow.

10.2.1 Hypotheses

This dissertation tested three different hypotheses. These hypotheses are interrelated—they build upon each other.

- Hypothesis #1: *A mathematical relationship exists between repair costs and age of heavy earthmoving equipment.*

  In fact, there were many suitable mathematical relationships between repair costs expressed within the cumulative cost index and age expressed in cumulative hours of use. Chapter 2 showed that many authors have attempted to quantify this relationship by various means (Nichols, Nunnally, etc.). Chapter 7 showed that there were many different suitable regression equations.

- Hypothesis #2: *It is possible to approximate the true equation for the relationship between cost and age by using linear regression techniques on existing data.*

  Chapter 7 of this dissertation presented the results of a detailed regression analysis to select the *best* regression equation for this purpose. The equation selected was:

  $$CCI = 1 + \beta_1 x + \beta_2 x^2 \qquad\qquad \textbf{Equation 10-1}$$

  It was determined that this equation used with a data set consisting of the CCIs of each machine interpolated to 500-hour intervals provided the best solution to the task.

- Hypothesis #3: *It is possible to incorporate repair cost regression equations into the Cumulative Cost Model (CCM).*

  This was proven to in Chapter 8 where it was shown how the L*, T*, and average hourly repair costs could be determined using the equations developed. Chapter 9 took the incorporation one step further by providing two tools that directly permit the visualization and quantification of the impact of the growth of repair costs within the CCM.

All three hypotheses were addressed. Significant evidence for their acceptance was provided.

10.2.2 The Contributions in Detail

In this section, the contributions will be discussed chapter by chapter.

Chapter 1 contributed a better understanding of the problems facing equipment managers. It also introduced the concept of a *Cumulative Cost Index* (CCI). The CCI is an invaluable tool in the comparison of machines that are not identical.

Chapter 2 combined pertinent information from the body of knowledge in a concise form that has sufficient breadth and depth to serve as an aid in the understanding of economic modeling and forecasting as they pertain to construction equipment.

Chapter 3 provided a fresh perspective on the Cumulative Cost Model (CCM) as developed by Michael C. Vorster. The myriad uses of the model were codified with understandable decision rules.

Chapter 4 contributed a detailed study into the nature of and problems with field data on construction equipment repair costs.

Chapter 5 presented an in-depth, statistically sound methodology for the development of regression equations using a state-of-the-art statistical software package (SAS®).

Chapter 6 showed how to process raw field data on construction equipment to a format that is suitable for analysis. A number of innovative techniques were presented. A process was identified whereby cumulative meter hour data could be associated with cumulative cost data through the use of oil-sampling databases.

Chapter 7 provided the single most significant contribution of this work—the selection of a regression model and recommendation of a data set for the quantification of the CCI in terms of cumulative hours of use.

Chapter 8 investigated the nature of this equation as applied to the data that were part of the study. There were a number of important contributions in this chapter. It was proposed that the β_1 component of the equation represents a static cost accumulation that is, in essence, a fact of life relating to the ownership of equipment. The β_2 component, on the other hand, represents a dynamic cost growth accumulation—that could possibly be a reflection of how well a company manages its maintenance and repair strategy.

It was shown that there is a significant relationship between the two β terms in the equation. The relationship is inverse—a relatively low value for one coefficient usually resulted in a relatively high value for the other coefficient. This relationship resulted in an even more significant relationship between the optimum life (L*) and optimum average cost per period (T*) values for differing fleets. There is an L* vs. T* continuum along which all fleets in the study were located.

The L* values which were lower seemed to provide more realistic estimates of optimum life than the higher L* values. It was proposed that collateral costs could be the discriminator. Collateral costs may not be that significant in the determination of L* for fleets of smaller, general purpose type equipment. Collateral costs may have a large impact on the L* values for fleets of larger, production-oriented equipment.

There is a strong relationship between CCI and L*. Most machines reach L* with a CCI value of approximately two. In general terms this could mean that a machine approaches the end of its economic life when 100% of the purchase price of the machine has been invested in repairs on that machine.

It was shown that the equations for estimating repair costs proposed by Nunnally (1993) do a good job of fitting CCI curves if they are given a starting point. The benefit of the cumulative repair costs curves developed in this dissertation is that no seed value is required. Optimizations can be performed without guessing at a starting point.

Chapter 9 provided two spreadsheet applications for the direct use of the cumulative repair cost equations within the CCM. One of these applications was an aid to making the rebuild decision. The other was a preliminary investigation of the Net Expenditure Line (NEL) based on historic residual values. A detailed guide on how companies can develop their own cumulative cost equations was provided. A framework for the establishment of industry-standard equations was presented.

10.3   APPLICATIONS AND BENEFITS

This research was pertinent and has produced some direct applications than can be applied in the construction industry. Construction firms that use heavy equipment should consider developing and employing their own cumulative cost equations. The equations can be developed within the constraints of existing data collection systems. All that is required is a personal computer with standard software (spreadsheet and database.) The regressions can be accomplished within the spreadsheet program—expensive scientific tools like SAS® are not required to develop equations.

The equations can be used to directly estimate average to date, average incremental, or average period repair costs or repair cost accumulation rates for specified fleets of equipment. The equations can also be employed within available applications of the CCM.

The benefits of using these equations include a better understanding of how repair costs accumulate as machines age. Equipment managers will be able to produce better estimates of average repair costs for their fleets of equipment. Better estimates can translate into less uncertainty about profit for the company under the competitive bidding process. Applications within the CCM can help the equipment manager maintain an optimum fleet of equipment. The CCM can help an equipment manager make decisions concerning acquisitions, maintenance, repairs, rebuilds, replacements, and retirements.

10.4   RECOMMENDATIONS FOR FUTURE RESEARCH

Throughout the course of this research, a number of areas were identified that could provide fruitful results if investigated further.

Definition of the NEL. A comprehensive study of residual values for construction equipment should be undertaken. Regression equations that can express residual value in terms of cumulative hours of use would provide a very important contribution to the cumulative cost model. All decisions cannot be made solely on the basis of the GEL.

Further Define GEL. The GEL might be further defined and made more accurate through the inclusion of more cost categories. All possible costs should be investigated as to the impact they

have on the determination of L* and T*. The quantification of collateral costs has proven to be a difficult and subjective task. It may be possible to reverse-engineer the collateral cost portion of the true GEL. This could be done with the help of experienced equipment managers. It would be necessary to assume that experienced equipment managers are able to incorporate collateral costs into the decision making process without solid, balance-sheet type numbers in front of them. If L^*_{actual} for a number of fleets can be provided by these equipment managers, the β terms within the equations can be adjusted to make $L^*_{predicted} = L^*_{actual}$. As a starting technique, β_1 should be held constant while varying β_2. It is felt that collateral costs grow at an increasing rate with the accumulation of hours.

Define Industry-Standard Benchmarks. The means for doing this were presented in Chapter 9. Industry-standard benchmarks would be invaluable if they can be developed. They could provide a basis against which to judge actual performance of a companies fleets or, more importantly, its maintenance and repair policies and strategies. The benchmarks could also lead to more concrete generalizations about concerning type and size of equipment. Additionally, such benchmarks could be employed by companies that do not have adequate decision support systems as aids to their decision making process.

Investigate other attributes. The attributes investigated during this study were equipment size, company, and type. It may be useful to study new vs. used equipment, brand "A" vs. brand "B" equipment, or the impacts of geographic location.

Fully develop tools for applications within the CCM. Prototypes of two of these tools were provided in Chapter 9. The tools for the rest of the equipment management decisions possible within the CCM should also be developed. The tools should be combined in one application that allows the user to access many different types of analyses with the touch of a button.

Further investigate important relationships. Relationships that merit further study are: CCI values at L*, the L* vs. T* continuum, and the β_1 vs. β_2 continuum.

Investigate other applications. The techniques developed for this research may be applicable to other industries besides construction. The mining industry, in particular, should be investigated.

10.5 CLOSURE

This dissertation has taken an in-depth, focused look at one central issue: quantifying the effect of machine age on the growth of repair costs. This issue was addressed through the use of regression analysis techniques. A suitable solution was found within the research objectives, scope, and limitations delineated at the beginning of this document.

The equations that quantify this effect have meaning beyond just a strict mathematical relationship. They provide a bridge that enables current data collection techniques to be used within the context of the cumulative cost model. This will eventually permit the direct application of economic theory to daily equipment management practices.

REFERENCES:

1. Anderson, D., Sweeney, D., Williams, T. (1991). *An Introduction to Management Science.* West Publishing Co., St. Paul, MN.

2. Birch, J. (1996). Statistics 5514 Course Notes.

3. Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control.* Prentice-Hall Inc., Englewood Cliffs, NJ.

4. Bureau of Labor Statistics, "Consumer Price Indexes Home Page", internet site address: http://stats.bls.gov/cpihome.htm, 6 Jun 1997

5. *Business Conditions Digest.* (1981). US Department of Commerce, Washington, DC.

6. Capen, E., Clapp, R., Campbell, W. (1971). "Competitive Bidding in High-Risk Situations." Society of Petroleum Engineers SPE 2993.

7. Carroll, R. J., and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New York, NY.

8. Casella, G. (1983). "Leverage and Regression Through the Origin." *The American Statistician*, 37 (2), 147-152.

9. *Caterpillar Performance Handbook.* (1995). Caterpillar, Inc., Peoria, IL.

10. Chester, T. (1995). *Mastering Excel for Windows 95.* Sybex, Inc., Alameda, CA.

11. Chiaccerini, L. (1996). Virginia Tech Statistical Consulting Center. Numerous visits between August 1996 and November 1996

12. Collier, C. A., Jacques, D. E. (1984). "Optimum equipment life by minimum life-cycle costs," Journal of Construction Engineering and Management, 110 (2), 248-265.

13. Consumer and Producer Price Indices. Bureau of Labor and Standards, Internet database, 1996.

14. *Cost Reference Guide for Construction Equipment.* (1996). K-III Directory Corporation, San Jose, CA.

15. Cox, E. A. (1971). "Equipment Economics." *Handbook of Heavy Construction*, J. A. Havers and F. W. Stubbs, eds., McGraw-Hill, New York, NY.

16. de Neufville, R. (1990). *Applied Systems Analysis*. McGraw-Hill, New York, NY.

17. Deshpande, J. V., Gore, A. P., Shanubhogue, A. (1995). *Statistical Analysis of Nonnormal Data*. John Wiley and Sons, New York, NY.

18. Douglas, J. (1970). "Replace For More Profit." *Construction Methods and Equipment*, Mar. & Apr. 1970.

19. Douglas, J. (1975). *Construction Equipment Policy*. McGraw-Hill, New York, NY.

20. Drinkwater, R. W., and Hastings, N. A. J. (1967). "An economic replacement model." *Op. Rsrch. Qtrly.*, 18(2), 121-138.

21. Drinkwater, R. W., and Hastings, N. A. J. (1967). "An economic replacement model." *Op. Rsrch. Qtrly.*, 18(2), 121-138.

22. *EP 1110-1-8 V.II.* (1995). Department of the Army, Washington, DC.

23. Fabricant, S. (1976). "Economic Calculation Under Inflation: The Problem in Perspective." *Economic Calculation Under Inflation*, Liberty Press, Indianapolis, IN.

24. *Fiatallis Owning and Operating Costs.* (1983). Fiatallis, Inc.

25. Freund, R. J., Littell, R. C., (1991). *SAS® System for Regression*. SAS Institute, Cary, NJ.

26. Freundlich, H. (1926). *Colloid and Capillary Chemistry*, Methuen, London, England.

27. Gibbons, J. D. (1993). *Nonparametric Statistics—An Introduction*. Sage University, Newbury Park, CA.

28. Gibson, J. L., Ivancevich, J. M., Donnelly, J. H. (1991). *Organizations: Behavior, Structure, Processes*. Irwin, Boston, MA.

29. Grant, E. L., Ireson, W. G., and Leavenworth, R. S. (1990). *Principles of Engineering Economy*, 8th Ed., John Wiley & Sons, New York, N.Y.

30. Grant, E., Ireson, W., Leavenworth, R., (1990). *Principles of Engineering Economy*. John Wiley and Sons, New York, NY.

31. *Green Guide for Construction Equipment*. (1996) K-III Directory Corporation, San Jose, CA.

32. Hahn, G. (1977). "Fitting Regression Models with No Intercept Term." *Journal of Quality Technology*, 9 (2), 56-61.

33. Hanke, J. E., Reitsh, A. G. (1995). *Business Forecasting*. Prentice-Hall, Englewood Cliffs, NJ.

34. Harris, F. C., and Olomolaiye, P. O. (1993). "A theoretical model for determining equipment service life." *Building Research and Information*, E. & F. N. Spon, 21(4), 243-245.

35. Hastings, N. A. J. (1969). "The repair limit replacement method," Operational Research Quarterly, Operational Research Society, London, England, 337-349.

36. Hotelling, H. (1925). "A general mathematical theory of depreciation," Journal of the American Statistical Association, ASA, 151, 340-353.

37. Howard, R. (1966). "Information Value Theory." *IEEE Transactions on Systems Science and Cybernetics*, 2 (1), 22-26.

38. Howard, R. (1967). "Value of Information Lotteries." *IEEE Transactions on Systems Science and Cybernetics*, 3 (1), 54-60.

39. Jaafari, A., Mateffy, V. K. (1990). "Realistic Model for Equipment Replacement", Journal of Construction Engineering and Management, ASCE, 116 (3), 514-532.

40. Jones, B. W. (1982). *Inflation in Engineering Economic Analysis.* J. Wiley, New York, NY.

41. Kannan, G. (1997). Unpublished research summary dealing with automated data collection prepared for construction affiliates meeting, Virginia Tech, Blacksburg, VA.

42. Kapoor, M. (1996). "Visualization of Data in the Construction Industry." Unpublished Project and Report, Virginia Polytechnic Institute and State University.

43. Kim, Y. H. (1989). "A forecasting methodology for maintenance cost of long-life equipment", Dissertation presented to the University of Alabama in partial fulfillment of the requirements for the degree of Doctor of Philosophy

44. Mahon, B. H., Bailey, R. J. M. (1975). "A proposed improved replacement policy for army vehicles." Operational Research Quarterly, Operational Research Society, London, England, 26 (3), 477-491.

45. Makridakis, S., Wheelwright, S. C. (1989). *Forecasting Methods for Management*. John Wiley and Sons, New York, NY.

46. Mays, J. (1996). Statistics 5616 Course Notes.

47. *Means Building Construction Cost Data.* (1995). R. S. Means Company, Inc., Kingston, MA.

48. *Minitab Reference Manual*. (1994). Minitab, Inc., State College, PA.

49. Mitchell, Z. W. (1997). "Economic Decisions in Equipment Management", Unpublished article. Virginia Tech, Blacksburg, VA.

50. Mitchell, Z. W. (1997). Unpublished survey of construction equipment managers. Virginia Tech, Blacksburg, VA.

51. Montgomery, D. C., Peck, E. A. (1992). *Introduction to Linear Regression Analysis.* John Wiley & Sons, New York, NY.

52. Myers R. H., Milton, J. S. (1991). *A First Course in the Theory of Linear Statistical Models.* PWS Kent Publishing, Belmont, CA.

53. Myers, R. H. (1990). *Classical and Modern Regression with Applications.* PWS Kent Publishing, Belmont, CA.

54. Noble, R. (1997). Unpublished "NOINT" SAS® IML® macro. Written for this research through the Statistical Consulting Center, Virginia Tech, Blacksburg, VA.

55. Nichols, H. L. (1976). *Moving the Earth.* North Castle Books, Greenwich, CT.

56. Nunnally, S. W. (1987). *Construction Methods and Management.* Prentice-Hall, Englewood Cliffs, NJ.

57. Nunnally, S. W. (1993). *Construction Means and Methods*. Prentice-Hall, Englewood Cliffs, NJ.

58. Ott, L. R. (1993). *An Introduction to Statistical Methods and Data Analysis.* Duxbury Press, Belmont, CA.

59. Paulson, B. C. (1995). *Computer Applications in Construction*. McGraw-Hill, New York, NY.

60. Peurifoy, R. L., Ledbetter, W. B., and Schexnayder C. J. (1996). *Construction Planning, Equipment, and Methods*, McGraw-Hill, New York, NY.

61. Preinreich, G. A. D. (1940). "The economic life of industrial equipment." *Econometrica*, 8 (1), 12-43.

62. Ratkowsky, David A. (1983). *Nonlinear Regression Modeling.* Dekker, New York, NY.

63. Ratkowsky, David A. (1990). *Handbook of Nonlinear Regression Models.* Dekker, New York, NY.

64. Rawlings, J. O. (1988). *Applied Regression Analysis: A Research Tool*. Wadsworth, Inc., Belmont, CA.

65. Rencher, A. C. (1993). *Methods of Multivariate Analysis*. J. Wiley & Sons, New York, NY.

66. *SAS/STAT® User's Guide* (1989). SAS Institute., Cary, NC.

67. *SAS® Technical Report P-229*. (1992). SAS Institute, Inc., Cary, NC.

68. Schulman, R. (1996). Statistics 5615 Course Notes.

69. Snaddon, D. R. (1990). "Maintenance Practices of Civil Construction Firms." *The Civil Engineer in South Africa,* 32(2), 57-63.

70. Snee, R. (1977). "Validation of Regression Models: Methods and Examples." *Technometrics*, 19 (4), 415-428.

71. Sprent, P. (1993). *Applied Nonparametric Statistical Methods.* Chapman and Hall, New York, NY.

72. Taylor, J. S. (1923). "A Statistical Theory of Depreciation", Journal of the American Statistical Association, ASA, 144, 1010-1023.

73. Terborgh, G. W. (1949). *Dynamic Equipment Policy.* McGraw-Hill, New York, NY.

74. *Terex Owning and Operating Costs*. (1981). Terex, Inc.

75. Vorster, M. C. (1980). "A systems approach to the management of civil engineering construction equipment." Thesis presented to the University of Stellenbosch, at Stellenbosch, South Africa, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

76. Vorster, M. C. (1986). "Equipment Management and Decision Making." Course notes.

77. Vorster, M. C. (1995). ACEM Executive Development Program at Virginia Tech, Course Notes.

78. Vorster, M. C., de la Garza, J. M. (1990). "Consequential equipment costs associated with lack of availability and downtime." Journal of Construction Engineering and Management, ASCE, 116(4), 656-669.

79. Vorster, M. C., Sears, G. A. (1986). "Model for retiring, replacing, or reassigning construction equipment." Journal of Construction Engineering and Management, ASCE, 113(1), 125-137.

80. Wilson, J. H., Keating, B. (1994). *Business Forecasting*. Irwin, Burr Ridge, IL.

Appendix A: Inflation Corrections

There are two general ways to account for inflation in economic calculations. They are known as current value accounting and price level accounting (Fabricant, 1976). *Current value accounting* attempts to incorporate general cost indices and specific appraisals to come up with a somewhat subjective value of the current market worth of specific goods or services. Current value accounting is market driven--the same assets could have very different current values in different markets (regions of the country). *Price level accounting* quantifies changes in the value of goods or services by incorporating fluctuations in the general purchasing power of the dollar. Price level accounting is the most appropriate and feasible method to use for this study.

The general formula to calculate inflated costs is (Jones, 1982):

$$p(t + \Delta t) = p(t) \cdot [1 + f]$$
Equation 0-1

Where $p(t + \Delta t)$ is the price of goods or services at some time in the future, $p(t)$ is the current price of goods or services, and f is the inflation factor for the given period of time. Unfortunately, f is not easy to define and can be different for different commodities. A better computational form of the inflation equation is given by the equation (Jones, 1982):

$$p(t_2) = p(t_1) \cdot \frac{I(t_2)}{I(t_1)}$$
Equation 0-2

Where $I(t)$ is an index that is specific to time t. In this equation, t_1 denotes the date that a transaction occurred—this will be called the *transaction date*. The other time parameter, t_2, denotes the time to which the transaction will be indexed, or the *base date*. These indices can be computed or obtained from existing sources. The US Bureau of Labor and Statistics computes a variety of statistics that are of great value when trying to estimate inflation rates (Business, 1982). Among these are the often-mentioned *Consumer Price Index* and *Producer Price Index*. The Consumer Price Index is based on the general prices of consumer goods. It is a good estimator for labor costs as many unions try to tie their wage increases to increases in this index. The Producer Price Index attempts to capture changes in the cost of producing goods. The Producer

Price Index is further broken down into broad classes of manufactured goods, the most appropriate of which is "Construction Machinery and Equipment." The periodical *Engineering News Record* (ENR) also publishes quarterly indices for general construction costs and equipment costs.

The best index to use for this study could be a composite one. In his book, *Construction Equipment Policy*, James Douglas recommends a composite index that contains mixes of indices for machinery price, prime rate of bank loans, labor, parts cost, petroleum, and overhead (Douglas, 1975). These indices are weighted, then applied to the overall operating cost to come up with an inflation correction. A similar composite index can be developed that is tailored to this research.

Adjusted Cost Indices
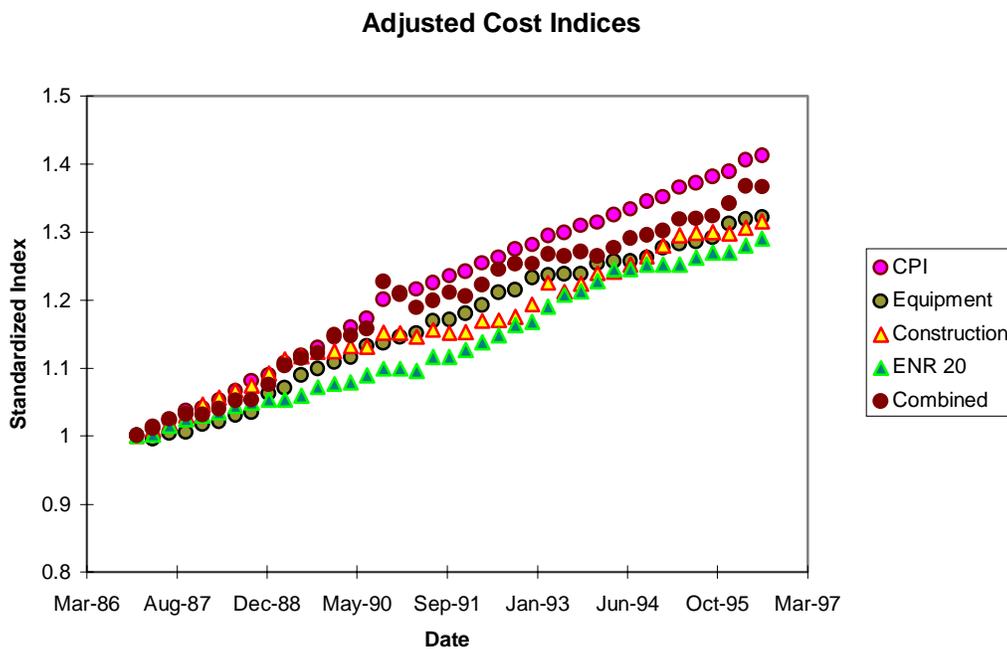
Figure A-1: Standardized Cost Indices (Bur. Labor & Stds., ENR)

All of the factors that Douglas recommends should not have to be taken into account for this research. Overhead and bank loans are not as important to this research as they would be to research that is looking at the entire equipment equation. An index that would seem to make sense for this research would be one that incorporates the cost of construction equipment and

245

labor. The initial purchase price of the machine could be indexed using solely an index for equipment. The repairs that take place would be indexed to the cost of equipment and labor in appropriate ratios.

Using data from one of the fleets in this study, estimates for the appropriate percentages of these items were developed. Labor was 45% of the repair costs and parts were 55% of the repair costs. The indices chosen to represent these two categories were obtained from the Producer Price Index Series series "construction machinery" (ID # PCU3531) and the Consumer Price Index series "all urban consumers" (ID # CUUR0000SA0) (Bureau of Labor and Standards, 1996).

These data are easily obtainable through the internet. The main internet address of the Bureau of Labor and Standards is http://www.bls.gov. The series are obtained from their statistical division. The current website for the indices is http://146.142.4.24/cgi-bin/surveymost?bls. The website has an interactive menu for selecting the information desired.

Data from the *Engineering News Record*, while developed specifically for the construction industry, does not differ significantly from that obtained from the Bureau of Labor and Standards (Figure A-1) and is not readily available in electronic format.

The indices shown in Figure A-1 include the Consumer Price Index, Producer Prices Index for construction machines, ENR top 20 U.S. cities construction index, the Bureau of Labor and Standards' construction cost index, and the combined index proposed earlier in this paper.

In their raw form, the indices had ranges from 0.9 to 530 depending on which index and which time period was being looked at. The reason for this is the indices had different base dates. The base date is the point where the index is equal to one—everything else is indexed to that date. To give a common start point for comparison purposes, all indices were adjusted to reflect January 1987 as the base date. Data from January 1987 to the present were plotted. This range of values covers the range of interest for the data used in this study. The two construction indices remain very close throughout the range of interest. The CPI increases at a rate slightly faster than most of the other indices, but all remain fairly closely grouped.

It should not matter which point in time is chosen as the base date—as long as all the transactions for the fleet are indexed to the same date. The reason it does not matter is that the numerator and the denominator of the CCI equation are both indexed to the *same* base date. The CCI is a unitless number.

The effect of inflation is substantial. Most of the indices show almost a 30% increase over the ten years of observation. This would mean that a repair that cost $100 in 1987 would cost around $130 in 1997. A correction of 30% must be applied to the later costs incurred. If it is not applied, it will not be possible to determine what happens to equipment repair costs in terms of real spending power.

The indices are applied to the data using Equation B above. The initial list price is adjusted once using the equipment index. The incremental monthly repair costs are adjusted using the combined index for the month in which they occurred. A problem arises when the cumulative repair data available on a machine starts at some time other than the initial purchase date. For the machines that fall into this category, the first value of cumulative repair cost is indexed to the halfway point of the range calendar months preceding it. This is not ideal, but some index must be applied to this figure. After the indices are applied, the CCI's are calculated and the equations can be developed.

The effect of the application of these indices should normally be a de-emphasis of the quadratic trends of the regression lines developed. This means that the β_2 term should be smaller than it would have been had the inflation correction not been made. Smaller β_2 terms correlate directly to larger L* values. The T* values for the adjusted line should be smaller.

Using one of the data sets from one of the fleets, trial regressions were performed to ascertain the numerical and graphical significance of the effects of inflation. Figure A-2 shows plots of the data, adjusted for inflation and not adjusted for inflation. The regression lines for each set of data are also depicted. The regression line for the adjusted data is flatter than that of the unadjusted data. The values obtained from this regression indicated a 16% increase in L* and a 15% decrease in T* when the data were adjusted.
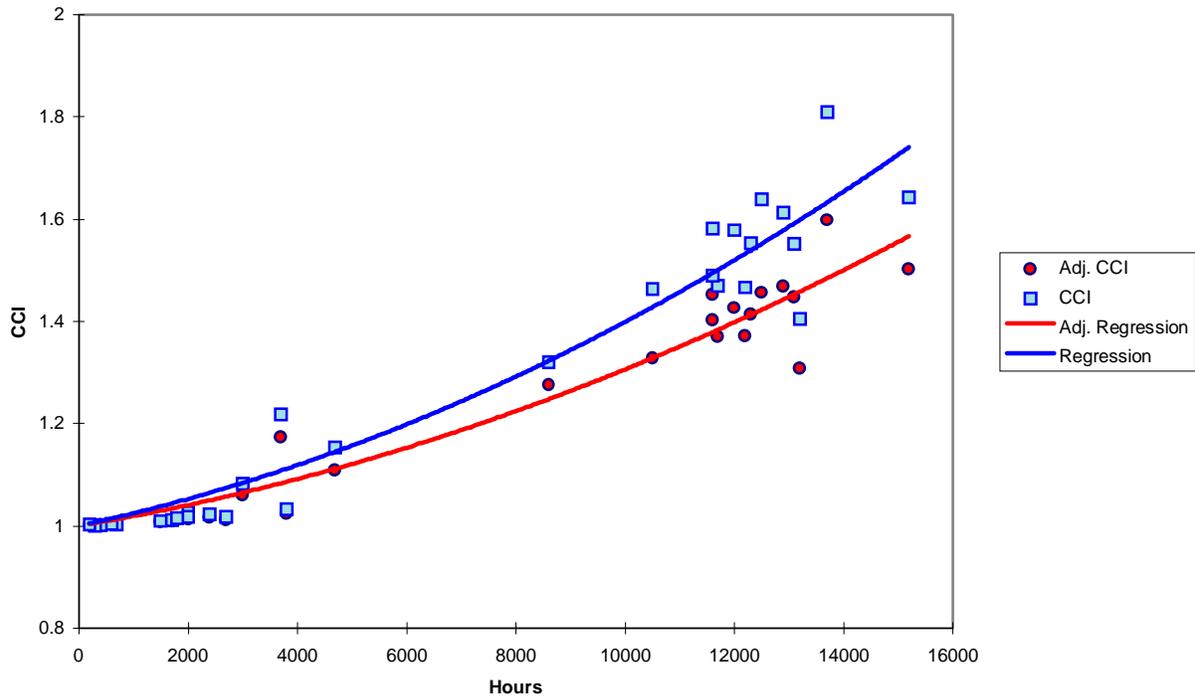
Figure A-2: - Regression Comparison

Inflation should not be ignored in this, or any other, economic forecasting model. The inflation indices during the time frame of interest for this study are not trivial. It has been demonstrated by example what kind of effect inflation has on results obtained. The impact is certainly measurable.

Appendix B: NOINT Macro

Shown below is the SAS® NOINT macro used in this research. It was originally developed by Robert Noble through the Virginia Tech Statistical Consulting Center. It was modified slightly by Zane Mitchell to adjust the PRESS statistic in the original macro to an R^2_{press} statistic.

```
options ls=200;

data d1;
input x y;
cards;

(Data in two columns go here)

;
%macro nointreg(data=d1);
%let data=d1;
proc iml;
  compnum=5;
  eqtype=83;
  setnum=3;
  use &data;
  read all into data;
  x = data(|,1|); y = data(|,2|);
  x = x || (x##2) || (x##3) || exp(x);
  ssep = (y - x*inv(x`*x)*x`*y)` * (y - x*inv(x`*x)*x`*y);
  s2 = ssep/(nrow(y)-4);

  result = .||.||.||.||.||.||.||.||.||.||.;
  do v1 = 0 to 1;
   do v2 = 0 to 1;
    do v3 = 0 to 1;
     do v4 = 0 to 1;
      check = v1 + v2 + v3 + v4;
      if check <> 0 then do;
       z=j(nrow(x),1,1);
       if v1 = 1 then z = z || x[,1];
       if v2 = 1 then z = z || x[,2];
       if v3 = 1 then z = z || x[,3];
       if v4 = 1 then z = z || x[,4];
       z = z[,2:ncol(z)];

       /* parameter estimates, ... */
```

```
        b = inv(z` * z) * z` * y;
        h = z*inv(z` * z)* z`;
        y_hat = h*y;

        /* sums of squares */
        sse = (y - z*b)` * (y - z*b);
        uss = y`*y;
        css = uss - nrow(y) * (sum(y)/nrow(y))**2;
        ssr = css - sse;

        /* degrees of freedom */
        dftot = nrow(y);
        dfreg = ncol(z);
        dferr = dftot - dfreg;

        /* regression stats */
        mse = sse / dferr;
        rsq = 1 - sse/css;
        adjrsq = 1 - mse*dftot/css;
        cp = sse/s2 - (nrow(y) -2*ncol(z));

        press = 0;
        do i = 1 to nrow(z);
          press = press + ((y[i,1]-y_hat[i,1])/(1-h[i,i]))**2;
        end;
        rsqpress = 1 - press/css;


        /* create output vector */
        temp = .||.||.||.;
        bloc = v1||v2||v3||v4;
        parm = 1;
        do i = 1 to 4;
         if bloc[1,i] = 1
           then do;
             temp[1,i]=b[parm,1];
             parm=parm+1;
           end;
        end;

        result                                        =
result//(temp||mse||rsq||adjrsq||cp||rsqpress||setnum||compnum||eqtype);
        end;
       end;
  end;end;end;end;
  result = result[2:nrow(result),];
```

```
   create statout var {x x2 x3 exp_x mse rsq adjrsq cp rsqpress setnum compnum
eqtype};
   append from result;
   close statout;

proc sort data=statout;
   by mse;
proc print data=statout noobs;
   var x x2 x3 exp_x mse rsq adjrsq cp rsqpress setnum compnum eqtype;
   title 'Results Sorted by MSE or adjusted R-square';
run;

title;
run;
%mend;

%nointreg(data=d1);
quit;
```

Appendix C: SAS® code

The following code was used to perform the intermediate analyses:

```
OPTIONS NODATE LS=120;
TITLE '';
DATA FLEET;
INPUT CUMHOURS CCI;
X = CUMHOURS;
X2 = X**2;
X3 = X**3;
EX = exp(x);
LX = log(x);
Y = CCI;
LY= log(Y);
CARDS;

(Data in two columns go here)

;
TITLE '1';
TITLE2 '1';
TITLE3 '1';
PROC REG;
      MODEL Y = X /NOINT P CLM CLI SS2 SS1 R INFLUENCE;
      MODEL Y = X X2 /NOINT P CLM CLI SS2 SS1 R INFLUENCE;
      MODEL Y = X X2 X3/NOINT P CLM CLI SS2 SS1 R INFLUENCE;
      MODEL Y = X X2 X3 EX/NOINT P CLM CLI SS2 SS1 R INFLUENCE;
      MODEL Y = X X3/NOINT P CLM CLI SS2 SS1 R INFLUENCE;
      MODEL Y = X2 X3 EX/NOINT P CLM CLI SS2 SS1 R INFLUENCE;
      MODEL Y = X2/NOINT P CLM CLI SS2 SS1 R INFLUENCE;
      MODEL LY = LX /NOINT P CLM CLI SS2 SS1 R INFLUENCE;

QUIT;
;
```

Vita

Zane Windsor Mitchell Jr. was born in Columbus, Ohio on 24 October 1961 to Zane W. Mitchell (SMSgt., USAF retired) and Dolores C. Mitchell. He lived in many different places while growing up, but spent most of his teen years in Carmichaels, Pennsylvania where he graduated from high school in 1979.

He attended the United States Air Force Academy from 1979 to 1983, when he graduated with a Bachelor of Science degree in Civil Engineering. Upon graduation, he was commissioned as an officer in the United States Air Force and attended Undergraduate Pilot Training. He is currently a Senior Pilot with over 3500 military flying hours.

He obtained an MBA from Rensselaer Polytechnic Institute, NY in 1992. He received his MS in Civil Engineering from Virginia Tech in 1993. He served as an Instructor and Assistant Professor of Civil Engineering at the United States Air Force Academy from 1993-1995.

From 1995 to the present he has been working on his Doctoral degree at Virginia Tech. Upon degree completion he will return to the United States Air Force Academy where he will serve as an Assistant Professor and Chief of the Construction Practices division in the Department of Civil Engineering.

Zane Mitchell married Robyn J. Krutchkoff of Blacksburg, Virginia on 17 December, 1993. They have twin 3-year-old boys—Keith D. Mitchell and Bryan K. Mitchell. Keith and Bryan love construction equipment as much as their father.