Chapter 1

Problem Statement

1.1 Introduction

Maintenance engineering is one of the fasted growing engineering disciplines in the world. Industry has only started to realize the importance of maintenance in the early 1980's and, ever since, there was no turning back the rapid development in the theory of maintenance. This theory is also more readily accepted by maintenance practitioners in industry as the mindset with regards to maintenance changes and greater successes are achieved by formal maintenance programs.

As is the case with most engineering disciplines, there is a drive in the field of maintenance engineering to optimize methodologies and practices. The maintenance fraternity has realized that the use of formalized maintenance models and tactics alone are not necessarily the optimal way to maintain equipment. One aspect of formal maintenance that needs optimization is decision making in life-limiting maintenance strategies, i.e. preventive maintenance, because of enormous losses industries are suffering due to a waste of residual life of equipment.

Preventive maintenance practitioners* have mostly reasoned along one of two schools of thinking. The first is to take action (replacement, repair or overhaul) based purely on an item's age as measured in time, miles, tons processed or any other convenient process parameter. The second is to assess the condition of an item through diagnostic measurements, which may include vibration monitoring, results of oil analysis, thermographic profiles, pressure, temperature, etc. This second viewpoint is referred to as predictive maintenance. Coetzee (1997) compiled a maintenance strategy tree that serves as a concise summary of possible

*Preventive maintenance, contrary to popular believe, is not necessarily the optimal maintenance strategy to apply. Any strategy's technical and economical feasibility should be determined before it is implemented. A methodology such as Reliability Centered Maintenance (RCM) or Total Productive Maintenance (TPM) should lead maintenance practitioners to the correct strategy.

maintenance strategies. See Figure 1.1.

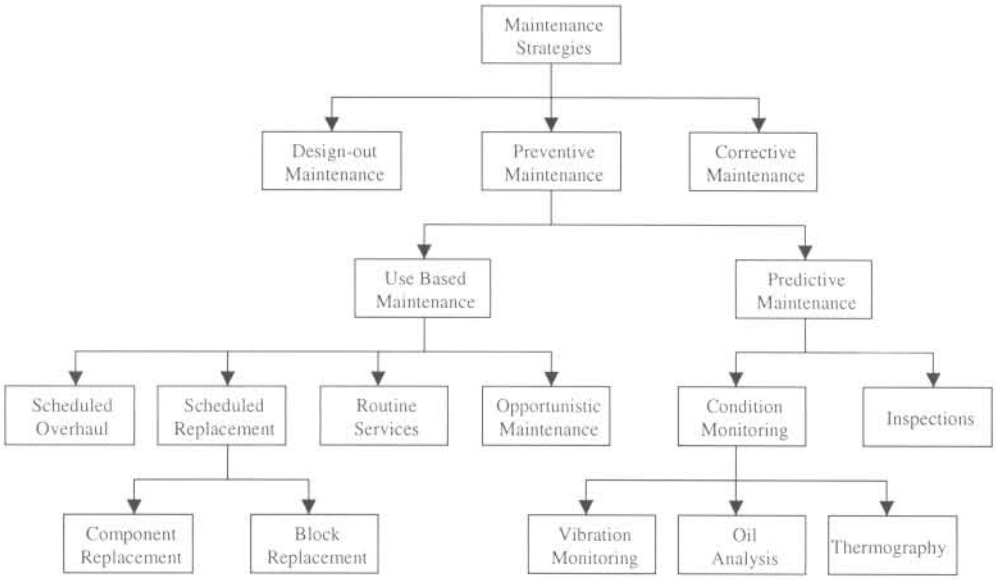

Figure 1.1: Maintenance Strategy Tree

Preventive maintenance is performed for one reason only: to prevent unexpected failure, which is in most cases considerably more expensive than planned preventive action. Unexpected failures often involve costly secondary damage to equipment, production losses, late delivery penalties, overtime labor costs and even loss of life. Preventive action is usually inexpensive relative to corrective action because of the planned nature of this type of action that eliminates many of the unwanted cost factors associated with unexpected failure.

In the case of use based maintenance, action is taken (by definition) only when an item has reached a certain age [†]. The time at which action is taken should be chosen in such a way that acceptably little residual life is wasted in the process but also such that the risk of unexpected failure does not rise unacceptably high. Optimization of use based maintenance thus involves a tradeoff between the waste of residual life and the risk of suffering an unexpected failure.

Predictive maintenance technologies, on the contrary, strive through continual [‡] assessment of an item's condition to warn those concerned of an imminent failure shortly before occurrence of failure. With advanced technology available at present, this seems to be a much more elegant approach than use based action. Closer investigation reveals that this is not necessarily true because, even with the advanced technology, there are still numerous unknowns

[†] *Time* will be used consistently to refer to an item's age but it should be emphasized that any convenient process parameter may be used

[‡] The word *continual* may be replaced by *continuous* in some cases

that cannot be eliminated and a tradeoff has to be made again. In this case the tradeoff is between the accuracy of the technology utilized to perform the condition assessment and the risk of running into an unexpected failure.

Both use based maintenance and predictive maintenance procedures have been optimized individually but very little work has been done to combine the advantages of the two schools of thought to produce an optimal solution. In this thesis, an methodology will be developed to merge the advantages of the two approaches into one approach that can be used as an authoritative decision making tool.

1.2 Conventional use based maintenance optimization

Lawrence (1999) studied mathematical use based optimization techniques in maintenance and concluded that most models address one of three questions,

(i) How often should a component be replaced?
(ii) How many spare parts should be kept in stock?
(iii) How should maintenance tasks be scheduled?

This section (and thesis) addresses point (i).

Many authors agree that the only scientific way to optimize use based maintenance strategies is through statistical analysis of event data. In this section, conventional optimization techniques are discussed, i.e. optimization through statistical models without covariates [§] or discontinuities. This field is poorly understood by maintenance practitioners, mainly because of the confusing terminology found in the literature. It is thus very important to define clear notation before any further discussion on optimization of used based maintenance through statistical modeling.

1.2.1 Terminology

The terminology of Ascher and Feingold (1984) will be used in this thesis. Ascher and Feingold's book was specifically written with the objective to clear some of the confusion in the field of statistical failure analysis. First of all it is important to distinguish between different types of items:

[§] Covariates are often also referred to as *explanatory variables*.

(i) *Part.* An item that is never disassembled and is discarded after first failure [¶].

(ii) *Socket.* A space that, at any given time, holds a part of a given type.

(iii) *System.* A collection of two or more sockets with their associated parts that is interconnected to perform a specific function(s).

(iv) *Non-repairable system.* A system that is discarded the first time it ceases to perform satisfactory, i.e. after first failure.

(v) *Repairable system.* A system that, after failure, can be restored to perform all of its function by any method other than complete replacement of the system.

After a system is repaired it could be in one of the following states:

(i) As good as new (GAN).

(ii) As bad as old (BAO).

(iii) Better than old but worse than new (BOWN).

(iv) Worse than old (WO).

The GAN and BAO assumptions are by definition the backbone of conventional statistical failure data analysis. Models with covariates or discontinuities are required to model BOWN or WO situations.

It is also very important to define appropriate time scales to measure life times of items. See Figure 1.2 for an example sample path of a failure process.

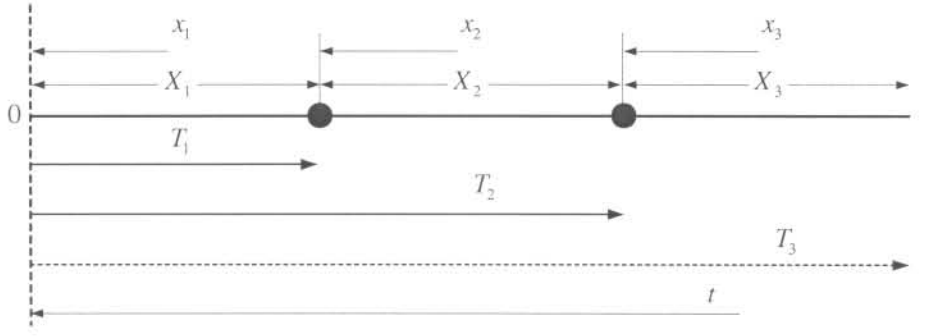

Figure 1.2: Example sample path of a failure process (Dots denote failures)

In Figure 1.2, X_i, $i = 1, 2, 3...$, refers to the *interarrival* time between the $(i-1)^{\text{th}}$ failure and i^{th} failure. X_i is a random variable (RV) with $X_0 \equiv 0$. This is referred to as *local* time and is

[¶]An item has failed when it no longer performs according to certain preset standards. This does not necessarily imply complete destruction.

convenient to use when analyzing non-repairable systems. The real variable x_i measures the time elapsed since the most recent failure. T_i, $i = 1, 2, 3, ...$, measures time from 0 to the i^{th} failure time. T_i is also called the *arrival* time to the i^{th} failure and is mostly used to analyze repairable systems. This time scale is referred to as *global* time.

Clearly, $T_k \equiv X_1 + X_2 + ... + X_k$. From this, a RV $N(t)$ can be defined as the maximum value of k for which $T_k \leq t$, i.e. $N(t)$ is the number of failures that occur during $(0, t]$. $N(t), t \geq 0$ is the integer valued *counting process* that includes information on both the number of failures in $(0, t]$, $N(t)$, and the instants of occurrence, $T_1, T_2, ...$

Another important concept used in survival data modeling is that of the *backward recurrence* time, $B(t)$. It is defined as the time from the arbitrary time t to the immediately preceding failure, i.e. $B(t) \equiv t - T_{N(t)}$. Similarly, is the *forward recurrence* time, $W(t)$, defined as $W(t) \equiv T_{N(t)+1} - t$.

1.2.2 Selecting an appropriate model type

The process of selecting the correct model type for a particular data set is totally ignored in many applications of statistical failure analysis theory. Ascher and Feingold (1984) have constructed an outline of this process based on fundamental statistics. See Figure 1.3.

Some comments will be made on Figure 1.3:

(i) *Chronologically ordered X_i's.* It is extremely important to keep data in chronological order when starting with the process of deciding on the model type. Very often, failure data is reordered by magnitude which makes the process appear to follow, for example, an exponential distribution according to Ascher and Hansen (1998).

(ii) *Trend testing.* A number of techniques exist to recognize trends in data. Graphical techniques include (a) plotting cumulative failure times versus cumulative time on linear paper (Nelson (1982)); (b) estimating the average rate of occurrence of failure (ROCOF, see Section A.3) in successive time periods; and (c) Duane plots as introduced by Duane (1964).

Mathematical tests generally suitable to identify trends in data include De Laplace (1773) (commonly referred to as Laplace's test), Bartholomew (1955), Cox (1955), Bartholomew (1956a), Bartholomew (1956b), Bates (1955), Boswell (1966), Cox and Lewis (1966), Boswell and Brunk (1969), Lorden and Eisenberger (1973) and Saw (1975). More recent examples are Bain, Engelhardt, and Wright (1985), Lawless and Thiagarajah (1996), Martz and Kvam (1996) and Vaurio (1999). Laplace's test is regarded as the most reliable test and is used most often because it produces useful results even for small samples and its result is easily interpreted. Laplace's test is discussed in

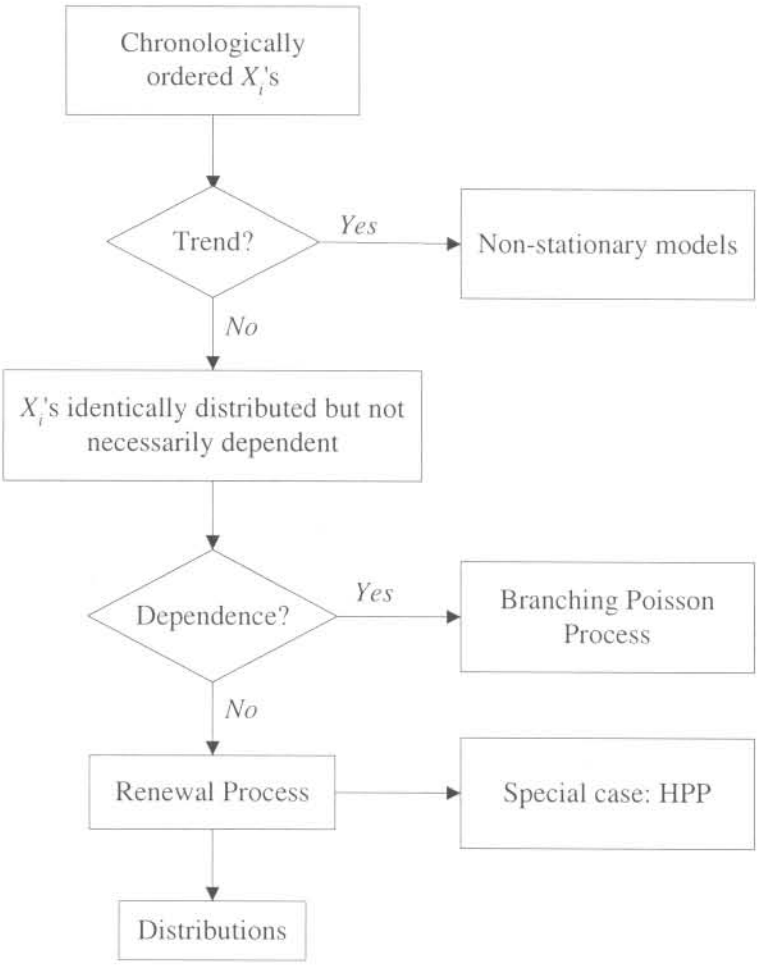

Figure 1.3: Statistical failure analysis of successive interarrival times of a system. (Adapted from Ascher and Feingold (1984)).

Section A.1.

(iii) *Non-stationary models.* Non-stationary models should be used to model data with a definite trend. The Non-homogeneous Poisson Process (NHPP) is used extensively for this purpose. Countless examples of the application of the NHPP are found in the literature, including Kumar and Westberg (1996b), Vineyard, Amoako-Gyampah, and Meredith (1999), Rhodes, Halloran, and Longini (1996), Percy, Kobbacy, and Ascher (1998), Newby (1993) and Lawless (1987). The NHPP is defined and described in Section A.3.3. Although the NHPP is used most often to model repairable systems' failure behaviour, there are also some other fundamentally different non-stationary models suitable for this application. See for example Cozzolino (1968), Singpurwalla (1978) and McWilliams (1979). These approaches were never very popular and are seldom cited in the literature.

Differential equations are also suitable to model non-stationary point processes in special cases. Schafer, Sallee, and Torrez (1975) have summarized a few differential equation models for repairable systems. Another approach occasionally used to model repairable systems' failure behaviour is time series models such as the Auto Regressive (AR) model (see Chatfield (1980)) and the Box-Jenkins Auto Regressive Integrated Moving Average (ARIMA) model (see Wals and Bendell (1987)). The Box-Jenkins model has been used on a few occasions to model software reliability. See for example Burtschy, Albeanu, Boros, Popentiu, and Nicola (1997) and Chatterjee, Misra, and Alam (1997).

(iv) *Testing for dependence.* Although testing for dependence of interarrival times are of extreme importance in reliability modeling, it is almost always ignored. Two reasons for this are (1) the need for large sample sizes; and (2) the complexity of interpreting dependency tests. Cox and Lewis (1966) propose a very natural technique to test for dependency by simply calculating the sample correlation coefficient of lag j, i.e. $\hat{c}_j$. Thus, the correlation between X_i and X_{i+j} is calculated for $i = 1, 2, ..., m - j$ and $1 < i + j \leq m$ where m is the total number of observed events.

(v) *Branching Poisson Process (BPP).* The BPP is described in Section A.3.4. "*The BPP potentially has wide applicability to reliability problems*" according to Ascher and Feingold (1984). However, no practical application of the BPP was found in the literature. This could be because of large data set requirements and that the reliability fraternity still has not accepted and understood a model like the NHPP.

(vi) *Renewal Process.* A renewal process describes an item that, after a failure, is simply replaced by a new item with the same characteristics, so that the life distribution of the item is enough to deduce all the properties of the item. Although it is very important to recognize renewal situations, it is seldom realistic for true life systems. Parts or non-repairable systems do, however, sometimes behave according to renewal processes. Some notes on renewal theory are presented in Section A.2.1.

(vii) *Homogeneous Poisson Process (HPP).* The details regarding the HPP are discussed in Section A.3.2. It is given as a special case of a renewal process in Figure 1.3 because it is numerically equivalent to the FOM of a renewal processes being represented by an exponential distribution. Other than this property, there is no relationship between the HPP and a renewal process.

(viii) *Distributions.* Distributions typically used to model renewal processes are presented as part of the discussion on renewal theory in Section A.2.2.

The outline in Figure 1.3 can be seen as a road map to the correct model-type and should always be used in failure data analysis. Guidelines for the appropriate selection of regression models are presented in by Kumar and Westberg (1996b) and are considered in Chapter 2.

1.2.3 Statistical models in conventional failure time data analysis

In conventional failure time data analysis it is either assumed that an item is totally renewed after maintenance (GAN), i.e. perfect maintenance was done [‖] or that the item is in the same condition after maintenance as it was shortly before failure (BAO), i.e. minimal repair was done. The GAN property is modeled by zeroing an item's Force of Mortality (FOM) after renewal while the BAO assumption is represented by equating an item's *intensity* shortly before and shortly after failure. These concepts are introduced in the sections to follow.

1.2.3.1 Renewal models

Suppose the interarrival times of a system follow a distribution $f_X(x)$ with cumulative distribution $F_X(x)$. $F_X(x)$ is referred to as the *unreliability* function since it gives the probability of failure up to a certain age x, i.e. $F_X(x) = \Pr[X \leq x]$. Similarly, the *reliability* function, $R_X(x)$, is defined as $R_X(x) = \Pr[X \geq x]$ or $R_X(x) = 1 - F_X(x)$, i.e. the probability of survival up to age x. From this it is possible to define the force of mortality (FOM) or hazard rate of an item that gives the probability of failure within a short time, provided that the item survived up to that time, i.e. $h_X(x) = \Pr[x < X \leq x + \mathrm{d}x | X > x]$. The FOM can also be expressed as,

$$h_X(x) = \frac{f_X(x)}{1 - F_X(x)} \tag{1.1}$$

The FOM is further known as the *full intensity* or *conditional intensity* of the failure process of a non-repairable system. These concepts are defined in detail in Section 2.2. The FOM is often erroneously described as a conditional probability density function. The FOM is clearly not a conditional PDF because,

$$R_X(x) = e^{-\int_0^x h_X(\tau)\mathrm{d}\tau} \tag{1.2}$$

and since $R_X(\infty) = 0$ it implies that

$$\lim_{x\to\infty} \int_0^x h_X(\tau)\mathrm{d}\tau = \infty \tag{1.3}$$

For an increasing FOM, an item has an increasing probability to fail as time progresses and use based preventive renewal will be a definite option to consider, although cost will be the decisive factor. Preventive renewal will usually only be used if the total cost of a failure is considerably higher than the total cost of preventive actions. If equation (1.1) yields a constant risk, the component is said to have a random shock failure pattern because the risk of failure of the component remains the same throughout the item's life. Corrective

[‖] This could imply complete replacement.

renewal will be the first option to consider for this case, i.e. a Repair Only On Failure (ROOF) strategy. A ROOF strategy will also most probably be used for a component with a decreasing FOM, since the probability of component failure becomes less as time increases. It should be kept in mind, however, that condition monitoring could be used for any shape of the FOM. The GAN assumption implies that the FOM is zeroed after every failure. Figure 1.4 illustrates this concept.

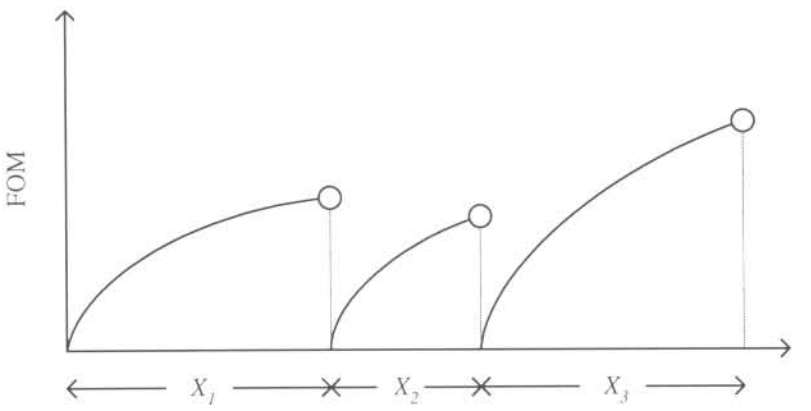

Figure 1.4: Illustration of the GAN assumption

Because of the assumption that interarrival times are part of an underlying distribution, only independent and identically distributed data sets can be used in renewal theory. This requirement is often totally ignored in the literature. In cases where the IID assumption holds, the Weibull distribution is usually most suitable to describe the data set because of its flexibility. Other distributions favored by analysts include the exponential, log-normal, log-logistic and normal distributions. Section A.2.2 gives more information about these distributions.

1.2.3.2 Models for repairable systems

For repairable systems it is assumed that the intensity of the failure process is equal shortly before and shortly after failure. To continue the discussion it is necessary to introduce the concept of intensity (also known as *full intensity* or *conditional intensity*) briefly at this point. This is done in detail in Section 2.2. The intensity of a counting process is generally defined as:

$$\iota(t) \equiv \lim_{\Delta t \to 0} \frac{\Pr\{N(t+\Delta t) - N(t) \geq 1 | H_t\}}{\Delta t} \tag{1.4}$$

where $N(t)$ is the observed number of failures in $(0, t]$ and H_t is the history up to, but not including, time t. Thus, $\iota(t)\Delta t$ is, for a small Δt, the approximate probability of an event in $[t, t + \Delta t)$, given the process history.

In conventional repairable systems modeling it is assumed that processes are orderly, i.e. simultaneous failures cannot occur, and also stationary, which implies $\iota(t) = v(t)$, where $v(t)$ is the so called Rate of OCcurrence of Failure (ROCOF), given by

$$v(t) = \frac{d}{dt} E\{N(t)\} \tag{1.5}$$

The above mentioned simplifications make the NHPP a very suitable candidate for modeling the ROCOF ** of repairable systems. The following forms are encountered most frequently: (1) $\rho_1 = \exp(\Gamma + \Upsilon t)$ (log-linear) and (2) $\rho_2 = \kappa\beta t^{\beta-1}$ (power-law) or even a constant ROCOF. A few authors that used these models are Balakrishnan (1995), Shin, Lim, and Lie (1996), Hokstad (1997), Jensen (1990), Ledoux and Rubino (1977), Kobbacy, Percy, and Fawzi (1994) and Hasser, Dietrich, and Szidarovszky (1995). Figure 1.5 illustrates the BAO assumption for an item.

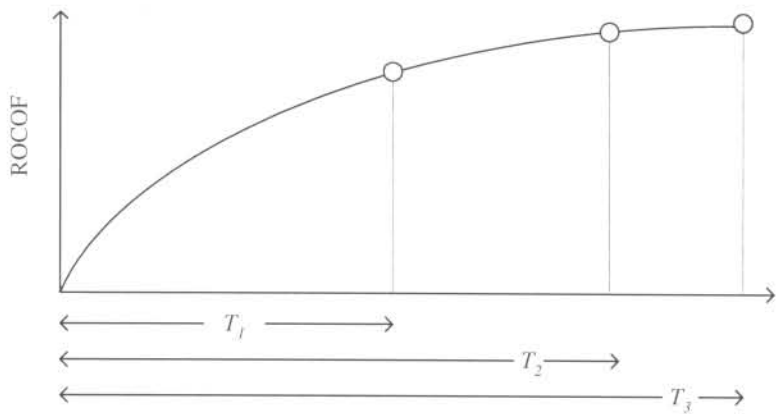

Figure 1.5: Illustration of the BAO assumption

Even though the BAO assumption is much more realistic than the GAN assumption, it could still be a very limited approach according to Ascher and Feingold (1984), since in practice the highest probability of failure is often directly after maintenance.

1.2.4 Conventional replacement/repair cost optimization models

Conventional cost optimization models strives to minimize long term operational cost of equipment by using the statistical models mentioned above. This optimum is often referred to as the minimum Life Cycle Cost (LCC) of an item. The term LCC could be somewhat confusing in this context since it is commonly used in capital replacement studies where the total cost of ownership is taken into account, including operation and maintenance cost, the time value of money, depreciation, etc. To be consistent with the majority of literature in

**The ROCOF of an NHPP is referred to as the *peril* rate and is denoted by $\rho(t)$.

this field, the term LCC will also be used in this document even though only operational costs are considered. In this section, some examples of different approaches are presented to explain the concept.

1.2.4.1 Optimization models for renewal situations

Here, the risk of wasting residual life is balanced with the risk of suffering an expensive unexpected failure in terms of cost. At the point of balance, the LCC per unit time will be a minimum. The costs involved are C_p, the cost of preventive replacement (or renewal) and C_f, the cost of unexpected failure.

The principle of these models is fairly simple to understand. Suppose a component is always replaced at time X_p or at failure time X, whichever comes first. The total cycle cost is then given by $C_p R_X(X_p) + C_f[1 - R_X(X_p)]$. If it is assumed that it takes a time units to perform preventive action and b time units to perform corrective maintenance, the expected duration of the component's life is $(X_p+a)R_X(X_p)+(X+b)[1-R_X(X_p)]$. Division yields the following relation for component cost per unit time (if the replacement rule is followed):

$$C(X_p) = \frac{C_p R_X(X_p) + C_f[1 - R_X(X_p)]}{(X_p + a)R_X(X_p) + (\int_0^{X_p} x \cdot f_X(x)dx + b)[1 - R_X(X_p)]} \tag{1.6}$$

The minimum cost is found where $dC(X_p)/dx = 0$. (See Jardine (1973) for details). For example, suppose a data set is described by a Weibull distribution with $\beta = 2.5$ and $\eta = 200$. Also, assume $C_p =$ R 5 000 (with $a =$ 2h) and $C_f =$ R 20 000 (with $b =$ 8h), then equation (1.6) will yield the graph in Figure 1.6. This graph shows that there is a clear optimum at around 111 days, i.e. R 77 per unit time .

Using the same methodology as above, a relation can be derived to optimize availability instead of cost. It is also possible to calculate the optimum preventive replacement frequency for component-blocks rather than for single components.

Many authors have made some minor refinements to the conventional optimization models for components, often to adapt to data constraints. Overviews of these refined models can be found in Sherif and Smith (1981), Aven and Dekker (1997), Aven and Bergman (1986), Dekker (1995), Zijlstra (1981), Sherwin (1999), Van Noortwijk (2000) and Schäbe (1995). A noteworthy extension of these models, is the model of Ran and Rosenlund (1976) in which the time value of money is taken into account. This model is obviously only useful in cases where equipment is expected to survive for several years.

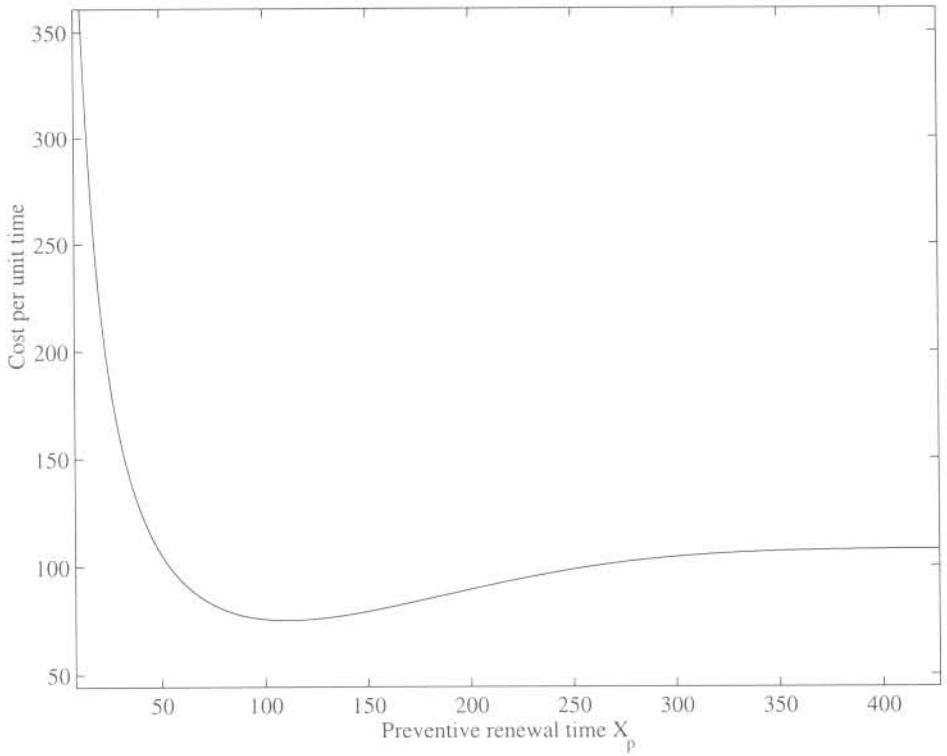

Figure 1.6: LCC of an item renewed after X_p time units or at failure (if $X < X_p$)

1.2.4.2 Optimization models for repairable systems

For repairable systems, a decision has to be made between minimal repair or complete replacement of a system when it has failed. The cost of minimal replacement, C_1, is expected to be considerably less than that of system replacement, C_2. It is hence required to calculate the number of minimal repairs that should be allowed before system replacement with the objective to minimize the LCC. The optimal solution can be expressed in terms of the number of minimal repairs, n, or as time, I. Ascher and Feingold (1984) showed that if the power-law process, $\rho_2 = \kappa\beta t^{\beta-1}$, is used to model the ROCOF of a process modeled by an NHPP, the optimal replacement time will be,

$$I^* = \left[\frac{C_2}{C_1 \cdot (\beta - 1) \cdot \kappa}\right]^{1/\beta} \tag{1.7}$$

and the optimal number of minimal repairs before complete replacement is:

$$n^* = \frac{C_2}{C_1 \cdot (\beta - 1)} \tag{1.8}$$

where I^* and n^* are the optimal solutions. Suppose $\beta = 1.7$ and $\kappa = 0.0015$ with $C_1 = \text{R}500$ and $C_2 = \text{R}8,000$, then $I^* \approx 288$ at R 67 per unit time and $n^* \approx 22$ at R 65 per unit time. Figures 1.7 and 1.8 show these results graphically. The costs per unit time resulting from the two policies above are often very similar except in situations where $C_1 \approx C_2$ (which is rare).

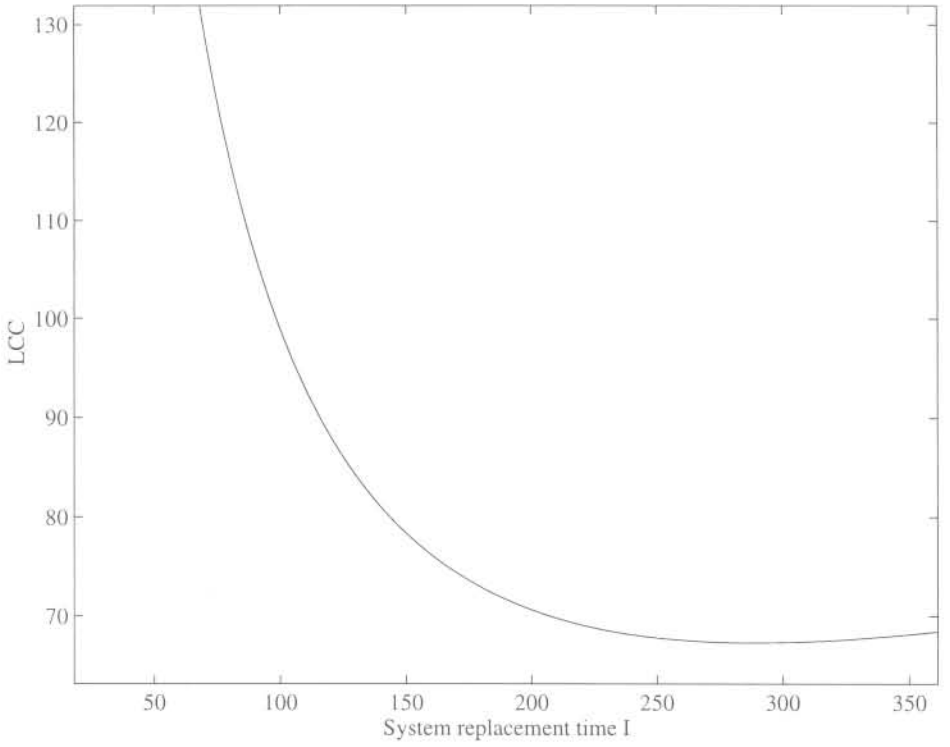

Figure 1.7: LCC of a system minimally repaired up to I* time units

Several minor improvements to minimal repair/replacement policies have been proposed since the introduction of conventional statistical failure analysis. For some early references see Barlow and Hunter (1960), Ross (1969), Morimura (1970) and Park (1979). More recent examples include Stadje and Zuckerman (1991), Yeh (1991), Lam and Yeh (1994), Hsu (1999), Sheu (1999) and Lim and Park (1999).

1.2.5 Shortcomings of conventional approaches

Limitations of conventional approaches do not so much lie in the techniques themselves but rather in the underlying assumptions. The renewal (GAN) assumption is probably the most unrealistic of the two assumptions discussed above. Deterioration of a system may influence the lifetimes of future components in certain sockets severely, even if components are completely renewed/replaced. Renewal theory deals with an important data type however, and certainly has its place in theory even though it is seldom practical. The minimal repair (BAO) assumption is much more realistic than the GAN assumption but still not completely practical. Human interference to improve the condition of a system is often the greatest cause of maintenance - a fact that the BAO assumption does not take into account.

Many authors have proposed models with discontinuities to incorporate the BOWN or WO

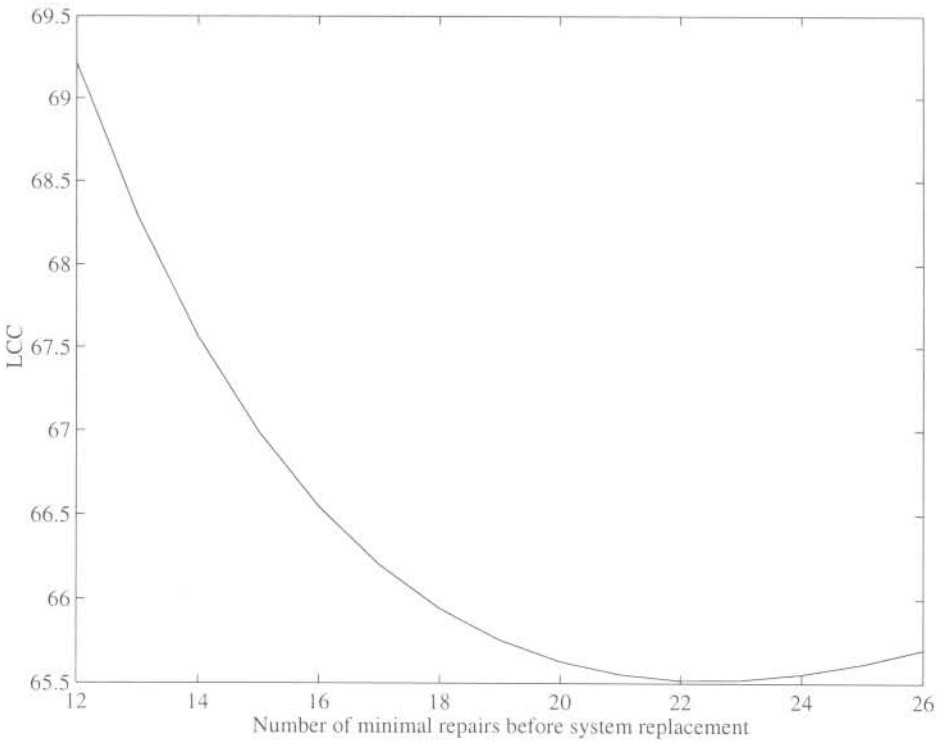

Figure 1.8: LCC of a system minimally repaired up to n failures

situations. These models are major improvements on the conventional approaches although seldom utilized in practice, mainly because of complexity. Models with discontinuities are discussed in Chapter 2.

The biggest shortcoming of models used for conventional failure time data analysis is their inability to include concomitant information in analyses. Diagnostic information recorded during lifetimes, such as Condition Monitoring (CM) results, is not included in models which certainly limits the accuracy of predictions immensely. Regression models solves this problem to a great extent because diagnostic information can be included in the form of covariates. Regression models are described extensively in chapters to follow.

A further serious disadvantage of conventional approaches is the long term nature of replacement/repair policies. Costs only converge to the statistical optimum after a few lifetimes and the minimum LCC approach is often rejected due to the impatience of maintenance practitioners. It is often also very difficult to estimate realistic values for C_p, C_f, C_1 and C_2 for use in the policies outlined in the previous section. The LCC is further poorly understood in industry and the optimum is commonly interpreted as a prediction of time to failure.

1.3 Preventive maintenance optimization through Condition Monitoring

Condition Monitoring has become increasingly popular in recent times. One reason for this is the present affordability of specialized condition monitoring equipment and the perception that advanced technology can solve all maintenance problems. A further reason is that maintenance strategy setting methodologies, such as RCM, recommend an on-condition task as default strategy, provided that the task is technically and economically feasible. (See Nolan and Heap (1978) for details on RCM). These and other factors contribute to a (often erroneous) drive towards condition monitoring in industry.

An item's condition can be assessed much better at present than a few years ago with technology of the day incorporated into techniques such as vibration analysis, oil analysis and thermography. This however does not imply that these techniques are perfect. A general investigation into typical condition monitoring practices revealed several shortcomings, which are discussed below.

1.3.1 Alarm trigger setting

CM techniques assess an item's condition in its present operating state and a maintenance decision has to be made based on the observed diagnostic information. This implies than limits have to be set for measured parameters and once one or more of the limits are exceeded (triggered), preventive action should take place. This may seem simple, but setting appropriate benchmarks is no trivial procedure.

Original Equipment Manufacturers (OEM's) often give guidelines as to what is acceptable operating conditions for equipment in terms of temperature, vibration, oil debris, etc. These guidelines usually form the basis of benchmarks although it is normally very conservative for obvious reasons. Initial benchmarks can then only be optimized through a trial and error approach that may be very expensive.

Many algorithms / techniques have been proposed by researchers in the various CM fields to determine optimal benchmarks in a process to eliminate trial and error approaches. These algorithms have one common underlying principle: to learn from observed diagnostic measurements taken in the past and then estimate optimal benchmarks in a scientific manner for a piece of equipment currently in operation. The most successful of these techniques is neural networks. Neural networks have a large appeal to many researchers due to their great closeness to the structure of the brain, a characteristic not shared by other modeling techniques. In an analogy to the brain, an entity made up of interconnected neurons, neural networks are made up of interconnected processing elements called units, which respond in parallel to

a set of input signals given to each. The unit is the equivalent of its brain counterpart, the neuron.

A neural network consists of four main parts:

(i) Processing units, where each processing unit has a certain activation level at any point in time.
(ii) Weighted interconnections between the various processing units which determine how the activation of one unit leads to input for another unit.
(iii) An activation rule which acts on the set of input signals at a unit to produce a new output signal, or activation.
(iv) Optionally, a learning rule that specifies how to adjust the weights for a given input/output pair.

Time failure data with CM information can be used as processing units to estimate and teach neural networks and additional data can then be used as inputs to predict future outputs. Recent attempts to apply neural networks in the reliability modelling field include Shyur and Luxhoj (1995), Rawicz and Girling (1994) and Lakey (1993). Neural networks have not made much ground in the field of reliability because of its general complexity, large data set requirements and its inability to eliminate insignificant observations.

Setting appropriate alarms for CM parameters is no easy task and CM techniques are seldom optimal from implementation. This is a significant shortcoming in the field of condition monitoring.

1.3.2 Significance of observed parameters

CM techniques use several parameters to assess an item's condition. This may be a frequency spectrum in vibration monitoring, a range of temperatures in thermography, the quantity of various foreign elements in an oil sample, etc. In some instances different CM techniques are combined to estimate equipment reliability. The reason for using more than one parameter is because it is very seldom obvious which parameter is the best indicator of approaching failure and no general technique exists in contemporary CM to isolate significant parameters. The inability of CM techniques to isolate significant parameters is closely related to the alarm trigger limit issues outlined in the previous section.

1.3.3 Lack of commitment towards CM

In general, there is a lack of commitment towards condition monitoring in the South African industry. In many cases, expensive CM equipment is used as the flagship of maintenance departments although inspections are done very irregularly and not recorded properly. Often the information supplied by CM is totally disregarded when a decision has to be made and experience or intuition is relied on. Even if CM information is considered, the final decision is frequently left to the discretion of technicians involved with the equipment.

It does not matter how technologically advanced CM is, if it is not practiced correctly, meaningful results are impossible to obtain. This is a maintenance management issue that is not directly addressed in this thesis.

1.4 Combining use based preventive maintenance optimizing techniques with CM technology

From the discussions above it follows that use based preventive maintenance optimization techniques complement CM technology extremely well. A technique that combines these strategies would have enormous potential. The solution lies in statistical regression models since this type of model allows for concomitant information with time to event data - in this context the concomitant information could be diagnostic information recorded by CM techniques.

Several regression models have been applied in reliability to estimate the risk of failure of an item and most of these models are discussed in the next chapter. Only the Proportional Hazards Model (PHM) is discussed in this section as an introduction to regression models, but also because this is the only regression model for which a scientific preventive maintenance decision model exist.

1.4.1 Proportional Hazards Modeling

The PHM was introduced by Cox (1972) and was considered to be a total revolution in survival analysis. This model was intended for the field of biomedicine but became increasingly popular in reliability modeling over the past two decades. The model uses a baseline hazard rate and allows a functional term containing covariates to act multiplicatively on the baseline hazard rate (or FOM), i.e.

$$h(x, \boldsymbol{z}) = h_0(x) \cdot \lambda(x, \boldsymbol{z}(x)) \tag{1.9}$$

where h_0 is the baseline FOM, λ is the functional term and $\boldsymbol{z}$ is a vector of covariates which may be time-dependent. Kumar and Klefsjo (1993) summarized the assumptions of the PHM as follows:

(i) Event data is IID.

(ii) All influential covariates are included in the model.

(iii) The ratio of any two FOMs as determined by any two sets of time-independent covariates $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ associated with a particular item has to be constant with respect to time, i.e. $h(x, \boldsymbol{z}_1) \propto h(x, \boldsymbol{z}_2)$. For time-dependent covariates, this assumption is not defined.

The exponential function is used most often for the functional term. This leads to a semi-parametric model. It is possible to calculate the semi-parametric model without making any assumption on the baseline hazard rate but this only yield relative risks. In reliability, the absolute risk is usually required and the model is hence parameterized by specifying some parametric FOM for the baseline, for example the Weibull FOM, i.e.

$$h(x, \boldsymbol{z}) = \frac{\beta}{\eta} \cdot \left(\frac{x}{\eta}\right)^{\beta-1} \cdot \exp\left(\boldsymbol{\gamma} \cdot \boldsymbol{z}(x)\right) \tag{1.10}$$

where β and η are the Weibull shape and scale parameters respectively and $\boldsymbol{\gamma}$ is a vector of regression coefficients. The influence of the functional term results in an improved estimate of an item's FOM. Figure 1.9 illustrates this concept.

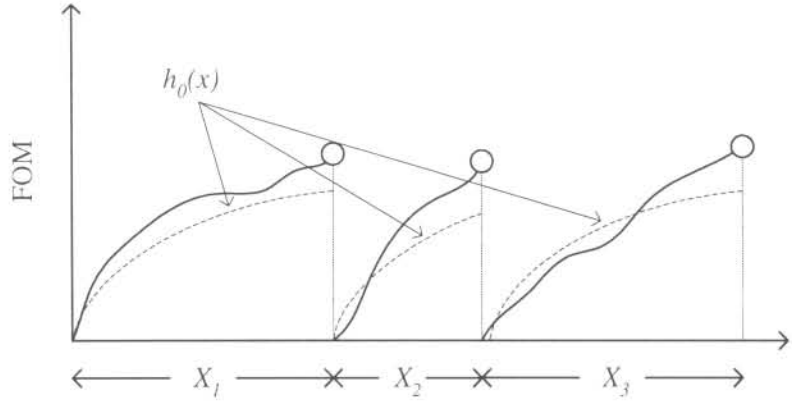

Figure 1.9: Illustration of the PHM with time-dependent covariates

The PHM has been applied successfully in diverse reliability applications because of the improved estimate of the FOM, including modeling component failures in a light water reactor plant by Booker et al. (1981), marine gas turbine and ship sonar by Ascher (1983), motorrettes by Dale (1985), aircraft engines by Jardine and Anderson (1988), high speed train

brake discs by Bendell et al. (1986), sodium sulfur cells by Ansell and Ansell (1987), surface controlled subsurface safety valves by Lindqvist et al. (1988) and machine tools by Mazzuchi and Soyer (1989). Other authors that published applications of the PHM in reliability include Jardine et al. (1989), Leitao and Newton (1989), Love and Guo (1991a) and Love and Guo (1991b).

The biggest criticism of the PHM is the fact that it is by definition only applicable for IID data. This shortcoming can be addressed by allowing for imperfect repair in the covariates, but this does not solve the problem completely and in some cases can even worsen the situation. Some authors, for example Kumar (1996), have applied the PHM with reasonable success on repairable systems, despite the requirement of IID data.

1.4.2 Decision making with the PHM

Estimating the optimum maintenance instant that will result in the minimum LCC of an item, based on the FOM as determined by the PHM, is no trivial procedure since the FOM is now dependent on time and the values of covariates. This implies that the optimum LCC instant must be specified in terms of risk and not in terms of a process parameter, such as time, as was described in Section 1.2.4.1.

Two attempts to calculate the optimal maintenance instant for a system with the PHM were found in the literature. The first was by Kumar and Westberg (1996a) that used the PHM together with Total Time on Test (TTT) plotting to estimate the optimum maintenance frequency. This paper was not very case-orientated and is, as far is known, the only of its kind. The second consists of a series of publications by, amongst others, Makis and Jardine. These authors have developed a technique for calculating the minimum LCC in terms of a system's risk as determined by the PHM. Makis and Jardine (1991) and Makis and Jardine (1992) proposed a semi-Markov approach to calculate the minimum LCC where covariate behavior is predicted by semi-Markov chains. Makis and Jardine's technique was then refined in several publications to follow, the most important being Banjevic, Ennis, Braticevic, Makis, and Jardine (1997) and Jardine, Banjevic, and Makis (1997).

Makis and Jardine's optimization technique produces a result that looks very similar to Figure 1.6, except that the cost is expressed as a function of $h(x, \boldsymbol{z})$. It is then required to allow an item to operate until the optimum risk level (as opposed to time) is reached before preventive action is taken. Figure 1.10 illustrates the policy in two dimensions with imaginary inspection data. The figure shows how the optimal risk is influenced by both time and the observed level of covariates.

Examples of successful applications of this replacement policy include Vlok (1999) and Jardine, Banjevic, and Makis (1997).

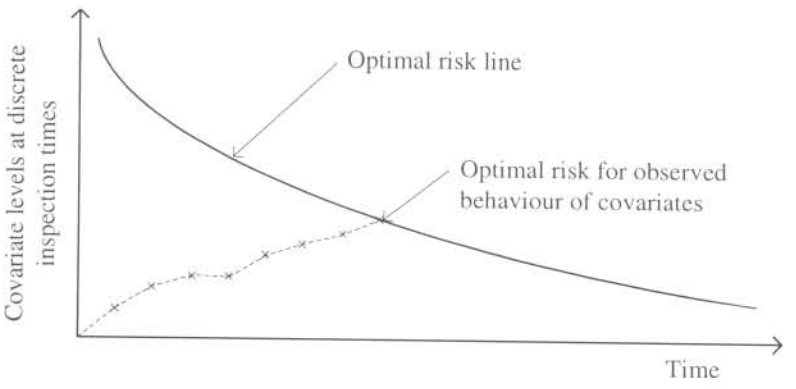

Figure 1.10: Illustration of the optimal policy with imaginary covariate levels

1.4.3 Shortcomings of PHM cost optimization

Even though the PHM cost optimization approach is an improvement on conventional techniques, there are still two major shortcomings:

(i) The PHM has an underlying assumption that data is IID. This limiting assumption can be overcome to a certain extend by including covariates that describe an item's failure history. The fact remains, however, that the model is not entirely suitable for repairable systems data and repairable systems data is expected much more often than IID data.

(ii) The minimum LCC cost approach for the PHM is a long term type approach, as is the case with conventional analysis techniques. This approach is also not accepted well amongst maintenance practitioners because the minimum is only reached after a few lifetimes of which some could be expensive unexpected failures.

Both the above-mentioned shortcomings should be addressed in order to make a truly valuable contribution to the field of statistical failure analysis.

1.5 Residual life

Maintenance would be a trivial affair if the exact times to failure of items were known. This would imply that no residual life is wasted and that expensive unexpected failures are totally eliminated. Although this view seems only feasible in a perfect world, the approach is certainly meritorious in a world striving for perfection. A scientific technique with the ability to estimate the residual life of equipment will be of great advantage to the field of maintenance engineering.

In CM, some empiric methods exist to estimate residual life. These methods are seldom generalized and are often only meaningful after many iterations. The methods also differ from situation to situation, even for nominally similar items, which makes it very risky to use a particular method to predict residual life.

For conventional renewal analysis, residual life estimation (in principle) only goes as far as conditional expectation or mean. If a distribution, $f_X(x)$, describes an item's failure history, the residual life, $\mu(x)$, can be calculated by

$$\mu(x) \equiv E[X - x | X \geq x] = \frac{\int_x^\infty (\tau - x) f_X(\tau) d\tau}{R_X(x)} = \frac{\int_x^\infty R_X(\tau) d\tau}{R_X(x)} \tag{1.11}$$

The simple statistical mean life, μ, of the item is given by

$$\mu = \int_0^\infty x \cdot f_X(x) dx \tag{1.12}$$

It is important to note that any differentiable $\mu(x)$ has to satisfy $\mu(x) \geq -1$ because of the identity

$$h_X(x) = \frac{\frac{d\mu(x)}{dx} + 1}{\mu(x)} \tag{1.13}$$

as described by Muth (1977). Ghai and Mi (1999) discussed mean residual life and its association with the FOM in detail. Other authors that worked on this subject include Tang, Lu, and Chew (1999), Baganha, Geraldo, and Pyke (1999) and Guess and Prochan (1988).

Equally little work has been done on estimating the residual life of items with conventional repairable systems theory. Calabria, Guida, and Pulcini (1990) proposed a point estimation procedure for future failure times of a repairable system modeled by a NHPP with a power intensity law. Suppose a repairable system has suffered n failures and it is required to estimate the $(n+m)^{\text{th}}$ failure, where $m \geq 1$ and $m \in \mathbb{Z}$. The Maximum Likelihood Estimate (MLE) of the expected value of the m^{th} future failure is given by

$$E_m = (n-1) \cdot \gamma \int_{t_n}^\infty \sum_{j=1}^m C_j \frac{n+j-1}{n} \left[1 + \frac{n+j-1}{n} \gamma \ln(t_{n+m}/t_n)\right]^{-n} dt_{n+m} \tag{1.14}$$

where $C_j = \prod_{i \neq j}^m (n+i-1)/(i-j)$. Schäbe (1995) followed a similar approach, as did Reinertsen (1996).

The theory above shows that conventional statistical failure analysis only yields a *mean* residual life estimate. This fact makes the use of residual life estimates very unpopular and unreliable in practice. A dynamic residual life estimate is required to be useful in practice, i.e. a technique that will adjust estimates based on certain observed influences. Statistical models that have the ability to incorporate concomitant information immediately seem to be a possible solution even though very few publications on this subject exist. Zahedi (1991)

proposed a proportional mean remaining life model analogous to the PHM where a baseline survivor function is influenced by a functional term containing covariates. No publication was found where this model was applied on real life survival data, however. Other contributions to multivariate residual life estimation include Nair and Nair (1989), Arnold and Zahedi (1988) and Zahedi (1985). In neither of these publications, practical illustrations of the theory were presented.

1.6 Problem statement

There is a need to optimize preventive maintenance decisions in today's ever increasingly competitive market. At present there are three established means for doing this namely, conventional statistical failure analysis, condition monitoring and Proportional Hazards Modeling. The following shortcomings were identified for the respective techniques:

(A) *Conventional failure analysis*

A-1. Only allows for the GAN or BAO assumption, which is extremely limiting.
A-2. Lack of ability to include concomitant information in the analyses.
A-3. Requires fixed estimates for C_p and C_f, which often varies for every failure.
A-4. The long term nature of optimal replacement/repair policies is often rejected by maintenance practitioners because unexpected failures are regarded as unacceptable.

(B) *Condition Monitoring*

B-1. It is very difficult to set optimal initial alarm trigger settings for CM techniques.
B-2. No scientific technique exists with which the significance of CM parameters can be calculated.
B-3. There is a general lack of managerial commitment to CM.

(C) *Proportional Hazards Modeling*

C-1. Assumes data to be IID.
C-2. The only replacement decision model found for the PHM is also based on costs and requires a few lifetimes before it converges to the minimum cost.

This thesis aims at improving all nine shortcomings listed above. It is proposed that this objective can be reached as follows:

(i) *Development of a combined Proportional Intensity Model (PIM) with the ability to address all the model-related shortcomings mentioned above*

It is proposed that a combined PIM, one for non-repairable and one for repairable systems, is developed that will include the majority of conventional PIM enhancements as special cases (including the PHM) to be able to model most of the typical wear-out/deterioration patterns found amongst industrial equipment. Such a PIM would be able to accommodate discontinuities in the failure intensity and to adapt to discontinuities or to scalings in its time scale which will be ideal for the WO and BOWN scenarios. By developing the combined PIM, shortcomings A-1, A-2, B-2 and C-1 will be addressed.

(ii) *Development of an algorithm to calculate residual life of an item based on the combined PIM*

A flexible and adaptive combined PIM will theoretically lead to a close representation of reality and hence realistic estimates of the residual life, provided that the future behavior of covariates can be estimated with relative high certainty. This could be a challenging task since very little work has been done in this field and the numerical implementation of the theory is fairly complicated. Successful completion of this goal would solve shortcomings A-3, A-4, B-1, B-2 and C-2.

(iii) *Comprehensible presentation of results*

To make a truly practical contribution to the field of reliability modeling, results produced by this study should be presented in a user-friendly and comprehensible manner. This step is required to address shortcoming B-3.

1.7 Thesis outline

In Chapter 2, a literature survey of advanced failure intensity models (including PIMs) in survival analysis is done. Terminology used in failure intensity models is defined and different models are categorized and evaluated. Chapter 2 serves as the foundation for the development of the combined PIMs in Chapter 3. In Chapter 3 the combined PIMs are derived and it is illustrated how these models can be reduced to most conventional PIMs. Parameter estimation techniques based on maximum likelihood are also discussed in Chapter 3. In Chapter 4 conventional techniques are applied to the combined PIMs to estimate residual life. Confidence bounds on estimates are also discussed. Chapter 5 contains a case study in which the theory developed in this thesis is applied to a typical data set from a South African industry. Results are compared to results obtained from a maintenance decision support tool similar to the residual life approach. In Chapter 6 the findings of this thesis are summarized with some recommendations for future research.