INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

RAMS analysis of railway track infrastructure

(Reliability, Availability, Maintainability, Safety)

Gonçalo Medeiros Pais Simões

Dissertation for the qualification in the Msc. degree in

Mestrado Integrado em Engenharia Civil

Jury

Presidente: Prof. Joaquim Jorge da Costa Paulino Pereira

Orientador: Prof. Paulo Manuel da Fonseca Teixeira

Vogais: Prof. José Manuel Care Baptista Viegas

September 2008

ABSTRACT

Railway infrastructures are assets which represent a high investment. They are designed to work in very demanding safety conditions and must display a very low occurrence of failures. Rail tracks are components which are subject to the development of rail defects such as head checks, corrugations, cracking, etc. These failures depend of the characteristics of infrastructure, train traffic and maintenance plans used. Maintenance plans must be devised and forecasting of failures is needed order to keep them within an acceptable level. Maintenance actions associated to the repair of these failures such as grinding, reprofiling or replacement of rails represent costs which are of difficult forecasting at the time the rail infrastructure is built. Methods for forecasting of failure are therefore needed.

Reliability, Availability, Maintainability and Safety (RAMS) techniques have been extensively applied to the electrotechnical engineering field. They have a high potential for application in the rail infrastructure management field but they currently lack standardization and stated procedures for the correct study of RAMS parameters. RAMS techniques allow reliability engineers to forecast failures from the observation of operational field data. RAMS parameters which may be forecast are reliability, availability, failure rate, Mean Time Between Failures (MTBF), Mean Down Time (MDT), among others. This study considers the possible applications of RAMS techniques to the case of railway track maintenance management, focusing on the rail component. Data collection and forecasting procedures are discussed and some considerations are presented on how to use failure forecasts to obtain life cycle costs.

Keywords: high speed rail infrastructure, RAMS techniques, life-cycle costs, rail defects

RESUMO

As infra-estruturas ferroviárias são bens que representam um elevado investimento. São desenhadas para trabalhar em condições de segurança muito exigentes e devem apresentar uma baixa ocorrência de falhas. Os carris são componentes que estão sujeitos ao aparecimento de rail defects como head checks, corrugations, cracking, etc. Estas falhas dependem das características da infra-estrutura, tráfego de comboios e dos planos de manutenção utilizados. Planos de manutenção devem ser traçados e as falhas devem ser previstas para serem mantidas a um nível aceitável. Acções de manutenção ligadas a estas falhas tais como esmerilamento, re-perfilamento ou substituição de carris representam custos que são de difícil previsão na altura da construção da infra-estrutura. São necessários métodos para a previsão de falhas.

As técnicas RAMS têm sido extensivamente aplicadas ao campo da engenharia electrotécnica. Têm um alto potencial de aplicação no campo da gestão de infra-estruturas ferroviárias, mas correntemente há uma falta de estandardização e de procedimentos descritos para o correcto estudo dos parâmetros RAMS. As técnicas RAMS permitem que os reliability engineers prevejam falhas a partir da observação de dados operacionais. Os parâmetros passíveis de previsão são: reliability, availability, failure rate, Mean Time Between Failures (MTBF), Mean Down Time (MDT), entre outros. Este estudo considera as possíveis aplicações de técnicas RAMS no caso de gestão de manutenção da infraestrutura ferroviária, com enfoque no carril. São discutidas técnicas para colheita de dados e elaboração de previsões, e são apresentadas algumas considerações sobre como utilizar estas previsões para a obtenção de custos de ciclo de vida.

Palavras-chave: infra-estrutura ferroviária de alta velocidade, técnicas RAMS, custos de ciclo de vida, *rail defects*

INDEX

List of Figures

List of Tables

List of acronyms and abbreviations

BWR	– Boiling Water Reactor
BS	– British Standard
CBA	– Cost-Benefit Analysis
CCDM	– Cause-Consequence Diagram Method
Cdf	– Cumulative distribution function
CQTM	– Consequence Tree Method
CNET	– Centre National d’Études des Télécommunications
CTM	– Cause Tree Method
CUSUM	– CUmulative SUM
CWR	– Continuous welded rail
EC	– European Commission
EPRI	– Electric Power Research Institute
ERDS	– European Reliability Data System
HBW	– Hardness Brinell Tungsten Carbide Ball Indenter
ISO	– European Norm International Organization for Standardization
FARADIP	– FAilure RAte Data In Perspective
FMEA	– Failure Modes and Effects Analysis
FMECA	– Failure Modes, Effects and Criticality Analysis

GADS	– General Availability Data System
GFCM	– Gathered Fault Combination Method
GIS	– Geographic Information System
IEC	– International Electrotechnical Commission
IEEE	– Institute of Electrical and Electronics Engineers
INPO	– Institute of Nuclear Power Operations
MDT	– Mean Down Time
MGT	– Million Grosse Tonnes
MTBF	– Mean Time Between Failures
MTTF	– Mean Time to Failure
MTTR	– Mean Time to Repair
MUT	– Mean Up Time
NCSR	– National Centre of Systems Reliability
NERC	– North american Electric Reliability Corporation
NPRDS	– Nuclear Plant Reliability Data System
NUCLARR	– Nuclear Computerized Library for Assessing Reactor Reliability
NUREG	– NUclear REGulatory commission
OREDA	– Offshore REliability DAta
pdf	– probability density function
PFD	– Probability of Failure-on-Demand
PHA	– Preliminary Hazard Analysis
PWR	– Pressurized Water Reactor
RADC NPRD	– Regional Air Defence Command Nonelectrical Parts Reliability Data
RAMS	– Reliability, Availability, Maintainability and Safety
RCF	– Rolling contact fatigue
SDM	– Success Diagram Method
SINTEF	– Stifelsen for Industriell og TEknisk Forskning
STRAIT	– STraightening of Rail welds by Automated ITeration
SYREL	– SYstem RELiability service data bank
TTM	– Truth Table Method
UIC	– Union Internationale des Chemins de fer
UK MOD	– United Kingdom Ministry Of Defence

Chapter 1: Introduction

This study will deal with the application of Reliability, Availability, Maintainability and Safety (RAMS) techniques to the railway infrastructure. More pointedly, it will be geared at the rail component, used in tracks which belong to high speed traffic networks, from a situation where the data available on failure rate is scarce and has been collected in short time series.

In Chapter 2, relevant RAMS techniques will be presented. An historical background of the application of RAMS is presented, as well as the current challenges within the reliability field. RAMS cycle is described, from component level to system level, and the transition to the life-cycle cost (LCC) modelling is shown. This is of special importance in the forecasting of costs. RAMS system level qualitative assessment methods (Preliminary Hazard Analysis and Failure Modes and Effects Analysis) are described. These are helpful for the collection of failure and maintenance related data. RAMS parameters such as reliability, availability, failure rate, Mean Time To Failure (MTTF), Mean Time Between Failures (MTBF), Mean Down Time (MDT) and Mean Time To Repair (MTTR) are presented, as well as the mathematical relationships between them. A special section is dedicated to the study of failure rate in all possible combinations of high/low number of failure and constant/non-constant failure rate. Issues regarding reliability data sources will be discussed and guidelines for future data collection are displayed. Finally, a section dedicated to reliability growth explains how to include this phenomenon in the forecasts made from current data sets.

In Chapter 3, Railway high speed track infrastructure will be presented. All components are listed from subgrade to superstructure. The following sections are dedicated to the rail, where all parameters and variables which may have an impact on rail defects are described. A listing of significant rail defects is presented, and maintenance actions associated to them are also displayed.

Chapter 4 is dedicated to the application of the RAMS techniques to the rail infrastructure. This is divided into two major sections, for it is concluded that the main reliability parameters may be obtained from failure rate and availability. These are also parameters with an easy translation into costs. The first major section (Chapter 4.2) is dedicated to the forecasting of failure rate (and reliability) through the Weibull distribution. A description of the failure data procedure and possible pitfalls is made, the statistical processing of field data and the translation to costs is suggested. The processes are illustrated through examples which use sets of fictitious data. The second major section (Chapter 4.3) displays a suggestion for the forecasting of availability, based on a break down process of MDT into several components. The scope for future work is presented in Chapter 5, following the sections in which Chapter 4 is divided.

This study shall **not**: define specific risk levels for rail tracks; define associated RAMS targets to fulfil regarding potential risks and hazards that might occur in rail tracks; present methods of RAMS simulations that allow for the verification of RAMS targets for the considered case study (rail tracks).

Chapter 2: RAMS techniques and their application

2.1 Brief introduction to RAMS – what is RAMS?

Since the beginning of the industrial age, engineers have strived to create reliable and durable equipments and systems. In that time, developments made in the design process occurred mainly due to a trial and error process. Despite all the improvements made since then, it is still impossible to expect any human activity to be flawless and thus failure and risk free. As the cost and risk associated to failures grew larger due to the increase in the complexity of equipments or systems and the shortening of their delivery deadlines, it became more and more important to assess failure and risk and try to make predictions on these as early as in the design step. A need arose to systematically study how solid would the behaviour of a new system, equipment or design improvement be, and that was the beginning of what is today called RAMS.

RAMS is usually looked at as a set of activities that encompass different fields, which are ultimately linked to the study of the failure, maintenance and availability of systems. Its main use is to predict, at any life cycle step of a system[1] what its expected failure rate or other RAMS criteria are, such as MTBF, MDT, Availability, Probability of failure on demand, etc. (Annex A for technical definitions). These parameters are then used for the prediction of life-cycle costs for there are costs associated with operation stoppage and maintenance actions.

As will be stated in the course of this study, the calculations involved in RAMS processes yield values which are not precisely what to expect on field. Nevertheless, RAMS will help in many aspects, for instance:

- It will provide indicators of how sturdy and reliable a system design can potentially be.
- It helps to identify which parts of a system are likely to have the major impacts on system level failure, and also which failure modes to expect and which risks they pose to the users, clients, or society.
- In the planning of cost-effective maintenance and replacement operations.
- The avoiding of hazards/accidents. Risk assessment helps to improve safety levels. RAMS has increasingly been called on use in the assessment of safety integrity levels.
- Assessment of how good a design enhancement, like implementation of a new part or redundancy is.

The resulting data from these assessments will enable the carrying of a life cycle cost-benefit analysis. The latter can be done in a strict economic perspective: if a system fails, it will be prevented to do its

[1] Hereby the word *system* shall be used, though it can also denote a product or equipment.

function correctly. This is translated into costs, due to production stopping but also to repair operations. Accidents, injuries and deaths – though this may be controversial and hard to account – and other aspects such as environmental damage, company image and costumer perception can also be economically translatable. The quantification of these aspects becomes harder and more controversial due to the increase of social awareness on human and environmental safety, and also on environmental sustainability aspects. The reader is referred to Smith (2005) for a deeper insight on these issues. Though it lies out of the scope of this study, it is important to say that studies on the value of environmental aspects and human life must be done, and a consensus about those values must be arrived at between the different parts which share responsibilities in a system (owners, operators, maintenance managers, maintenance crews, etc.) and the authorities regulating their functioning (safety authorities, governments, etc.).

2.1.1 Historical evolution of RAMS

As stated above, the principle of trial and error has, since the beginning, been a key aspect of design improvements even before conditions were set for formal collection of failure data. This process has evolved: by analysing failure modes and their causes, design improvements can be done in order to avoid these failures. This is termed **reliability growth** (Smith, 2005).

Formal collection of failure data started to gain importance with the implementation of mass production chains, and an increasing demand in reliability aspects combined with the pressure imposed by increasingly shorter delivery (as well as design) deadlines for systems. Mass production meant the components could be standardized, so the notion of extensive testing of components under the same conditions finally made sense. Before this they were produced in a craft-like approach, which wouldn't allow for statistical treatment, for they were inherently different.

Several collections of failure data appeared in the UK and US during the 1960s stimulated by poor field reliability experienced by military equipment during the Second World War (Smith, 2005; Villemeur, 1992-A). The notion that it would be better to make investments in preventing systems to fail instead of waiting for failures to happen and repair them was gaining importance (Villemeur, 1992-A).

Later on, from the 1970s, the notion that the trial and error principle had to be replaced started to gain strength as the consequences of accidents involving new technologies became too big, and started to pose risks to the society (i.e. nuclear or chemical industries). So besides the economical consequences of not having systems functioning properly (availability), lately reliability tools started to be heavily applied in the field of risk and hazard assessment (safety) (Smith, 2005).

2.1.2 New challenges in system modelling

As in any other engineering branch, the results of the predictions of a RAMS analysis are directly connected to the way the system is modelled and the input data supplied. This means the modelling should be done having in mind what the goals are and the gathering or collection of information must be focused barely on the relevant data required for those goals (Villemeur, 1992-A).

In any case, these models will take their input from failure data collections. As a consequence, nowadays the accuracy of prediction models is contested, due to the fact that the failure data itself are extremely variable depending on the collections (Smith, 2005; Nunns, 2002). This means that for recently implemented systems, where failure related data is of contested accuracy, the simplest models of simulation will suffice, since it isn't guaranteed that a more complex model will lead to a better prediction due to the uncertainty of the data used. Very complex models can become overwhelming in terms of work, impracticable in terms of data and calculations, and a waste of time and money (Smith, 2005; Villemeur, 1992-A). There might be a larger gain in cost connected to the use of more complex models than the gain in savings which result from their simulation results.

Even more disputable is the premise that the predictions will be accurate based solely on component failure data. System failures are more than the combination of the failures of the components which are part of the systems, other types of failure include (Smith, 2005; Villemeur, 1992-B):

→ Failures caused by software elements (systematic failures).
→ Human factors, e.g. misinterpreting of operating documentations.
→ Failures caused by environmental (common) factors.
→ Common cause failures where redundancy of components is defeated due to external factors that are the same for the replicated units.
→ Over-stressing of components during field operation or drift of parametric specifications of the same, i.e., lower component resistance than expected.

The current reliability data collections do not account for the factors mentioned just above, and are kept at the level of random component failure. Recent techniques are being developed to include in the simulation models common mode failures, human and environmental factors. This is normally achieved by considering these failures as a percentage of the total failures. These can be included in the models for instance using a new series block before the final output in the Success Diagram Method (Annex D.1). For a deeper insight on these subjects the reader is referred to Smith, 2005 and Villemeur, 1992-B. Nevertheless, the key aspect here is the uncertainty of the analyses. There is no proof that models including those aspects will produce more accurate results (Smith, 2005).

These difficulties call for the use of high safety factors which result in the **predicted reliability** or **design reliability** being much smaller than the actual achieved reliability or **field reliability**, normally in one or more orders of magnitude less (Smith, 2005).

As a conclusion, it can be said that RAMS simulation models are to be used with caution. Their usefulness lies mainly in the ability to compare the results of simulated reliability parameters for different component arrangements and maintenance policies, to understand which will lead to a more efficient functioning. The absolute values of reliability parameters returned by RAMS simulations are contested not only because of the simulation methods themselves but mainly due to the variability of component reliability data used as input (Smith, 2005). Therefore, these absolute values should provide only indicative values and lack a confirmation with system testing and on-field operation.

2.2 The RAMS cycles: data collection, simulation, and impact on system life cycle

The goal of RAMS is to create input data for the assessment of the suitability of a system in a life cycle. That is, to provide data on failure rates of the system, possible failure modes, MDT, maintenance operations, hazards and their consequences, etc. This output of RAMS simulations enables the life cycle specialists to calculate costs and to perform Cost-Benefit Analyses (CBA). But before the RAMS simulation at system level, input on known component failure data must be provided. There are then 3 different steps in the application of RAMS, all of which are interlinked. This is illustrated in Figure 2.1.

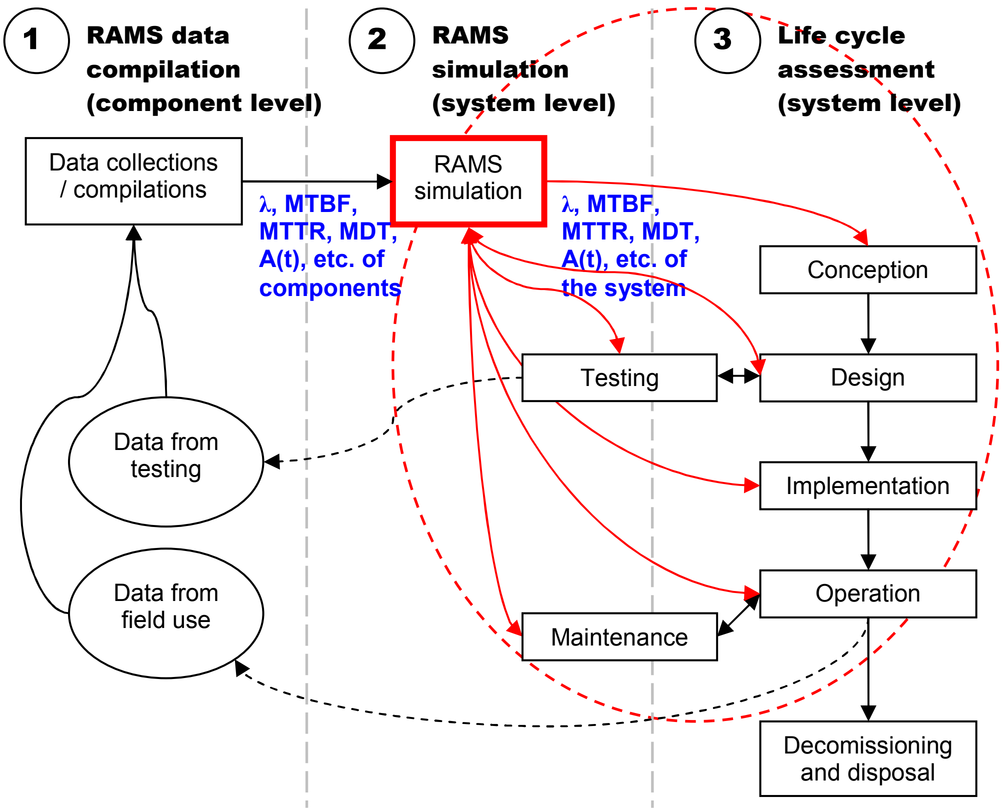


Figure 2.1 Representation of the steps which comprehend RAMS and life cycle. (Adapted from Smith, 2005, Villemeur, 1992-A)

Each one of the steps shown in Figure 2.1 shall comprise different methods. Brief references on steps 2 and 3 shall be made, and a special focus shall be given to step 1. This is the step which is directly connected with the main goal of this study – to be able, from short failure data time series on rail tracks for high speed networks, to create reliable input for steps 2 and 3 (failure rates, MTBF, MDT, Availability, etc. leading to costs). Steps 2 and 3 are not separable; one will influence the other. In the figure, the red ellipse stresses this connection, which is of iterative nature. RAMS targets keep being revised according to life cycle predictions, which depend on the targets. This shall be discussed in chapter 2.2.3.

2.2.1 Step 1: RAMS data compilation (component level)

Failure data compilation stands as the foundation of any RAMS simulation or process. Whether it is made through testing or through observed field operation and maintenance feedback, the study of individual component failure provides data on failure rate and all other reliability parameters. This data is used as input for step 2, RAMS simulation.

This is of utmost importance, for as stated above, in the first paragraph of chapter 2.1.2, a great number of authors state that the use of more complex reliability models (in step 2) does not necessarily mean there is a better result, for there is generally a lack of relevant reliability data (step 1). Reliability data has more impact on the RAMS simulation results than the RAMS simulation models themselves.

The issues discussed from chapter 2.3 through 2.8 are related to this step.

2.2.2 Step 2: RAMS simulation (system level)

In this step, the goal is to model the system in terms of reliability aspects. That is, from values of failure rate, MTBF, MDT, among others from individual components, be able to compute values which refer to the same components but while interacting with each other in a system and a stated environment. The first methods developed to do so – like the Preliminary Hazard Analysis and the Failure Modes and Effects Analysis – are of qualitative nature and are mainly intended to identify failures at system level. Other methods which appeared later on are of quantitative nature and enabled the reliability engineers to predict through mathematical formulations several values for failure rate and related parameters in systems. These simulation result values are termed **design** or **expected reliability**.

The main simulation methods for RAMS are listed below. They are listed in chronological order of appearance:

Qualitative approach:

- Preliminary hazard analysis (PHA), 1960s
- Failure modes and effects analysis (FMEA), 1940s

Quantitative approach[2]:

- Success diagram method (SDM)
- Cause tree method (CTM)
- Truth table method (TTM)
- Gathered fault combination method (GFCM)
- Consequence tree method (CQTM)
- Cause-consequence diagram method (CCDM), 1971

Examples of RAMS software analysis packages used nowadays are RAM Commander, which is used for project (system) level analysis and Weibull++, which is used for the statistical studies of components with non-constant failure rates (chapter 2.5.3).

2.2.2.1 Preliminary hazard analysis (PHA)

This method has been used since the 1960s, being originally developed in the US. Its goals are to identify potential hazards in any system (in its first application cases, in industrial installations), as well as identifying their causes, and finally to determine the severity level of the potential hazards.

After this is done, engineering judgement is used to develop measures that may control, avoid, or ultimately eliminate these hazards. These measures are of qualitative nature, for the method does not allow for a quantitative measuring of the hazards identified.

The presentation of results consists in a table which summarizes all system components, possible failures, their likely causes, as well as the hazards that are associated with the failures.
The method *per se* is only useful in terms of safety, and not so much in reliability assessment in its strict sense. As its name indicates, this is a method which should be used in a preliminary step of other reliability assessment studies, being useful to understand and list which are the failure modes and the undesirable events leading to those failures.

[2] Only qualitative approach methods will be presented in the main text of this study, due to their importance in helping to understand failure modes of the rail track. Here a brief introduction is made and further developments shall be made during the application of these methods in chapter 4. Quantitative approach methods are presented in Annex D.

2.2.2.2 Failure modes and effects analysis (FMEA)

This method was also developed in the 1960s, geared at the analysis of aircraft safety, but has been since then applied to many other fields, including nuclear power plants and the military. A flow chart presenting its steps is shown in Figure 2.2.

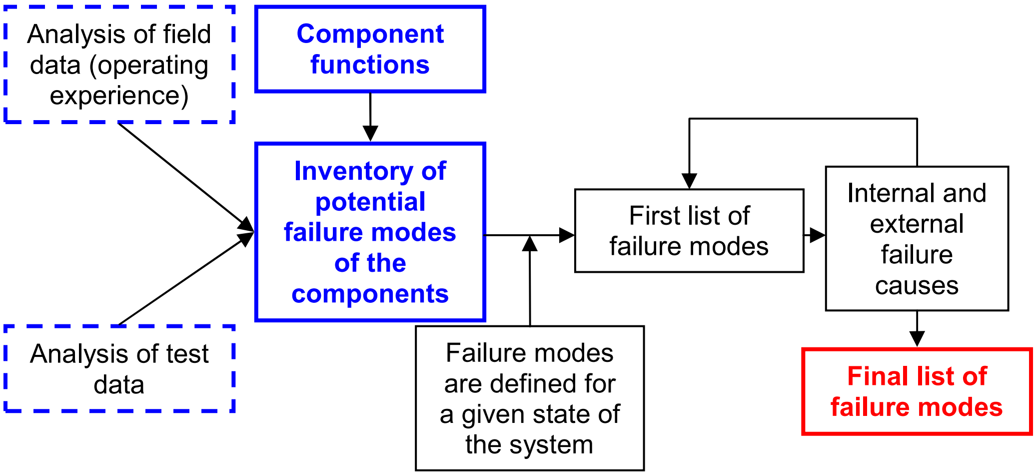


Figure 2.2 Schematic representation of the FMEA. (Adapted from Villemeur, 1992-A)

FMEA's goals are to assess the effects of failure modes of the components in the functions of a system and identify those that affect significantly the aspects of reliability, availability, maintainability or safety of the system. Failure modes are defined by the reliability engineers. Failure modes are different from faults. Failure modes may either be defined by looking at the consequences of faults on a system or by defining a certain fault degree which leads to an action in the system (like a repair action, for instance). Thus, failure modes are dependant on the degree of severity of faults. To avoid ambiguity, physically measurable fault parameters should be defined and linked to each failure mode. In a mechanical system, such as the railway track infrastructure, failure modes may be defined by the crack-opening or wear level of different rail defects (faults).

Due to the definition of failure modes, each type of fault may lead to different failure modes, according to its degree of severity. Also, the combination of different faults may lead to different failure modes. This is the reason for the plurality of failure modes. For instance, if a track defect (fault) like transversal cracking is looked at, different failure modes may include for example "crack opening of 1 mm from rolling surface" to "crack opening of full section height". Whilst the first leads to a repair action such as reprofiling, the other leads to a full system stop and rail replacement action.

In a first step, the system must be defined (functions and components). This means the breakdown or resolution level must be settled, system boundaries, limitations and specifications must be identified. Most importantly, the system state (operation, standby, maintenance, testing…) on which the FMEA is going to be carried must be decided, for normally FMEAs can only be carried out for one system state at a time.

After this the components' failure modes and causes leading to those must be identified (here, the PHA can be useful). This step must not be rushed or overlooked, for the way the system is modelled and therefore all the analyses' results will depend greatly on it. The failure modes and causes must be exhaustively inventoried.

Then a study of the failure mode effects is carried. These should be described methodically and exhaustively, assuming whenever possible that one failure mode happens by itself, while all the other components are working.

Both last steps have room for improvements in an iterative manner since the description of components' failures may help the analyst to realize about second failures and chain effects and improve the system. The problem of interdependence of failure modes may be addressed by the listing of the possible system states, assuming different combinations of component failures and studying the impact each system state has on each failure mode. Nevertheless, it is important to keep goals in sight, otherwise the work may become too burdensome and ineffective (Villemeur, 1992-A).
In the end, conclusions and recommendations can be drawn. If the FMEA is carried out correctly, then there is a reassurance that all the effects of single failures have been identified, how they lead to failure modes, and what are the effects of the latter ones on system functions (and which chain effects lead to the impact on other failure modes). These results can then be used as guidelines for the implementation of alarm systems, new design specifications and methods, design redundancy or maintenance schemes.

This analysis is generally provided in tables, (such as what happened with the PHA). There is an extension of this method called FMECA – the "C" introduced here stands for criticality, meaning an analysis of the criticality of failure modes – degree of severity – is also carried.

2.2.3 Step 3: Life cycle assessment (system level)

In the life cycle step, the goal is to assess the benefits the system will bring. In order to do so, costs and incomes must be assessed. Incomes are relatively well known – in terms of a railway system, the bigger share belongs to faring on passenger tickets or freight services. Another part to quantify is costs. RAMS parameters will play a very big role in determining the costs, since as stated above, failures will bring costs to the system, i.e. the cost of a train to be stopped due to a signalling system failure, the cost of a corrective maintenance operation on a rail track which forces the impediment of traffic, the cost of accidents which might involve serious injuries or death, among others.

So at this step, the idea is to take RAMS simulation results and create targets for the system studied. The targets will influence the system's mission profile and the life cycle's CBA, and the results of the CBA might indicate that the RAMS targets need to be revised. This is thus an iterative process, for the RAMS targets influence the CBA and the CBA influences the RAMS targets. It is then needed to go back to step 2, and rerun the simulations in order to revise RAMS targets. Finally, the testing and operation steps in the life cycle will influence step 1, for data can be collected through the testing and operation phases of the life cycle. This is a way to achieve **reliability growth**.

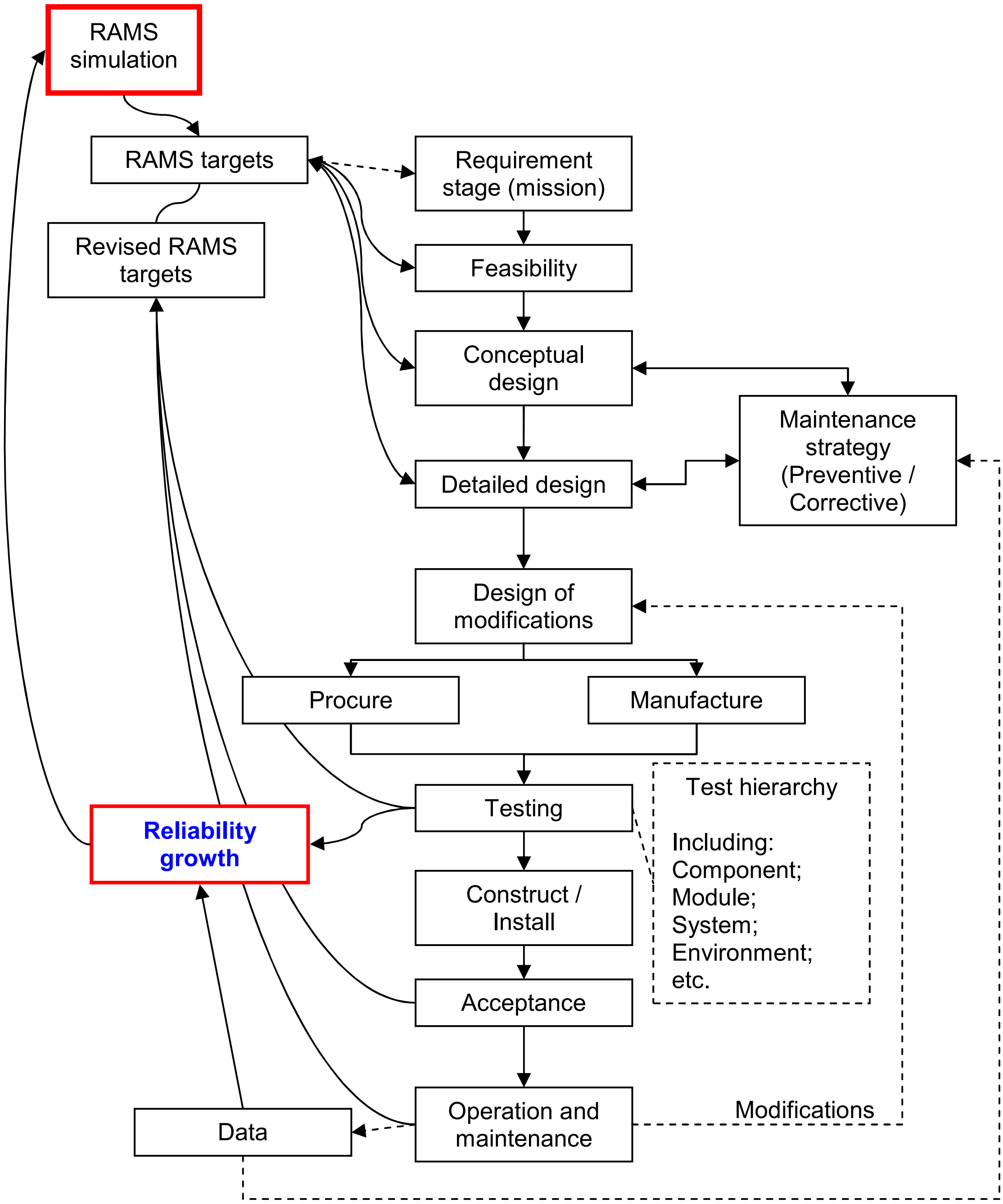


Figure 2.3 The RAMS cycle – connection between RAMS simulation and the life cycle. (Adapted from Smith, 2005)

The RAMS cycle (Figure 2.3) illustrates the presence of RAMS throughout the life cycle of a system. The first thing to do in any project is to define its RAMS targets. It is of utmost importance to define these targets contractually during the requirement step, otherwise they will just be a nuisance in terms of budget and time resources, and they will be purely disregarded. In the railway infrastructure case, there are three main players: the infrastructure owner (normally the government or local authority), the train operator, and the track maintenance manager. There are therefore two contract levels. The first is the concession between the infrastructure owner and the train operator. The owner normally assigns train traffic operation to a train operator, demanding nevertheless comfort levels, being safety levels normally demanded as the occurrence of no accidents. This happens for it is the governments and local authorities' mission to protect the clients (passengers and cargo). The train operator then normally hires a track maintenance manager, or commissions this job to its maintenance department, and turns the desired safety and comfort levels into RAMS targets in terms of failure modes and their occurrence. It is also very important to define the staff to which these tasks will be commissioned. The staff must be technically capable of performing the tasks.

Next, a feasibility assessment is made, which may change the established targets, depending on the trade-off between the difficulty of keeping up to the targets' levels and the available production technologies and resources.

The design step follows simultaneously with the maintenance strategy, and the first predictions of RAMS targets are made using the simulation methods and models listed on the previous chapter. Maintenance strategy is also conceived having in mind RAMS targets. This is of utmost importance for it will influence the predictions made on reliability. Obviously, design and maintenance must be linked, for the ease of maintenance operations will be dictated by design conditions and the logistic support behind it (storing of spares, access to the components, etc.).

Manufacturing and procurement follow, and the first real demonstrations and testing of the systems can take place. If results are satisfactory then installation may start. Finally, in the phase of field operation, data can be acquired to provide feedback to the RAMS simulation methods and models, which may change the first RAMS targets and design premises, eventually leading to revised RAMS targets and design modifications for later versions.

The scheme shown in Figure 2.3 shows all the possible steps where RAMS may be applied but this will obviously depend on the project's scale and on contractual requirements. The scheme presented is the broadest possible, and in smaller projects some phases may be omitted due to time or budget constraints. Contractual issues on RAMS requirements must be settled clearly to avoid pitfalls such as conflicts of interest, undefined responsibilities, misinterpreting of RAMS targets, ambiguity in their definition, or demanding of impossible or immeasurable features.

There are a number of recommendations which can be suggested for each life cycle phase in order to achieve a reliable and safe product or equipment.

In the design step, it is common sense to say that the path to reliability goes through a design which is simple (meaning that it has few different types of components), includes redundancy (though this may improve certain failure modes at the expenses of others), and can withstand greater stresses than the ones required to function properly.
In the manufacturing step the main concern should be about quality control systems, and finally in the field use, adequate operating instructions should be provided, maintenance plans should be followed, and it is important to keep in touch with the costumer through technical supply and costumer feedback.

Cost and benefit quantification is out of the scope of this study. CBAs basically take the forecasts of both and compare them reduced to the same year in an economical point of view, to evaluate whether a project will be worthy or not. Completing with what was said in chapter 2.1, environmental damage and accidents can also be economically quantifiable, and further studies should be carried on this field. They are phenomena of stochastic nature, which may be treated as having independent probability distributions. One idea for their inclusion in RAMS simulation results is to create an additional series block (Annex D.1) to allow for their treatment.

2.3 Reliability parameters

When conducting a RAMS study, the RAMS engineer is faced with a wide choice of parameters to assess. Different reliability parameters are useful in different situations, and as an advantage, it is possible to deduct some parameters from others.

2.3.1 Which reliability parameter to choose?

Data collection should be geared at a set of specific reliability parameters. This must be done having in mind what the goal of the collection is (that is, having in mind what is ultimately sought for in step 3 – chapter 2.2. A list of reliability parameters is provided, and the situations where they should be studied are mentioned:

- **Failure Rate (λ)**: It is the probability of occurrence of a failure at a specific component age, defined in terms of failures per time, load cycle, or cumulative load. Can be useful in almost every component. It is used in steps 1 and 2 of the RAMS life cycle to compute values of component/system failure and unavailability. At system level, it is normally assumed that the failure rate is constant, but at component level this might not be the case. This will be subject to further discussion in Chapter 2.5.
- **MTBF and MTTF**: Whilst **λ** is the number of failures in a certain span of time or load, MTBF is the time or load span between two failures (MTTF is the time span until the first failure, it is

used in non-repairable components). Used while simulating the effectiveness of different maintenance policies and/or the costs linked to them. The higher the value, the less maintenance actions are needed in the same time span.

- **Reliability/Unreliability**: Reliability is the probability of the system to fail over a time or load span. Used mainly where there are security concerns and therefore where the probability of failure is of interest (ex: probability of failure of a braking system in a train).
- **Maintainability**: Seldom used as such (Smith, 2005).
- **Mean Time To Repair**: Useful when calculating costs associated with loss of income due to outage of the system, and with maintenance operations. It is normally expressed in percentiles. Maximum MTTRs are meaningless, because they are log-Normally distributed (lower bound is null and higher bound is ∞).
- **Mean Down Time**: Used in similar situations as MTTR, since they may overlap. In the same manner as with MTTR, maximum MDTs are meaningless. Both MTTR and MDT are known to be log-Normally distributed (Smith, 2005)[3].
- **Availability/Unavailability**: Very useful when calculating costs associated with loss of income due to outage of the system (availability is linked to MDT through equation 2.7). Combines reliability and maintainability. Unavailability is the Probability-of-Failure-on-Demand, extremely useful when assessing safety aspects, such as the behaviour of hazard mitigation components which must remain ready to work in standby while waiting to be triggered (ex: airbags).

2.3.2 Relationships between reliability, unreliability, failure rate, MTTF and MTBF

Reliability (and as such unreliability), failure rate and MTBF (or MTTF) are interrelated. Therefore some of the parameters can be inferred from knowledge about others. The algebraic relationships are presented on Annex B. The results are presented here:

Relationship between reliability R(t) and unreliability Q(t):

$$R(t) + Q(t) = 1 \tag{2.1}$$

Relationship between reliability R(t) and failure rate λ(t):

$$R(t) = \exp\left[-\int_0^t \lambda(t)dt\right] \tag{2.2}$$

[3] For statistical purposes, IEC 61508 states that access and diagnostic test time intervals may be omitted from the sum of MTTR if the diagnostic tests are run almost continuously (Lundteigen & Rausand, 2006), a feature common in more modern equipments.

Relationship between MTTF θ(t) and reliability[4]:

$$\theta(t) = \int_0^{\infty} R(t)dt \tag{2.3}$$

If the failure rate $\lambda(t)$ is assumed to be constant (equal to a value λ), then taking equation 2.2, it is possible to write:

$$R(t) = \exp\left[-\int_0^t \lambda(t)dt\right] = \exp\left[-\lambda t\Big|_0^t\right] = e^{-\lambda t} \tag{2.4}$$

And MTBF can be simplified taking 2.3 and 2.4 to:

$$\theta = \int_0^{\infty} R(t)dt = \int_0^{\infty} e^{-\lambda t} dt = \frac{1}{\lambda} \tag{2.5}$$

2.3.3 Relationship between MTTR and MDT

MTTR and MDT are interconnected. Some literature suggests “down time” can be defined in different ways, overlapping more or less partially with “repair time”. It is important to clarify this.

Down time refers to the time while the system experiences outage, that is, it is not functioning according to its required goals. Repair time is the time the repair actions take to be performed. It can include (depending on the applicability) the operations of access, diagnosis, waiting for spares, replacement, checking and alignment. Thus, down time is bound to be equal or larger than repair time. This happens for several reasons (Smith, 2005):

- There is normally a time for failure detection, though this time is gradually shortening with the introduction of alarm systems in the components themselves;
- There might be a maintenance policy which states that maintenance (and thus repair) operations are done according to pre-defined time intervals or cycles, or even number of failures. This may happen for economical reasons. In this case, failed components might remain un-repaired until a maintenance operation happens;
- In some systems, after repair is completed, there might be a time for checking and alignment of settings of the recently repaired component before setting back the system into a functioning state. This only applies if the definition of “repair time” does not include these operations.

[4] The expression given here is valid for a non-repairable component’s MTTF and also for a repairable component’s MTBF, if it is considered that the repair operation returns the entity to a state equivalent to the state it had before it was new (Nahman, 2002).

A note on the inclusion of checking and alignment operations in MDT and MTTR: in some systems these operations are considered to be out of the definition of “down time”, for they may happen while the system is considered to be functioning. Whenever these two operations require manpower they are considered to be inside of the definition of “repair time”.

A parameter related to MTTR is the repair rate μ. It expresses down time as a rate:

$$\mu = 1/MTTR \tag{2.6}$$

2.3.4 Relationships between availability, unavailability (PFD) and MDT

As it was stated in the definitions (Annex A), unavailability, also known as probability of failure on demand, can be written as:

$$\overline{A} = 1 - A = \frac{\lambda MDT}{1 + \lambda MDT} \cong \lambda MDT \tag{2.7}$$

A special case of simplification is the one of components with unrevealed failures, i.e. failures which are only detected and simultaneously repaired in scheduled maintenance operations. Here, the MDT shall be half the interval between two of these maintenance operations (plus its mean time to repair, which may be neglected if it is small in comparison to the MDT). Thus, if the proof interval (time lapse between two maintenance operations) is T, then, applying equation 2.7, we obtain:

$$\overline{A} = \lambda MDT = \lambda\, T/2 \tag{2.8}$$

Finally, availability may be seen as a ratio of Up time over Down time (Annex A):

$$A = \frac{Up\ Time}{Total\ Time} = \frac{Up\ Time}{Up\ Time + Down\ Time} \tag{2.9}$$

2.4 Distribution of failure rate in time & the bathtub distribution

In some mathematical expressions written above, simplifications were made assuming that the failure rate is constant. This is what is generally assumed at system level, but at component level this may not be the case – and generally speaking, it isn’t. One of the most used distributions for failure rate of components in RAMS is the bathtub curve. This curve owes its name to the shape it creates when plotted.

The bathtub curve sits on the following premises:

- In an early lifetime, components experience more failures due to a period of adjustment of their parts and interaction with other components – take for instance a car-engine's break-in period (also known as burn-in).
- In the mid and most part of their lifetime – useful time – the components experience a constant failure rate.
- Close to the end of a component's lifetime, the failures increase due to wearout of parts and other ageing related problems.

Here the mid part relates to the above mentioned failure classification of Random hardware failures and the non-constant parts to the Systematic failures (Smith, 2005).

In fact, the bathtub distribution may be seen as the sum or juxtaposition of three different distributions. This is illustrated in Figure 2.4:

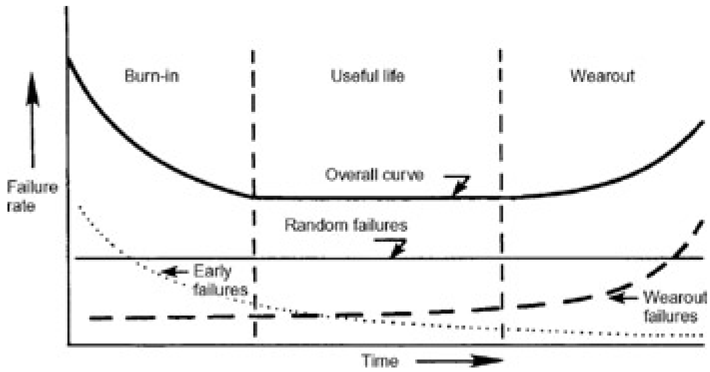


Figure 2.4 The bath-tub distribution. (Source: Smith, 2005)

This distribution, or in other cases just the wearout curve, are believed by a large number of authors to be the dominant failure distributions in components, though some authors disagree (Mather, 2005). This was stated in RCM, or Reliability Centered-Maintenance, a report written by Stanley Nowlan and Howard Heap, from United Airlines for the US Defence Department in 1978.

According to Nowlan and Heap's results, there is only one predominant failure rate distribution out of six that in fact shows failure rate to be constant throughout the entire life-time of components (Nowlan and Heap, 1978, pg. 67).

This is the case in railway infrastructure components, where most of the failure rates grow with load. Therefore during their study, formulae for non-constant failure rates must be used (Chapter 2.5).

2.5 Dealing with failure rate data

Different expressions or cumulative distribution functions[5] can be written and assigned for the failure rate from empirical data sets. The expressions or distributions will differ from case to case, and accordingly testing data will need to adapt to the calculation demands. Four different cases are presented on Table 2.1:

Table 2.1 Different cases of combinations of failure rate and number of failures.

Relative number of failures / Constant or non-constant failure rate over component's life (time dependency)	High number of failures	Low number of failures
Constant	$\lambda = k/T$ (chapter 2.5.1)	Statistical interpretation – χ^2 (chapter 2.5.2)
Non-constant	Use Weibull distribution (chapter 2.5.3)	Inadequate data (chapter 2.5.3)

2.5.1 Constant failure rate (non time-dependant), high number of failures

Suppose a set of *N* non-repairable components is tested during time *T*, and that a number *k* (≤N) of failures occur. Failure rate may be written as $\lambda = k/T$, in the case of a constant failure rate (that is, if failures occur at approximately similar time intervals). The statistical meaning of this formula cannot be disregarded, and so, if the components present a low failure rate, there are bound to be less *k* failures. This can create a problem when trying to yield data from tests, since if different samples display several different small *k* values for rather high similar *T* values, there are going to be large differences in the values. So the formula $\lambda = k/T$ is mostly used for cases with a high number of failures (i.e., cases with high values of *k* when compared to *N*) with constant failure rate displayed.

2.5.2 Constant failure rate (non time-dependant), low number of failures

The only way to always yield the correct value of MTBF (θ) or failure rate (λ) is to wait until all the N components being tested to fail – such a test is then said to be "uncensored" (Villemeur, 1992). Since nowadays' components are designed to achieve low failure rates, in a magnitude order of 1 to 10 per million hours, it becomes very impracticable to let all of them fail during one of these tests. It is also a matter of time-saving during the project phase – nowadays' project deadlines are becoming

[5] A cumulative distribution function is the function that completely describes the probability distribution of a real-valued random variable. The return is a number between 0 and 1. A probability density function is a function which represents the probability distribution in terms of integrals.

increasingly shorter. Due to this need, it was observed that it is possible to truncate these tests either in:

1. A given time limit.
2. An *a priori* set number of failures n (k=n).

This corresponds to the second case, where constant failure rate is combined with a low number of failures. Statistical treatments are then needed to yield results from these tests.

The most common approach takes the fact that the expression:

$$\frac{2k\hat{\theta}}{\theta} \qquad (2.10)$$

where *k* is the number of failures observed during test time and θ is MTBF, follows a chi-square (χ^2) distribution, provided that the failure rate is constant (Smith, 2005, chapter 5.3). Degrees of freedom will be discussed later. Since $\theta = T/k$ (*T* is the test time) it is possible to apply this to expression 2.10, returning:

$$\frac{2k\hat{\theta}}{\theta} = \frac{2T}{\theta} \qquad (2.11)$$

The latter term in expression 2.11 will accordingly follow a chi-square distribution as well. Since test time *T* is known, it is then possible to obtain MTBF (*θ*) through the chi-square distribution, with whichever degree of confidence $(1-\alpha)$ chosen. Normally the confidence interval is set with the upper bound of MTBF as infinite – for this is the worst case scenario –, and the other value as the lower bound, though two-sided confidence intervals may also be used.

Degrees of freedom vary. If the test is failure-truncated, then there are *2k* degrees of freedom. If it is time-truncated there are *2(k+1)* degrees of freedom for the lower bound of MTBF, since it is considered that there could have been a failure just after tests stop. This also represents the practical solution for tests that display no failure whatsoever.

2.5.3 Non-constant failure rate (time-dependant)

Non-constant failure rate must be handled with thorough care. It is commonly solved by working with reliability and MTBF directly. It is of little worth working with the actual failure rate. As shown in chapter 2.3.2, reliability is calculated through expression 2.2:

$$R(t) = \exp\left[-\int_0^t \lambda(t)dt\right] \qquad (2.2)$$

Since now we are dealing with the non-constant case, the failure rate function can be extremely complicated and even difficult to write analytically, and moreover it will be even more complicated to integrate to calculate reliability.
What is normally done in the case of a non-constant failure rate – especially in mechanical components (Villemeur, 1992-A) – is to use the Weibull method, which owes its name to Prof. Waloddi Weibull. He created a versatile function which allows modelling of reliability by using expression 2.12:

$$R(t) = \exp\left[-\left(\frac{t-\gamma}{\alpha}\right)^{\beta}\right] \quad (2.12)$$

The function has three calibration parameters: γ, β and α. These are adjustment parameters that allow the function to produce Reliability and MTBF results from test values. The parameter γ is normally known as the location parameter. Its value is normally set to null on the first iterations of the Weibull function, and its function is to displace the plot along the horizontal axis. Parameters β and α are known as shape and scale parameters respectively. The function of β is to provide shaping to the plot, which can vary a lot with it in normal axis scales. In log-log paper a Weibull function correctly adjusted will be a straight line and β will simply deliver the value of the slope. The special case of β=1 and γ=0 is a simplification of the Weibull function which yields the exponential distribution. For β<1, the failure rate decreases with time. For β>1, it increases (Villemeur, 1992-A). Parameter α is also known as the characteristic life parameter.

The process to obtain the Weibull function for a given set of data is presented in chapter 4.2.3. Weibull parameters may be obtained with any data processing software such as Excel through a linear regression of the plot.

A warning must be done regarding Weibull distribution and non-constant failure rates in general. Plotting is **not significant** if it has less than 6 points. 10 points are recommended. 0, 1, 2 or 3 points do not suffice to understand if the Weibull function applies, i.e., if it is a non-constant failure rate case. Thus, **a low number of failures for a failure rate which is known to be non-constant is insufficient as data**, and it is generally better to assume a constant failure rate and use the chi-square method stated above (Smith, 2005).

The function of failure rate with parameter *t* given by the Weibull function is given by expression 2.13 (Villemeur, 1992-A; Nahman, 2002):

$$\lambda(t) = \frac{\beta}{\alpha}\left(\frac{t-\gamma}{\alpha}\right)^{\beta-1} \quad (2.13)$$

2.5.4 Other distributions

Other time distribution functions can be obtained from different cumulative distribution functions. The exponential case mentioned above yields a constant failure rate λ, as said before. A summary of Weibull and exponential, as well as other distributions are presented in Table 2.2. Description of the main domains of usage of each distribution also follows:

- Exponential distribution: The probability of future failures from a time instant t_1 does not depend on up-time until that time instant. Distribution lacks "memory". Constant failure rate. Cumulative Distribution Function (CDF) may be used to model up times of systems subject to unusual random stresses which usually come from an environmental source – components of electronic and electrical systems. Notice these components behave mostly according to the bath-tub curve distribution, so this is applicable to their middle lifetime.
- Weibull distribution: already discussed above. β=1 and γ=0 yield the exponential distribution.
- Uniform distribution: sometimes used to model preventive maintenance duration.
- Normal distribution: μ is the location parameter (-∞<μ<∞) and α is the scale parameter (α^2>0). The Probability Distribution Function (PDF) is symmetrical about the mean value, M=μ, which means that there is the same probability that the variable is greater or smaller than its mean. It's also a consequence of this that the variable may have negative values too, and thus it cannot be used to model system state residence times. It is instead used to estimate errors of prediction of several quantities subject to uncertainty. The standard Normal distribution is presented for ease of calculations, and the truncated normal distribution is defined for modelling of system state times (it is truncated for positive values only).
- Gamma distribution: β is the shape parameter and Θ is the scale parameter. For β=1 the distribution converts also in the exponential distribution, and for large values it approaches the normal distribution. If β=k and k is a positive integer, then it converts into Erlang's distribution of order k, of which the CDF is presented in the table. Erlang's is able to describe, for instance, the up time of a system composed of k components whose lifetimes are all exponentially distributed with a failure transition rate being equal to kΘ (Nahman, 2002). Of these components, only one is working while the others are in a standby state which does not allow failure to happen, being merely replacement units for the working component – the up time of the system will then be the sum of components' up times.
- LogNormal distribution: used sometimes for modelling of repair duration of assets.

Table 2.2 Statistical distributions applicable in RAMS.

Name	CDF (Cumulative Distribution Function) ≡ Unreliability 1-R(t)	PDF (Probability Distribution Function) ≡ Failure density U(t)	Failure rate λ(t)
Exponential	$F(x)=\begin{cases}1-e^{-\lambda x} & x\geq 0\\ 0 & x<0\end{cases}$	$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & x\leq 0\end{cases}$	$\lambda(t)=\lambda$
Weibull	$F(x)=\begin{cases}1-e^{-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}} & x\geq\gamma\\ 0 & x<\gamma\end{cases}$	$f(x)=\begin{cases}\frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta}e^{-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}} & x\geq\gamma\\ 0 & x<\gamma\end{cases}$	$\lambda(t)=\frac{\beta}{\alpha}\left(\frac{t-\gamma}{\alpha}\right)^{\beta-1}$
Uniform	$F(x)=\begin{cases}0 & x\leq a\\ \frac{x-a}{b-a} & a<x\leq b\\ 1 & x>b\end{cases}$	$f(x)=\begin{cases}\frac{1}{b-a} & a<x\leq b\\ 0 & otherwise\end{cases}$	$\lambda(t)=\begin{cases}0 & t\leq a\\ \frac{1}{b-t} & a<t<b\end{cases}$
Normal	$F(x)=\int_{-\infty}^{x}f(u)du$	$f(x)=\frac{1}{\alpha\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\alpha^2}}$	$\frac{U(t)}{R(t)}$
Standard Normal	$\Phi(z)=\int_{-\infty}^{z}\phi(u)du=0.5+\int_{0}^{z}\phi(u)du \quad z=\frac{x-\mu}{\alpha}$	$\phi(z)=\frac{1}{2\pi}\exp\left(-\frac{z^2}{2}\right)$	-
Gamma	$F(x)=\begin{cases}1-e^{-k\Theta x}\sum_{i=0}^{k-1}\frac{(k\Theta x)^i}{i!} & x>0\\ 0 & x\leq 0\end{cases}$ (see description above)	$f(x)=\begin{cases}\frac{\beta\Theta}{\Gamma(\beta)}(\beta\Theta x)^{\beta-1}e^{-\beta\Theta x} & x>0\\ 0 & x\leq 0\end{cases}$	-
LogNormal	$F(x)=\Phi\left(\frac{\ln x-\mu}{\alpha}\right) \quad x\geq 0$	$f(x)=\frac{1}{x\alpha\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\alpha^2}\right)$	$\frac{U(t)}{R(t)}$

2.5.5 General method given the number and time distribution of failures

This method is presented by Nahman (2002). It states that unreliability may be seen as the ratio between number of entities failed (taking non-repairable entities as an assumption) and number of entities that originally existed, as expressed in 2.14:

$$Q(t) = \frac{n(t_n)}{N(0)} \qquad (2.14)$$

Where $n(t_n)$ is the number of failures registered until time t_n and N(0) is the number of entities working at time t=0 (i.e., at the beginning of the reliability test).

The probability density function is then given by 2.15:

$$f(t) = \frac{n(t_n)}{N(0)\Delta t_{n,n+1}} \qquad (2.15)$$

And failure rate is defined by 2.16:

$$\lambda(\Delta t_{n,n+1}) = \frac{n(t_n)}{N(t_n)\Delta t_{n,n+1}} \qquad (2.16)$$

The test time should be split into time intervals $\Delta t_{n,n+1}$, which must have the same length. The idea is to distribute them in a way which doesn't allow for uneven concentration of failures in the intervals, and that, according to Nahman (2002), is achieved through splitting of the total test time in a number of K intervals, obtained from either of the 2 empirical formulae:

$$K = 1 + 3{,}3\log(n) \quad or \quad K = \sqrt{n} \qquad (2.17; 2.18)$$

An example of a comparison of the use of two values for the same test is shown in Figure 2.5.

This leads to a table-like calculation of the several values of $Q(t)$ and $\lambda(t)$, which can then be plotted. Reliability values are achieved through $R(t) = 1 - Q(t)$.

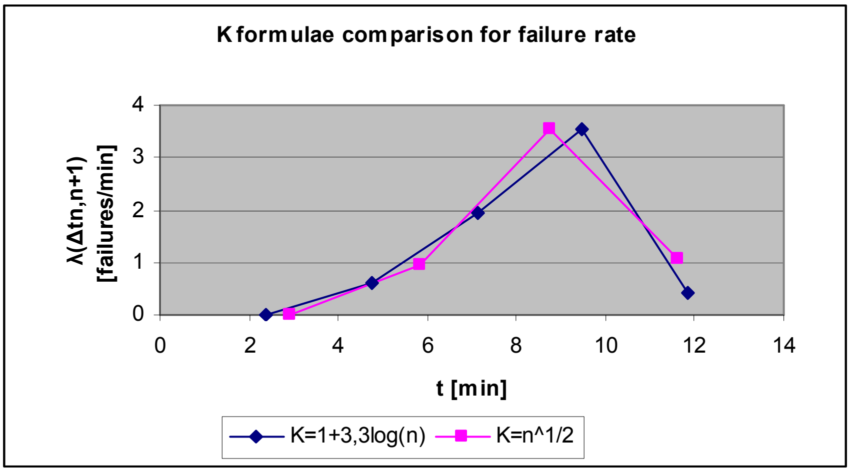


Figure 2.5 Comparison of the use of formulae 2.17 and 2.18 for test-time splitting. Results for $\lambda(\Delta t_{n,n+1})$. Test with 11 failures out of 20 components over 12 minutes.

In a similar way to what was stated above, MTBF or MTTF (in the case of non-repairable entities) can be calculated when all possible failures occur, and its calculation is simply given by 2.19:

$\theta = \frac{\sum_{i=1}^{N(0)} T_i}{N(0)}$, where T_i are the time instants when the i failures occur. (2.19)

2.5.6 Method regarding the mechanical behaviour of components

This method is presented by Villemeur (1992-B) or the assessment of failures of mechanical components. In this type of components, failure happens whenever stress becomes greater than strength. Typically components are designed according to standards which impose a ratio α between the expected value of strength S and the expected value of applied stress s, given by 2.20:

$$\alpha = \frac{E[S]}{E[s]} \quad typically > 1 \qquad (2.20)$$

This ratio is known as a safety margin and will provide a relative degree of safety. The problem lies in the fact that both strength and stress are not constant; they are rather described by distributions. There is a slight chance that a peak in stress s might overcome a low in strength S, and when this happens a failure occurs. Since reliability is the probability of a failure not happening, then reliability may be written as:

$$R = P[s < S] \qquad (2.21)$$

Several methods to solve equation 2.21 are suggested by the author. One is to apply the Monte Carlo method. Monte Carlo simulations are ran whenever a complex algorithm or equation with a probabilistic input needs to be solved. The simulation generates several values of the probabilistic variables involved, provides them as input and returns the final result. In this case several values of stress s and strength S are generated based on their distributions and compared. Reliability is then given by expression 2.22:

$$R = \frac{number\ of\ times\ S > s}{number\ of\ runs} \tag{2.22}$$

Another method is the analytical solution shown in expression 2.23, easier if the relation α is known.

$$R = \frac{1}{1+\alpha} \tag{2.23}$$

There are also methods of dealing with fatigue, which basically use S-N curves with different probabilities, that is, the curves which define all the pairs of ultimate tensile strength and number of cycles (S-N) which correspond to a given probability value of failure. Reliability may then be estimated from these curves. The method is fully described by Villemeur (1992-B).

2.5.7 Warnings and recommendations

Independently of the method chosen, it is very important to use engineering judgement over statistical judgement. The latter looks barely at numbers without considering the actual way the systems or components work, so engineering judgement may provide the first hints to figure out which method to apply. It is also very important to clearly separate failure modes. It has been noticed – Drenick's theorem[6] (Smith, 2005) – that if there are at least 3 different failure modes combined in a data set, the behaviour read in it will be the one of a constant failure rate, even though each of the failure modes might be non-constant with very different β and α parameters.

When there are many different components with non-constant failure rate – which might behave in the same way – that are subject to replacements and maintenance at different time moments, the combination of the different failure rate functions will turn into one continuous and constant failure rate phenomenon. There are methods to deal with this, like CUSUM and Duane – two reliability growth models – or the Laplace test which verifies if there is a trend in the failure rate (Smith, 2005).

[6] Drenick's theorem states that even if individual components fail according to different distributions, over time, the successive times of a repairable series system follows the exponential distribution, which displays a constant failure rate

2.6 Reliability data sources

Different listings of component or system reliability parameters have been gathered over the last years by several entities or organizations. These are meant to be used as input data for RAMS simulation models used in step 2 (chapter 2.2.2).
This type of listings flourished in the 1980s and beginning of 1990s, but since then most of the lists have not been updated. Since reliability is a comparatively new field, meaning that sometimes collection processes lack standardizations or comparison methods, different databanks may often show failure rates of very different values, even magnitude orders, for the same type of components. Other point that deserves special attention is that unless otherwise specified, the values contained in these lists refer to random failures (please refer to the definition of "random failures" in Annex A), meaning that these represent constant failure rates (Smith, 2005). If for instance, the bathtub distribution is considered, then these failure rates refer to the constant part of the bathtub.

Each list must be carefully looked at, and each author's recommendations must be followed through thoroughly. It is always best to use the most specific type of reliability data source for the purposes meant, i.e. to use company or site specific data instead of industry generic data.
Whether field data or testing data is being used, there is a large range of variation in reliability parameters. Specific sources of variability for reliability parameters in field or testing data are:

- Some databanks include components replaced during preventive maintenance actions. Ideally these should be excluded from testing data and dealt with separately, but such does not always happen since sometimes they are hard to identify;
- The definition of failure might vary. In some databanks failure may simply be a drift from the expected value of a certain parameter. In other databanks, this may not be accounted as a failure if no repair actions are needed. This happens due to design tolerance, the ability of components to perform in a range of parameters.
- Environmental and quality parameters must be defined, for there might be a broad range for each of them.
- Reliability parameters are often referred to "component types". If a broad range of components that fall into the same type is tested, large ranges of values for parameters are likely to be yielded.
- Different failure modes are sometimes mixed in the data.
- Where field reliability data is collected, reliability growth is likely to improve reliability failure rates over time, thus leading to a gradual change in the failure data collections.
- Failure rate can be inflated/deflated due to the way the failures are accounted. If the data collection procedure is ill-devised, lacking specific indications of what to account and what not to account as failures, then the accounting of failures will sometimes be left to the analyst's decision. This must be eliminated through stating of collection procedures and the use of data forms with unambiguous codification.

Data banks will often present tabled results, but might also provide mathematical equations that require some input data in order to return the values (normally failure rate values). The latter are to be used extremely carefully since they provide a totally wrong perception of accuracy (Smith, 2005). Interpolations among the data series may be estimated to a certain extent, but it is extremely ill-advised to try to create extrapolations using these equations.

Most of the times data ranges are presented in tables, often two columns presenting a lower and higher bound and a middle column presenting the most predominant value. If the RAMS engineer does not have any special indication to use any of the boundary values, the best practice is to use the predominant value.

If no predominant value is supplied, arithmetical or geometrical mean can be used. It is normally best to use the geometrical mean instead of the arithmetical mean when the boundary values are presented with one or more magnitude orders of difference (Smith, 2005). This happens for the arithmetical mean favours the bigger magnitude order value, whilst the geometrical mean evens both terms out.

The lower bound values represent what can be expected in terms of reliability parameters after reliability growth is achieved. The most predominant values are of great credibility for they are usually confirmed by several sources. The higher bound values are normally applicable to systems with unrevealed failures, that is, failures that are not revealed during normal operation, only during maintenance. If in one of these systems the maintenance screening operations are very frequent, this may lead to a bigger number of failures in the same time period than a system with a lighter maintenance plan. In Annex C, a list of significant reliability databanks and sources is presented, listings which have sprung from various branches of industries and which deal with many different components.

Worth of mention is the FARADIP.THREE bank, which is a data range, i.e., a data bank collection comprising many other databases. Its purpose is to show the ranges of possible failure rate values due to the plurality of collections. The author of FARADIP.THREE also worked in presenting the confidence levels of values related to the mean values of failure rates, and this was done for each of the three general types of databanks mentioned below (Smith, 2005).

There are three general types of databanks: generic, industry specific, or company/site specific. Obviously, it is preferable to use the most specific banks when available, since they show, as a rule, a better degree of precision. It is important to once again state that system failure is not the same as component failure, though there is a strong link between both. Interaction between components must be accounted for, especially in more complex systems (i.e. common failure modes). Other factors such as human error or environmental conditions may also contribute to this dissociation of the terms. So the data presented on these data sources must be used with thorough care, not in an absolute way, but more in a relative way as to provide indicators of reliability levels, helping in decision-making processes. In other cases, they may be used to predict reliability or safety levels expressed as requirements – for instance, contractual requirements.

Several authors point out that the choice of failure data is more relevant to the failure prediction results than the refining of the statistical calculations themselves while modelling (Smith, 2005; Villemeur, 1992-A; Lundteigen & Rausand, 2006; Nunns, 2002).

Thus the engineer's tasks are: to choose the most specific databank available, or if none specific enough is available, to stick to a coherent choice in the databank or various databanks chosen; to carefully process the data in order to present the full scope of possible results; and finally to develop and present in the modelling reports the degree and extent of applicability and the vulnerabilities of the results obtained using values from those databanks.

2.7 Reliability data collection procedures

As it was stated above (Chapters 2.1.2 and 2.6), one of the biggest problems in RAMS resides in the inaccuracy and variability of the data used. A big part of this is due to a lack of stated procedures for the collection of test and especially field data. This must be avoided at all costs, once again mainly on field data, for this type of data is much more representative of the actual behaviour of components in systems when subject to real environmental and operational conditions and maintenance strategies.

Field data recording is ultimately done by humans – especially in the railway system case – and as such it is subject to errors. Good practice suggests that forms for data collection should be created, using unambiguous codification for every aspect dealt with. Also, staff training is essential and finally motivation should be provided. If there is a lack of budget for a RAMS assessment, one of the first fields to be overlooked is data collection. The RAMS engineer will barely skimp it or oversimplify the descriptions in the reliability data reports. If there is a lack of training and consensus regarding failure recording, then some engineers might account failures where others do not, leading to inflation/deflation of failure data.

Smith (2005) suggests that the following information should be recorded:

- Repair time (active and passive).
- Type of failure (primary/secondary, random/induced, etc.).
- Nature of failure (stress peak, wearout, design deficiency, etc.). An assessment of redundancy parts should be made with the purpose of ascertaining if it is a common cause failure.
- Effects and consequences of the failure (including accidents, injuries and possible deaths) and, much more importantly <u>the physical mechanism leading to failure.</u>
- Failure location (in the system and component).

- Environmental conditions at the time of failure.
- Action taken against failure.
- Personnel involved.
- Equipment used.
- Spares used.
- Cost of outage divided in loss of production or operation, personnel costs, equipment costs and spare costs (may be done in post-processing).
- Component's installation date and running time since last failure.

The importance of registering what the physical mechanism that led to the failure is, and in which component, is connected to the fact that the application of the Weibull distribution works best with only one failure mode in one component at a time. Component installation date is crucial to understand if it is expectable to have more burn-in, random or wearout failures, i.e. to know in which step of the "bath-tub" the component is. Running time may refer to calendar time or to operational time. In some components one may be more important than the other, and a consensus must be reached about this in beforehand. Also in mechanical components subject to fatigue and to loading cycles, number of cycles or cumulative stress may be the units used to measure failure rate.

The presentation of the data is normally done using a Pareto analysis, which ranks the failures regarding their impact – usually, the impact of costs. The process is to count the frequency of each failure and multiply that number by a pre-determined constant which represents the impact (cost) of a single failure. In this way the failures that are more significant have a higher final value as a result. The top ranked failures may then be studied in deeper insight and measures towards their elimination can be taken (re-design, implementation of redundancies, etc.).

2.8 Predicting reliability growth

As it was mentioned before, reliability growth is expected in many systems. The methods presented above which deal with reliability parameters provide a collection of data regarding past events. It is known that reliability growth leads to an improvement in reliability parameters. The question is: how to predict reliability growth?

The most widely applied model is the one of J.T. Duane (Smith, 2005), created in 1962. The model states that the improvement in MTBF is proportional to T^{α} where T is the total equipment time and α is a growth factor. This is represented in expression 2.24:

$$\theta = kT^{\alpha} \Rightarrow \frac{\theta_2}{\theta_1} = \left(\frac{T_2}{T_1}\right)^{\alpha} \qquad (2.24)$$

The growth factor α and the constant k can thus be obtained by taking any two known pairs of values of θ and T. Then, assuming that the growth factor does not change, predictions for MTBF can be made. The parameter α is variable, and depends on the type of components.

Chapter 3: The rail

3.1 Constituents of the track system

The typical rail track infrastructure system is represented in Figure 3.1.

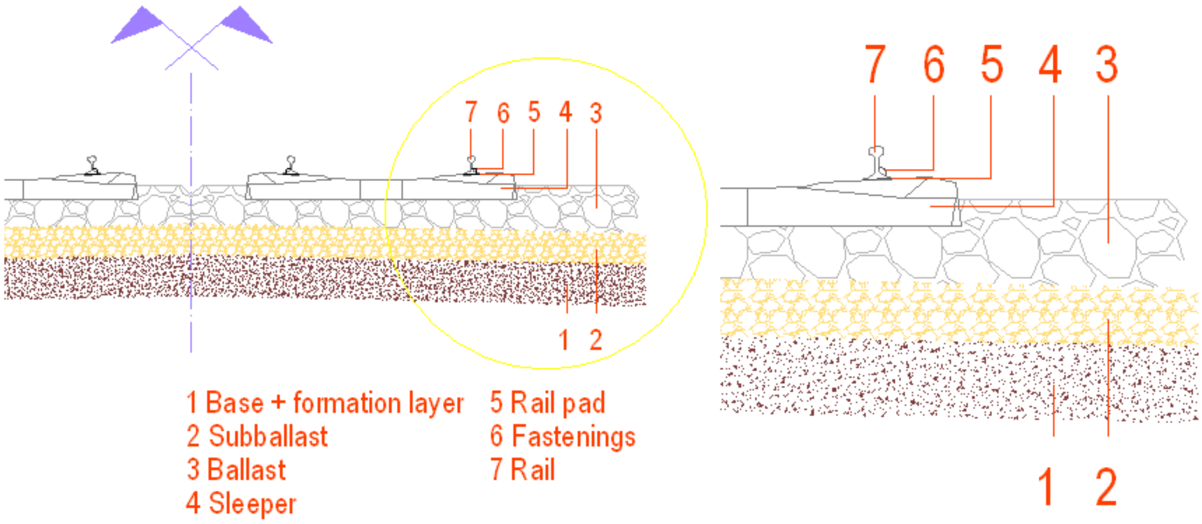


Figure 3.1 Typical cross section of a railway double track infrastructure.

Using Figure 3.1 as a reference, from bottom to top, subgrade is composed of:

- *1.a Base*. Normally composed of on-site soil (or soil transported from nearby excavation sites), which may be consolidated through mechanical means or stabilized by chemical means (Profillidis, 2000).
- *1.b Formation layer*. This layer is used whenever the base layer is composed of soil of poor quality, sitting below the subballast and above the base.

The subgrade is designed according to the superstructure design, geotechnical and hydrogeological conditions, soil type and mechanical strengths involved. It should provide flexibility enabling a smooth ride, limit/control eventual settlements and provide stability and durability of the formations for the imposed loadings of adjacent embankments, constructions and trains at the design speed.
It should also provide good drainage of ground water whether it is caused by rain or melted snow.

According to the UIC standards, there are 4 different subgrade classifications:

- → Subgrade S_3: Low settlements and very good support of trains.
- → Subgrade S_2: Medium behaviour regarding settlements and the withstanding of train loads.
- → Subgrade S_1: Large settlements and less satisfactory performance in withstanding loads.
- → Subgrade S_0: Large, extensive settlements and poor performance in withstanding loads.

From bottom to top, the track superstructure is composed of[7]:

- *2. Subballast or ballast concrete layer*. Both are used for a further distribution of loads to the subgrade. The subballast consists of gravel-sand and prevents ballast stones from penetrating into the subgrade. Ballast concrete layer is usually made of poor grade concrete and is used to create a levelled surface before the final, more precise levelling created by the concrete slab.
- *3. Ballast or concrete slab*. The ballast normally consists of crushed stone, only exceptionally of gravel. Its function is to distribute the loads from the sleepers to the subballast or ballast concrete layer, dampen vibrations and increase ride comfort, reduce noise and ensure a quick drainage of rainwater. The concrete slab is an alternative to ballast which takes up less room (has a smaller cross section) – see the description below.
- *4. Sleepers (*or *ties* – U.S.*)*. Normally made from timber, concrete and in few cases of steel (temporary or very low tonnage networks), they are used to distribute loads applied to the rail to the ballast or concrete slab, and to keep the rail in the correct gauge and geometrical conditions. Timber sleepers are not used in high speed networks for their transverse resistance is not great enough to distribute loads from high speed trains. The sleeper types used in high speed networks are mono-block pre-stressed and twin-block reinforced concrete, the latter displaying a superior transverse resistance, but the first a better elasticity (adaptation to alternating stresses) (Profillidis, 2000).
- *5. Rail pads.* These pads are used to transfer loads from rails to sleepers while damping high frequency vibrations, reducing noise and problems with ballast particles. They can be made from rubber, cork, or other soft materials.
- *6. Fastenings.* These are the components which ensure the connection between rail and sleeper. Types are divided into rigid and elastic fastenings. The use of elastic fastenings is mandatory when concrete sleepers are used (no rigid fastenings may be used in this case). For the rest of the situations – timber or steel sleepers – any fastenings may be used. Elastic fastenings currently in use are the DE clip, Pandrol standard clip, Pandrol Fastclip, Vossloh fastening and the Nabla clip (Esveld, 2001).
- *7. Rails*. Made from steel, may differ in shape, size, weight per length unit, inertia, steel grade (ultimate tensile strength) and regarding placement, they can differ in gauge according to the rail network. They are used to guide and support the rolling stock and locomotives.

[7] Draining systems, bridges, tunnels, power supply (catenary) and related, switches and signalling systems are not going to be discussed in this study, and the reader is therefore referred to Profillidis and Esveld (Profillidis, 2000; Esveld, 2001) for a description of these.

Regarding the layers below sleepers, two systems may exist. The flexible support consists of ballast+subballast and the inflexible support consists of concrete slab+ballast concrete layer. The flexible support allows a certain degree of readjustment in case of settlements – which is of extreme importance – and reportedly provides a more comfortable ride and is very effective in dampening noise. Inflexible supports may be built using a smaller (narrower) cross section, being helpful in tunnel sections, and since they have no ballast, they require much less maintenance (Profillidis, 2000). Sometimes lines have a mix of both supports, for instance some tracks are built mainly on flexible supports but in tunnels change to the inflexible support. The transition of supports can be felt by passengers and as such rubber pads must be placed along appropriate extensions at the entrance and exit of tunnels. In some high-speed lines or lines with frequent traffic, concrete slabs are used due to the large imposed stresses. This is for instance the case of the majority of the extension of the Nürnberg-Ingolstadt-München high speed line (source: SPG Media PLC).

Also, the imposition of large stresses creates more maintenance problems for tracks which use ballast systems. When stress is applied to the rail, it then passes on to the sleepers and subsequently to ballast and the subgrade. In the contact points between ballast and subgrade, infiltration of subgrade particles and ballast particles happens on each of the layers. The infiltrated subgrade particles cause problems in water drainage. Also, due to the repetition of stress cycles, ballast particles settle and the ballast bed geometry is affected.

Several attempts have been made to redistribute and reduce the stress applied from the sleepers to the ballast. Normally this is done by increasing rail profile height – redistribution of stress happens – and reducing the spacing of sleepers. Two mentions are to be made: the first to the German wide sleeper, which successfully reduces stress through the use of a wider contact area, and the Austrian frame sleeper, which uses a framework of sleepers (longitudinal and perpendicular), rather than just the perpendicular sleepers.

Finally chemical binders are used mostly in tracks where ice formation on carriages is possible. When ice blocks formed underneath train carriages become detached from them at high speeds, they hit the ballast and produce an explosion-like reaction.

This study is focused on the rail, since it is a fundamental component in service level and safety aspects. It is a component suitable to be studied with RAMS tools. Different failure modes (degree of severity of rail defects) may be defined by rail infrastructure maintenance managers and individual RAMS studies can then be carried. Finally, these defects may be corrected with specific maintenance actions, which enable the restoration of the rail to an operable state. Therefore, the next sections are dedicated to rail characteristics, rail defects and rail maintenance operations.

3.2 Rail characteristics

Two main types of rails are currently in widespread use:

- Grooved rail
- Rail with base (also known as flat bottom rail or Vignole-type rail)

The grooved rail is used whenever the pavement surface and the rail surface are aligned at the same level. This is the case of most tram and light rail systems. The rail with base is the most common rail type in all other situations (ballast or concrete slab support), including high speed rail networks.

3.2.1 Axle load and traffic load

These two parameters are decisive to assess the stresses and therefore the type of infrastructure that should be implemented on a railway. Axle load is the load each axle of the train carries. It is typically divided into 4 categories (Profillidis, 2000):

- A: Maximum load per axle 16 t
- B: Maximum load per axle 18 t
- C: Maximum load per axle 20 t
- D: Maximum load per axle 22.5 t

Some networks may use larger axle loads, for instance in the US, where the railway network is mainly used for freight transport, maximum axle load spans from 25 t to 32 t. In Russia where broad gauge is used the maximum axle load is 25 t.

Traffic load or tonnage is a way to sum all the loads of all kinds of passing trains in a line during a time period – whether these are passenger trains, freight vehicles, locomotives, or others – and turn them into equivalent passenger train loads. It depends on the passing train loads, type of traffic and the speeds on the line (V). The higher this value, the more stresses applied to the railway line. Traffic load is represented by T, and given by the expression 3.1 (Profillidis, 2000):

$$T = S \times T_{th} \tag{3.1}$$

Where S=1.0 for lines without passenger traffic

S=1.1 for lines with mixed traffic and $V_{\max} < 120 km/h$

S=1.2 for lines with mixed traffic and $120 km/h < V_{\max} < 140 km/h$

S=1.25 for lines with mixed traffic and $V_{\max} > 140 km/h$

And T_{th} is the theoretical load of the track, given by 3.2 (Profillidis, 2000):

$$T_{th} = T_p + k_{fr} \times T_{fr} + k_{tr} \times T_{tr} \qquad (3.2)$$

Where T_p is the daily passenger vehicle traffic

T_{fr} is the daily freight vehicle traffic

T_{tr} is the daily traction engine traffic

k_{fr}=1.15

k_{tr}=1.40

As an addition, tonnage is also the parameter used to calculate the failure rate of the infrastructure components, for it is considered to be an adequate measure to the fatigue to which the railway infrastructure is subject.

3.2.2 Flat bottom rail cross-sections

Flat bottom rails can differ in section, that is, their geometrical characteristics. This rail section always consists of head, web and base (also called foot).

An effort of standardization of rails by the International Railways Union (UIC) led to the creation of the UIC rail cross-sections. They are always named after their weight “m” per length unit. As heavier types of rail were implemented, an increase in inertia was sought for in such a way that the ratio I/m would increase faster than “m”. As a result, rail profile height grows with “m”. A list of common rail cross-sections is presented in Table 3.2.

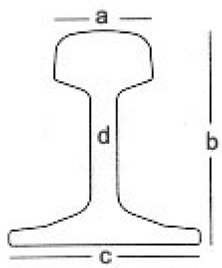
Rail types are used according to the level of traffic load and the time gaps between renewal sessions. For instance, according to Profillidis, UIC 60 and 71 are used for medium and heavy traffic load, which normally corresponds to the high speed case. In conventional railway lines, the choice of the rail section was usually made according to the traffic load as shown in Table 3.1.

Table 3.1 Choice of rail profile according to traffic load. (Source: Profillidis, 2000)

Daily line traffic load (t)	< 25000 t	25000 to 35000 t	>35000 t
Minimum required weight m of the rail per metre length	50 kg/m	50 kg/m for timber sleepers, 60 kg/m for concrete sleepers	60 kg/m

In Europe, for conventional lines with speeds over 160km/h and maximum axle loads of 22,5tonne, as well as for all high-speed lines, Rail profile UIC 60 E1 is the profile used, with monoblock or twin-block concrete sleepers. For conventional lines with lower speeds, UIC54 is usually recommended.

Table 3.2 Rail profiles in commercial use in the UK (BS), Germany (S) and other countries (UIC). (Adapted from Profillidis, 2000 and H.J. Skelton Canada, Ld.)

Section	Weight *m* [Kg/m]	*a* Head [mm]	*b* Height [mm]	*c* Base [mm]	*d* Web [mm]	Shape
BS60A	30.618	57.15	114.3	109.54	11.11	
BS75R	37.041	61.91	128.59	122.24	13.1	
BS80A	39.761	63.5	133.35	117.47	13.1	
BS800	39.781	63.5	127	127	13.89	
BS113A	56.398	69.85	158.75	139.7	20	
S41-R10	41.38	67	138	125	12	
S41-R14	40.95	67	138	125	12	
S49	49.3	67	149	125	14	
UIC50	50,18	70	152	125	15	
UIC54	54.43	70	159	140	16	
UIC54E	53.81	67	161	125	16	
UIC60	60.34	72	172	150	16.5	
UIC60E1	60.21	72	172	150	16.5	
UIC71	71.19	74	186	160	18	

3.2.3 Rail steel grades

Ultimate tensile strengths vary between 70 to 120 kg/mm^2 according to the steel grade used in rail profiles (Profillidis, 2000). Higher ultimate tensile strengths mean there is a larger resistance in terms of stress but may cause brittle (sudden) failure. Ductility, wear resistance and ease of welding are characteristics sought for in rails, and these can be modelled by adjusting the percentage of carbon, manganese and chromium.

An older reference used for the choice of steel grades was Codex UIC 860 V, 8th edition, dated from 1986 (Vitez, Oruč, Krumes, Kladarič, 2007; Vitez, Krumes, Vitez, 2005). Nowadays, ISO 9002 quality standards in manufacturing and acceptance of Vignole-type rails above 46 kg/m are applied through European Norm EN13674 Part 1. This norm specifies seven different steel grades, which are presented in Table 3.3.

Table 3.3 Steel grades described in European Norm EN13674.

Steel grade	R200	R220	R260	R260 Mn	R320 Cr	R350 HT	R350 LHT
Minimum tensile strength [MPa]	680	770	880	880	1080	1175	1175

Norm EN13674 Part 1 approaches the subjects of qualifying tests and acceptance tests. The steel grade naming is based on the lower limit of their hardness range (in HBW units). Other parameters for each grade are defined, like for instance the quality parameter of fracture toughness K_{Ic}, measured in $MPa.m^{1/2}$. Typical values of relevant acceptance tests are also presented. The tests included in the norm ensure quality when carried on rail profiles also prescribed in the norm. EN13674 contains 21 different profiles with linear masses ranging from 46 kg/m to 60 kg/m. The corresponding profiles are designated 49=49E1 and UIC 60=60 E1.

3.2.4 Welding methods

In high speed railway track infrastructures and high-performance conventional tracks, no discontinuities should exist on the track and therefore rail joint components such as fishplates are not used. This means that in these types of infrastructures, welding methods are currently in use. These eliminate discontinuities between standard lengths of manufactured rail, but at the same time they make the possibility of accommodation of expansion displacements impossible. This problem exists especially due to the thermal expansion of steel, which is rather high ($\alpha=10^{-5}$). Tracks on which welding methods are applied are known as continuous welded rail (CWR) tracks.

A list of common welding methods is presented below.

1. Electric flash-butt welding

This method can be applied either on a rail depot with a stationary welding machine or with a mobile welding machine directly on track. Fatigue strength obtained by both methods is similar, but not always good geometrical properties are achieved. The method consists in applying electric voltage to rail ends so that they are heated up to forging temperature and upset under high pressure. The resulting weld collar is then stripped, and rails are aligned and ground. Flash butt welds are in terms of quality better than thermit welds (Esveld, 2001).

2. Thermit welding

The thermit weld method provides a good metallurgical weld which can be made directly on track. After a pre heating of the rail ends with propane burners until a temperature of around 900ºC, aluminium powder and iron oxide are mixed and converted to alumina and steel at high temperatures. The conversion reaction releases high amounts of heat, causing temperature to rise to about 2500ºC, creating the weld. As in the flash butt weld, the resulting weld collar is then stripped and the rail ends are ground (Esveld, 2001).

3. Electric arc welding

This is a poor quality weld which is basically done after pre heating rail ends with propane burners to a temperature of around 250ºC. The weld is then made from the rail foot using a copper casing, which is removed in the end of the procedure (Esveld, 2001).

Other methods are also used. The reader should be referred to Manly, 2006; Weman, 2005; AWS, 2004; Dowson, 1999 and Mills, 1956, for a deeper insight on these:

4. Oxyacetylene (autogenous) welding – a method which burns a pressurized combination of oxygen and acetylene at very high temperatures. It is the only welding method which is able to melt all commercial metals. As a disadvantage it creates very rough welds.

5. Pressurised welding – in this method pressure is applied to the working pieces causing friction which heats up the working pieces, leading to forging.

6. Induction welding – In this method an electromagnetic field is generated by an induction coil. In the case of an electrically conductive material such as steel, the heating effect is caused by electrical resistance.

7. Resurfacing welding – Oxyacetylene flame is combined with melt rods of ferrous alloy or of bronze to recreate the surface of old or damaged parts.

3.3 Rail defects

As an important pre-emptive note it must be said that *rail defects* are caused by initial imperfections. These might be due to manufacturing errors which cause initial microscopic imperfections, or imperfections caused by external damage to rails – for instance ballast imprints. The initial imperfections then grow larger in time due to the passing of trains, i.e. they are rolling contact fatigue (RCF) phenomena. This means rail defects include internal discontinuities which may grow larger due to fatigue and rail alterations of mechanical nature occurring under the influence of loading cycles imposed by the passing of rolling stock, and especially locomotives (due to their usually higher axle load). The problem of *track defects* is treated separately.

Track defects are defined as deviations of actual from theoretical values defined for the track's geometrical characteristics (Profillidis, 2000). Track defects are exclusively due to train traffic, macroscopic and geometrical in nature and are generally reversible if correct track maintenance procedures are applied.

Track and rail defects and their speed of propagation are influenced by parameters such as (Vitez, Oruč, Krumes, Kladarič, 2007):

- Radius of curves. Defects which appear in low radii curves where trains pass at high speeds are: lateral wear (UIC 2203), corrugation of the lower rails of curves (UIC 2201, 2202), head-checking (UIC 2223), and other rail contact fatigue (RCF) phenomena.
- Daily or annual mega gross tonnage (mgt) in the studied stretch.
- The transition zone between falling and rising gradients (of around 20 ‰).
- Train speed and cant on curves. Different ranges of traffic types (pulling or pushing with cant) cause lateral wear on the high rail. Excess cant may cause crushing on the low rail.
- Axle load of trains. High axle loads may also cause crushing of the low rail.
- Some rolling stock types and wheel radii may exert higher forces on rails, especially on curves. Also, wheel-rail tensions may be larger when wheels of smaller radii are used.
- Lubrication and correctly scheduled maintenance actions such as grinding may help in combating wear and rail contact fatigue phenomena.

3.3.1 UIC coding system for the classification of rail defects

A list of significant rail defects is presented in Chapter 3.3.2. These defects were studied and classified (numbered) by the International Railways Union. The numbering follows the codification presented on Figure 3.2.

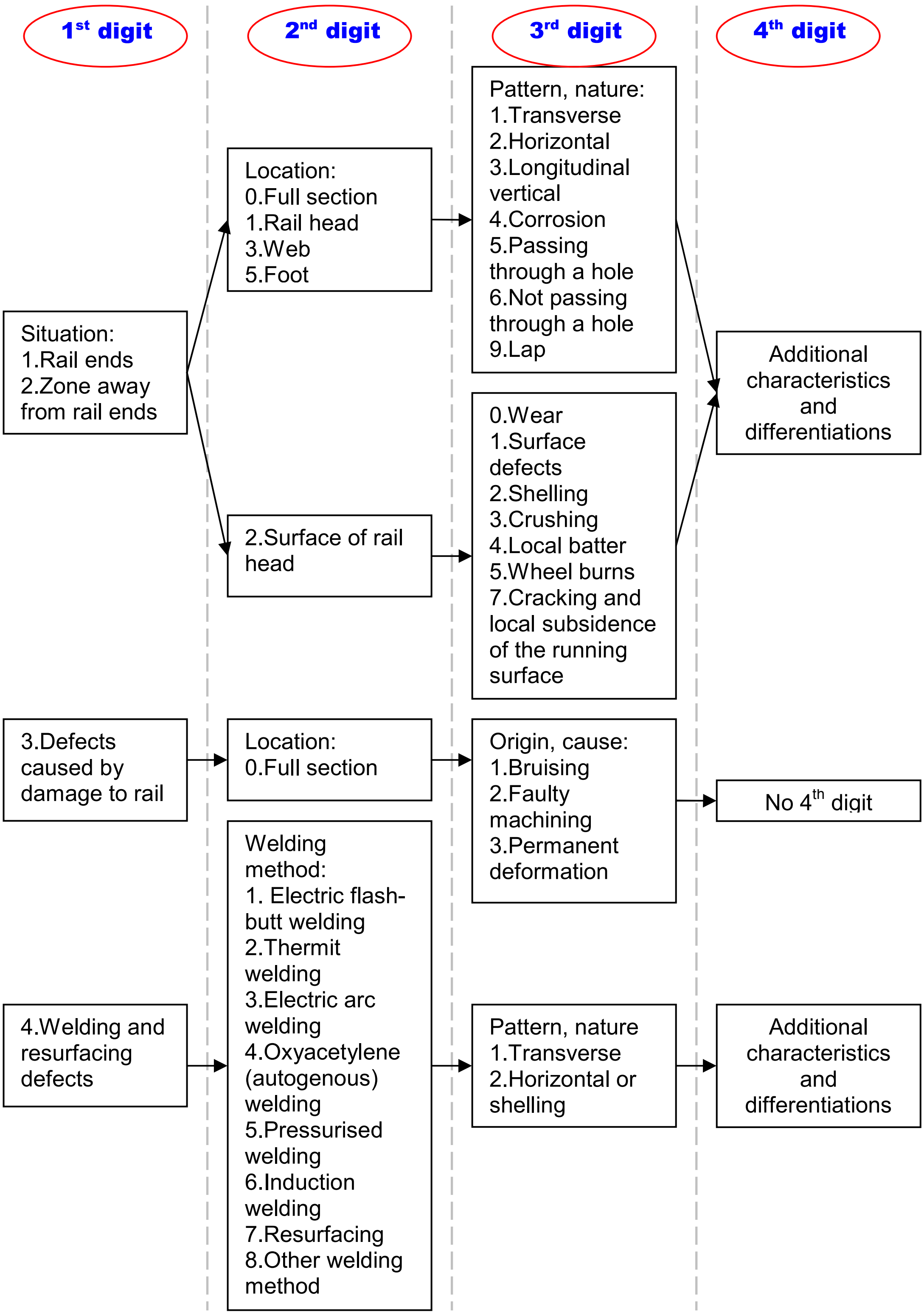


Figure 3.2 UIC rail defect codification system. (Source: Vitez, Oruč, Krumes, Kladarič, 2007)

3.3.2 Description of significant rail defects

The following is a list containing the most common rail defects:

- *Tache ovale (Rail defect UIC 211).* An initial internal oval discontinuity appears in the manufacturing stage inside the rail head, some 10 to 15 mm below the surface. With repeated loading it reaches the rail surface and causes its immediate failure. May be extensively present in stretches if rails of the same manufacture are used. It is caused by accumulation of hydrogen during manufacturing or poor welding of the rails. It is traceable with ultrasonic equipment or visual inspection. See Figure 3.3.

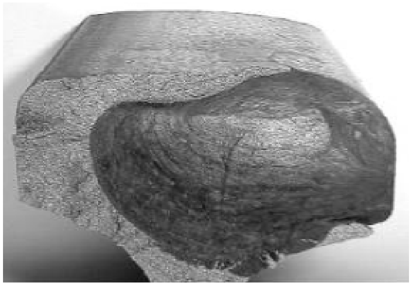


Figure 3.3 Tache ovale. (Source: Khumar, 2007)

- *Horizontal cracking (Rail defect UIC 212 or 412).* Horizontal cracks appear on the side of the rail head and may cause peeling of the rolling surface. Traceable visually or with the aid of ultrasonic equipment. In the case of the defect in a flash butt weld section (code 412), the fracture assumes a curved shape in the weld, which may spread up or downwards leading to fracture of the rail (Profillidis, 2000).

- *Short-pitch corrugations (Rail defect UIC 2201).* These are corrugations of wavelength between approximately 3 to 8 centimetres. They are particularly dangerous for they cause high frequency oscillation of the track which may lead to resonance and high rail stresses, concrete sleeper fatigue together with cracking at the rail support, loosening of fastenings, higher rate of wear at insulators and clips, premature failure of ballast or subgrade and a substantial increase in noise level. The defect may be detected by rail defect recording equipment or visual inspection. It may be repairable by smoothing of the rail with the passage of special equipment. See Figure 3.4.

Figure 3.4 Short pitch corrugations. (Source: http://www.corrugation.eu/images/site/track004.jpg)

- *Long-pitch corrugations / waves (Rail defect UIC 2202).* These corrugations have a wavelength of approximately 8 to 30 cm. They mainly occur on the low rails of curves with radii below 600 m and are therefore more common in underground or suburban rail lines which carry large amounts of traffic. Detection and repair are similar to the short pitch corrugations.

- *Longitudinal vertical cracking (Rail defects UIC 113, 133, 213 and 233).* Vertical cracks appear on the rail head (113 and 213) or web (133 and 233), which upon expanding may split the rail head or web in two. May be detected by ultrasonic equipment. Rails with this type of defect should be replaced immediately. When cracking occurs on the rail web only, this defect is termed “piping” (Figure 3.5).

Figure 3.5 Cross sectional schematic view of a rail where piping occurred.
(Source: U.S. Dept. of the Army, 1991)

- *Lateral (gauge corner) wear (Rail defect UIC 2203).* This type of wear is caused by snaking – the lateral movement of the trains due to the conic shape of the wheels. After a certain amount of wear, rails must be replaced for the rail gauge is affected. See Figure 3.6.

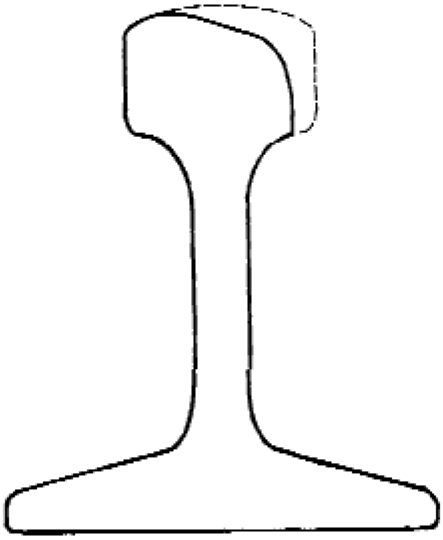


Figure 3.6 Cross sectional schematic view of gauge corner wear.
(Source: U.S. Dept. of the Army, 1991)

- *Rolling surface disintegration (Rail defect UIC 221).* The name of the defect corresponds to its description. The disintegration process happens gradually. Affected sections must be replaced.

- *Shelling of the running surface (Rail defect UIC 2221).* Deformations on irregular patterns on the rail surface appear prior to the formation of shells which span through several millimetres of depth in the metal. This defect always happens over large spans of track. It is detectable by visual inspection or with the aid of ultra sonic equipment. See Figure 3.7.

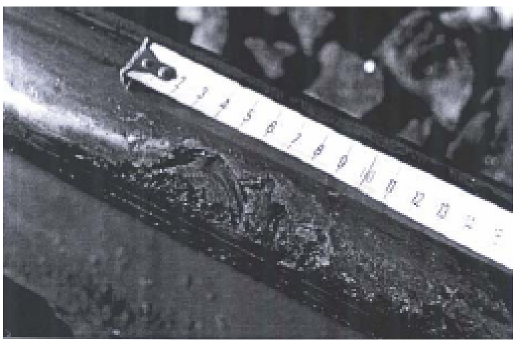


Figure 3.7 Shelling: a “shell-like” fracture is observed. (Source: Khumar, 2007)

- *Gauge-corner shelling (Rail defect UIC 2222).* In a first stage, dark spots appear along the gauge corner of the rail head. This corresponds to disintegration of layers of metal just below the surface of the rail which eventually evolve into cracking and shelling of the gauge corner itself along large stretches. This defect appears normally on the outside rails of curves lubricated to avoid lateral wear. This defect may be eliminated in an early stage by grinding.[8]

- *Head checking (Rail defect UIC 2223).* Head checking appears due to the high stresses in the rail gauge corners in curves. It appears in curves of radii < than 1500m (Khumar, 2007). Head checks display angles of 15-60º in relation to the longitudinal axis of the rail, with a spacing of 2-7 mm, which eventually can reach 0.5 mm in rails with hardened heads (Krull, Luke, Thomas, Pohl, Rühe, 2002). Grinding may be used to correct this defect. See Figure 3.8.

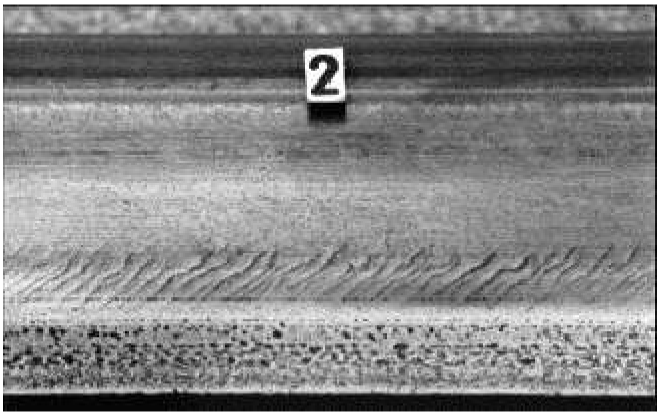


Figure 3.8 Head checks. (Source: http://www.ndt.net/article/v07n06/thomas/thomas.htm)

- *Evolution of head checking: Spalling (Rail defect UIC 2223).* This defect is characterized by the loosening of small chips of steel, caused by the intersection of cracks formed at the rail surface. It appears in a late stage of crack development, and is more frequent in cold climates (Khumar, 2007). See Figure 3.9.

Figure 3.9 Spalling of the rail surface. (Source: *www.arrt-inc.com/jpg/spalling.jpg*)

- *Squats (Rail defect UIC 227)*. Sometimes a two crack development process occurs, where there is a leading crack and a trailing crack in the crown (surface) area of the rail. The leading

[8] The initiation of shelling defects is helped by the presence of traces of oxide inclusions in the steel and also by residual stresses during the rolling of the rail profile, so the control of the manufacturing process is critical. The cracks formed during shelling processes may evolve downwards and cause a transverse cracking of the rail (Khumar, 2007).

crack propagates in the direction of train traffic, while the trailing track propagates in the opposite direction, at a faster rate. Sometimes the trailing track may propagate downwards into the rail web. Horizontal cracking leads to detachment of the rail surface from the rest, creating a depression on the rail surface.

- *Plastic flow and tongue lipping.* Plastic flow occurs in the outside part of the rail head of inner (lower) rails of curves. It happens due to overloading. Tongue lipping is a form of plastic flow which is triggered by initial cracks on the surface. Plastic flow is able to create a plastic deformation on the rail with a deflection of about 10-15 mm. It can be prevented by grinding. See Figure 3.10.

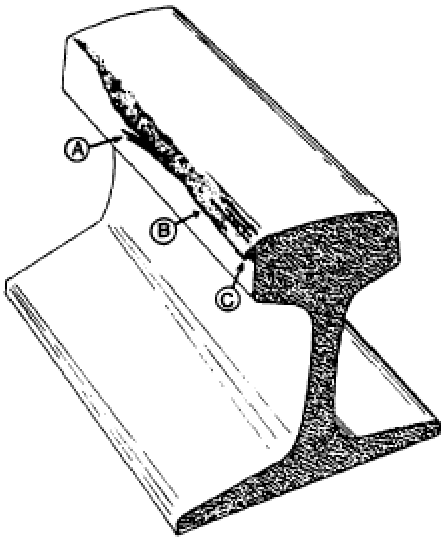


Figure 3.10 Schematic representation of plastic flow. **A:** Flow of the surface metal to the field side of the rail. **B:** a lip of metal is formed along the field surface corner. **C:** In advanced stages, the lip becomes blade-liked and jagged. (Source: U.S. Dept. of the Army, 1991)

- *Isolated / single (UIC 2251) and repeated (UIC 2252) wheel burns*. Damaged areas on the rail running surface may appear due to the slipping of wheels. Repeated wheel burns appear in sections where trains usually break or are subject to high traction forces. Wheel burns develop cracks from the surface into the rail head, which may then evolve upwards or downwards. This may either cause immediate failure or enhance the risk of brittle failure, especially in cold climates. See Figure 3.11.

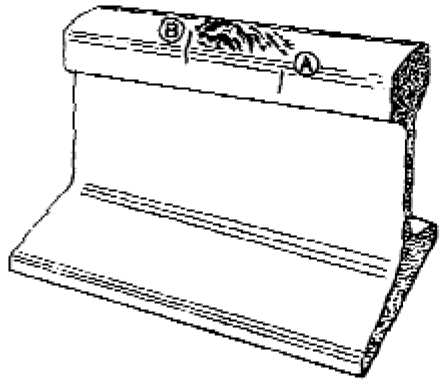


Figure 3.11 Wheel burn. **A:** Hairline crack in the vicinity of the burn. **B:** Thermal crack produced by the burn, extending down to the gauge corner of the rail. (Source: U.S. Dept. of the Army, 1991)

- *Transverse cracking (Rail defect UIC 111, 211 or 411)*. On rail ends, transverse cracks appear usually in defective flash butt weld sections (411), i.e. sections which are misaligned, have pores or inclusions (Khumar, 2007). Away from rail ends (211), transverse cracks develop from zones of wear, corrosion, or manufacture anomalies (Cannon, 2003). Normally the cracks are initiated in the rail head or foot and develop due to fatigue. Transverse cracking may be triggered by *tache ovale* as well, leading to a kidney shaped transverse fracture. See Figure 3.12.

Figure 3.12 Profile with full transverse crack. (Source: Transportation Safety Bureau Canada)

- *Buckling*. On continuous welded rails (CWR) there is no joint room to accommodate the phenomenon of thermal expansion. Stress builds up with the rising of temperature and can lead to lateral buckling. The desired lateral buckling strength of the rail profiles should therefore be cautiously calculated taking in consideration the thermal gradients, geometrical characteristics of the profile and sleeper or other lateral constraint spacing.

- *Transverse cracking of profile in Thermit weld sections (Rail defect UIC 421)*. Cracks are located in normal cross sections near the weld.

3.4 Rail maintenance

Maintenance is necessary due to wear of components with the passing of trains. Generally speaking, maintenance may be necessary to improve ride comfort of passengers or to improve safety levels (risk of derailment, etc.). There is a complementarity between both, for safety related maintenance actions must be preventive, while comfort related actions are normally corrective. There is also a distinction between maintenance related to rail defects and to track defects. The latter may be from 5 up to 15 times more frequent than maintenance related to track defects. Normally in lines with average traffic loads, geometrical characteristics' restoration happens after loads of 40 to 50 million tonnes, while rail replacements or grinding actions occur after about 500 to 600 million tonnes (Profillidis, 2000).

This explains why there are more maintenance operations destined to correct track geometry. These basically consist of reshaping of the ballast layer (tamping, ballast regulation) and its stabilization (compacting of ballast, addition of chemical binders). Ballast may also experience problems of weed intrusion, which are solved through ballast cleaning operations (Profillidis, 2000; Esveld, 2001).

Rail centred maintenance actions consist barely of rail grinding and re-profiling (correction of rolling contact fatigue defects which exist on the rolling surface), and weld straightening. Some rail defects require, after a certain degree of development, the full replacement of the rail lengths affected (Profillidis, 2000; Esveld, 2001).

3.4.1 Visual inspection and ultrasonic inspection

Visual inspection is the traditional method regarding track inspection. It provides the first overall picture of the track constituents' condition, and of the track geometry itself. Any defects which are not caused by internal discontinuities in the rail can be spotted and actions can be taken to correct them. Some modern tracks are already equipped with video or photo imaging inspection systems (Esveld, 2001).

Ultrasonic inspection allows maintenance personnel to assess the internal structure of rails. The purpose of the method is to detect defects caused by internal discontinuities in rails which are not visible from the outside. This is important mainly due to security reasons, for some internal discontinuities may cause brittle failure. Comfort is mainly disturbed by defects which are visible from the outside. Ultrasonic inspection also allows the detection of failures in an early stage, which in turn allows the timely scheduling of maintenance operations, avoiding disruption of regular train traffic.

Ultrasonic inspection may be carried out by hand equipment or with the use of equipment mounted on special ultrasonic trains, which may reach an inspection speed of 100 km/h. Ultrasonic trains are used over large stretches of rail networks. Handheld equipment (Figure 3.13) is mainly used on special parts like switches, joints and bridges, among others. It is also used to double check sections indicated by the ultrasonic trains, in order to detail measurements and the actions to take against the detected defect.

Figure 3.13 Ultrasonic handheld equipment. From left to right: Eurailscout SPG 2 and MT95 ultrasonic handheld equipments, Geismar SIRIUS ultrasonic handheld equipment.
(Sources: www.eurailscout.com, www.geismar.com)

One example of an ultrasonic train is the EURAILSCOUT UST96 (Figure 3.14), which is able to perform measurements at a maximum speed of 100km/h (speed depends on rail conditions). The precision of measurements is very high. More than 95% of the reported defects are confirmed to be actual defects (Esveld, 2001). The measuring car is provided with D-GPS satellite navigation and an odometer with magnetic sensors, which allow to locate the position of reported defects with an accuracy of around 1 meter. The train is equipped with 2 blocks of ultrasonic probes per rail. These blocks are made up of 8 probes which are aimed in different angles, allowing inspection of almost all parts of the rail. Water is used as sonic contact fluid. The in-rail depth of defects is obtained with an accuracy of about 1 mm (Esveld, 2001).

Figure 3.14 Eurailscout UST96 ultrasonic train. (Source: www.flickr.com – Leen Dortwegt)

3.4.2 Rail grinding

Rail grinding allows the correction of the rail profile in the case of rolling contact fatigue (RCF) defects which arise on the rail surface. The purpose of rail grinding is to eliminate surface-initiated cracks which are in an initial stage. It is especially indicated in the case of corrugations, and is also able to correct or prevent cases of wheel burns, shelling, head checks, spalling, plastic flow and tongue lipping.

Two types of machines are used for the operation of rail grinding: machines equipped with rotating stones and stones oscillating longitudinally. The movement of the stones against the rail abrades and smoothes out the rail surface. Rail grinding machines may be equipped with equipment which performs the STRAIT weld grinding operation (i.e., the Plasser GWM 250). The current grinding technique uses pivoting units, which allow the re-profiling of the rail surface in a polygonal shape. This shape quickly merges into a continuous curved profile (Esveld, 2001). See Figure 3.15. Manual operation is also possible, using portable multi-purpose rail grinding machines.

Figure 3.15 Rail grinding equipment. **A:** TMS rail grinding unit. **B:** Fairmont rail grinding machine. (Sources: www.tms-rail.com, www.lrta.gov.ph)

3.4.3 Rail reprofiling

When rail grinding is insufficient to correct highly developed defects, rail reprofiling is used. The principle is to use a heavy machine equipped with a planing cutter mounted on a system of quick release fastenings and cooled by high pressure jets, which allows a full reprofiling of the running surfaces and edges. In the Plasser & Theurer rail planing machines, cuts are made in 4 different angles and radially, and are adjusted by guiding rollers which ensure that a perfect longitudinal movement is always achieved. The reprofiling operation takes several passes (Esveld, 2001).

3.4.4 Weld correcting

One possible way to correct weld defects is to use the STRAIT (Straightening of Rail welds by Automated Iteration) principle, which uses two displacement transducers to measure the step between rail ends on the weld. A displacement tolerance interval is set with a 0.2mm lower bound and a 0.5mm higher bound. STRAIT measures the difference in displacement to the higher bound of the tolerance interval, which is the maximum overlift of the weld. The section is then bent iteratively in high tension (yield), until weld displacement exceeds the lower bound limit of the tolerance interval, converging normally in 3 to 4 iterations.

The process is normally accompanied of tamping of the sleepers around the weld area. Overlift weld material is ground afterwards. Hence, STRAIT machines are usually mounted on tamping cars or grinding cars (like the Plasser GWM) (Esveld, 2001).

3.4.5 Rail replacement

Replacement of rail may be done individually if manual techniques are used. Full mechanized track renewal cars also exist, which are able to renew a part of or all track elements.

The manual method consists basically in:

- arrival of the train carrying new rails, followed of unloading (normally at night);
- loosening of fastenings;
- cutting and moving of the old rail to the inside of the track;
- positioning of new rails;
- welding of the rails;
- weld straightening and lengthening of rails with hydraulic tensioning equipment. This is made so the rails reach a length which corresponds to a 25ºC temperature, which is considered to be neutral (rail renewal is normally done during night periods);
- fastening of rails;

- grinding of weld overlift;
- cutting of old rails. This is made in order to transport the old rail sections, and can be done during the day time. The old rails, fasteners and sleepers may be removed at night by the special hauling trains which had previously brought the new rails.

The mechanized processes involve machines which are long enough to operate on the newly laid track, making track renewal a continuous process (Esveld, 2001).

3.4.6 Rail lubrication

The primary goal in rail lubrication is the reduction of wear, especially in the gauge face corner of curves with low radii. It also reduces railway locomotive fuel consumption and noise levels. Lubrication may however cause RCF defects to develop due to the presence of fluids (fluid entrapment), so it should be used cautiously. Another problem which may arise is the migration of lubricant material from the gauge face corner to the rail head surface, leading to wheel slipping – this may stimulate the appearance of head checks. Several lubrication materials are available, such as grease or oil based, solid lubricants, or combinations of several types (Dowson, 1999). These may be applied by on-bogie mounted systems, or more frequently, by track-side permanent systems with automatic grease dispensing units.

Chapter 4: Application of RAMS techniques to the rail track infrastructure

4.1 Introduction

The prediction of failures in a railway track infrastructure is essential to the estimation of its life cycle costs. If correctly applied, RAMS techniques may, from a sufficient amount of failure data, predict the number and distribution in time of failures on the rail infrastructure, which in turn provide an estimate of availability. Availability will dictate service capacity and track maintenance operations, therefore leading to cost assessment.

The main question asked in the scope of this thesis is:

- How can RAMS techniques be applied to the rail in order to forecast failures and their associated costs?

The special case asked to be dealt with is:

- How can RAMS techniques be applied to the rail in order to forecast failures and their associated costs when failure data sets are short?

As stated in chapter 2.3, different RAMS parameters may be obtained from others, using the relationships stated. From failure rate, it is possible to obtain MTBF and reliability. From the time between maintenance operations T, it is possible to obtain MDT, availability and to define MTTR.

There are therefore 2 items to study, which answer the main question:

1. **Prediction of failures and reliability.** This item shall be discussed in chapter 4.2.

2. **Prediction of availability and MDT.** This item shall be discussed in chapter 4.3.

4.2 Failure prediction

Reliability is obtained from failure rate, as described in chapter 2.3.2. The latter is obtained by employing the statistical methods described in chapter 2.5. For this to happen, failures must be recorded using the recommendation of railway and RAMS authors. In most of nowadays' databases, failure data sets are scarce or they have a short background. A suggestion of collection procedures and the statistical distributions to use on failure data will be presented. The whole process is summed up in the chart presented in Figure 4.1.

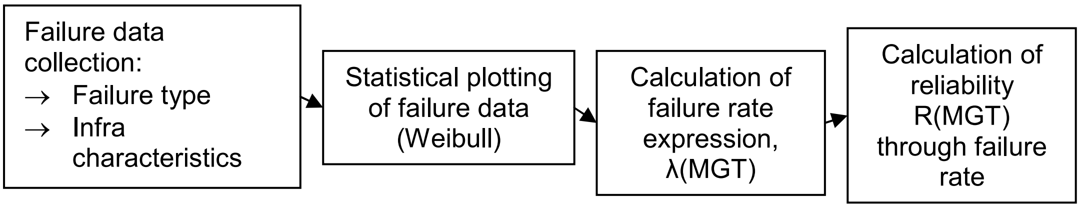


Figure 4.1 Reliability and failure rate forecasting procedure.

These steps are included in the failure data compilation step of the RAMS cycle (Step 1 – Chapter 2.2.1).

4.2.1 Failure data collection

Firstly, the most important step is to contractually define failure (chapter 2.2.3). Different failure modes are associated to wear, cracks or other type of damage, and each rail infrastructure maintenance manager should have the limits for this damage defined. The suggested way to do this is to define, for each failure mode, limits for crack number, size and opening (in mm, for example).

As a probabilistic plotting of failures is the desired goal, the failure data collection should clearly separate each failure mode (Chapter 2.5.7; Chapter 2.7). In railway engineering, plotting is usually done indicating the MGT (tonnage) to failure. Furthermore, to provide the data sets with statistical meaning, plotting should be made in track units of the same characteristics, i.e. failure data should be grouped into comparable sets. According to Esveld (2001), the recording of failures should include:

- Classification of failure – Chapter 3.3;
- Cumulative tonnage to failure – see below;
- Time to failure;
- Method of repair (action taken);
- Welding process at repair;
- Post-weld treatment;
- Time to repair;
- Failure location (Infra data, specific location on the rail network).

Cumulative tonnage, or cumulative traffic load, normally measured in MGT, is the main unit with which failure rate grows, for most of the failure modes are due to mechanical fatigue and cracking. It is also for this reason that it is crucial to correctly classify failures, for different failure modes are subject to different crack models. The method to calculate traffic load was presented in chapter 3.2.1.

Esveld lacks specifications on infra data. Since failures occur in different parts of the rail network, these may only be studied together if the track lengths in which they occur are comparable. For this reason, criteria on infra and type of traffic should be provided when failures are recorded.
In addition, the guidelines presented on chapters 2.7 and 3.3 were taken in consideration, especially in terms of train traffic. The recommendation of RAMS authors, to use engineering judgement over statistical judgement (chapter 2.5.6) should not be overlooked, and therefore some characteristics were introduced while having in mind the characteristics of different failure modes.

The criteria suggested to identify comparable track stretches (excluding switches and crossings) are:

- **Type and quality of rail:** Rail profile height and shape (dimensions) and weight. Type of steel used (chapters 3.2.2 and 3.2.3). Stronger steels and heavier, larger rails with greater inertia and cross-sectional area should endure more stress and be more resilient. Track gauge should also be recorded.

- **Type and quality of fastenings:** In the case of high-speed rail, the type is always elastic. Models may differ (chapter 3.1), which can lead to different failure distributions with MGT.

- **Type and quality of sleepers:** In the case of high-speed rail, they are usually made of concrete, either mono-block pre-stressed or twin-block reinforced sleepers (chapter 3.1). Spacing and sleeper weight should be recorded.

- **Ballast or slab technical characteristics:** Ballast or slab thickness should be recorded (chapter 3.1).

- **Subgrade quality:** Geotechnical and foundation aspects have a large influence on the speed of the degradation of track geometrical characteristics. The latter may increase the development of RCF defects such as head checking, corrugations, etc.

- **Curve radius and super-elevation:** It is important especially in the case of low radii where there is a tendency for the development of head checks and low pitch corrugations. Curves with high super-elevation may also experience larger stresses in the low rail when trains pass at lower speed than design speed, and on the high rail for the opposite case (Profillidis, 2000).

- **Track gradient:** More traction force and therefore more wheel-rail friction are required to overcome uphill gradients. Lubrication must be reduced in order to avoid wheel slipping. These lead to more wear than in downhill gradients (Khumar, 2007).

- **Special stretch of track (when applicable: bridge, tunnel):** Track characteristics may change dramatically or special considerations should be taken (like joints or rail gaps in transition zones, different support conditions, etc.) These stretches should be accounted in separate data (chapter 4.2.2).

- **Installation date of the components:** A part of the components' deterioration comes from weather or other age related-phenomena (i.e., corrosion, degradation of concrete) (Profillidis, 2000; Esveld, 2001). This also allows the study of the mean life of the different components.
- **Traffic speed and direction:** Speed increases dynamic loads, especially contact forces in curves. The direction of traffic is also crucial in defects such as corrugations (they usually propagate along traffic direction - Vitez, Oruč, Krumes, Kladarič, 2007);

- **Axle load of trains / locomotives:** Rail defects evolve faster when exposed to higher loads for they are mostly caused by crack propagation processes (Khumar, 2007);

- **Maintenance history on the analyzed stretch (operations performed and scheduling):** This item should include inspection frequency, rail grinding frequency, ballast cleaning and tamping frequency (Profillidis, 2000; Esveld, 2001), rail lubrication process. Number of replacements may be accounted to understand the promptness of the track stretch to the development of defects;

Information should be collected regarding manufacturers and manufacturing dates for each of the abovementioned components. Switches and crossings should be accounted in different data, for they are inherently different from regular track, being subject to different actions from the passing of trains.

As a guideline, an **infrastructure and traffic form** and a **failure identification form** destined to field use by the reliability engineer are presented in Annex E. These forms should be filled in by members of staff which had relevant RAMS training, who are familiar with the rail infrastructure and who are competent in identifying failure modes (chapter 2.7).

4.2.2 Statistical plotting of failure data

As stated above, failure plotting should be based on data sets of:

- → The same failure mode, and;
- → Comparable track stretches (that is, the technical characteristics of the infrastructure and train traffic collected previously fall in the same range).

The most interesting way to present data is per failure mode, so that the impact of each rail defect in the network may be assessed. It is assumed that the presence of one type of failure in a stretch or vicinity of a rail doesn't influence the development of another failure. This allows to state that data sets are independent, meaning statistical distributions may be applied on their study.

Lengths of track from different parts of the rail network which are similar can be grouped. The criteria shouldn't be too demanding in order to avoid the over-splitting of rail lengths, which would present very low numbers of failures. Therefore when studies are carried the characteristics should be grouped in intervals. These intervals for infrastructure characteristics are to be defined by the railway infrastructure manager. This falls out of the scope of this thesis and should be subject of further discussion. In any case, a suggestion of intervals to be used is presented in Annex E.

In the study of rail, failure rates usually grow with MGT, for rail defects are phenomena of a mechanical nature. If the reliability engineer doesn't know the MGT to failure at the time of recording, an estimate of tonnage may be obtained from train traffic information up to the date/time of failure.

Since the phenomena are of mechanical nature, it is implied that failure rate should grow with time. Generally speaking, rail is a highly reliable component (Profillidis, 2000), and therefore a low number of failures is expected (even if the definition of failure changes from rail break to a slight rail defect). The reader is referred back to chapter 2.5, where it is stated that the Weibull distribution is the one which suits best this type of data sets.

Other distributions are presented in chapter 2.5, and these may be applied to the data sets to verify the adjustment level, though in a first approach, if engineering judgement is considered, Weibull should provide the best results. The method quoted in chapter 2.5.6 could have also been applied, but research lacks on the value of α for the specific case of the rail infrastructure. This should be subject of further studies and falls out of the scope of this thesis.

4.2.3 Calculation of failure rate expression, λ(MGT), and reliability, R(MGT)

Failure rates and reliability are obtained using the expressions of the Weibull distribution. For this to happen, a minimum of 6 recorded failures of the same mode in a comparable stretch of track is the recommended by RAMS authors for the plotting. The process is summed up in:

→ Plotting of failures and cumulative MGT in log-log charts;
→ Obtaining of reliability and failure expressions.

This can be done inexpensively in an Excel spreadsheet. The procedure presented here is widely based on the method presented by Dorner (1999). Its first step is to obtain the Weibull expression of reliability and failure rate is to create a table with the MGTs to failure recorded in the "comparable stretch of track". An example of a stretch of track with 10 recorded failures is presented on Table 4.1, where failure is defined as "gauge corner wear > 2 mm":

Table 4.1 Example of (fictitious) failure data.

Example: Track stretch "A" (20.3 Km long) – MGT to "gauge corner wear > 2 mm"										
MGT	473	688	591	532	812	654	513	709	620	454

The data should be sorted in a column, in ascending MGT order, while adding a rank column (integers from 1 to 10). An estimate of the proportion of the population which shall display a failure at the number of the recorded MGTs is needed. This may be done by adding a column of median ranks, which can be obtained by Bernard's approximation, given by expression 4.1.

$$\frac{r-0.3}{N+0.4} \tag{4.1}$$

Where r is the "rth" failure and N is the total number of the population which may exceed the last r[9]. Two more columns are needed. These are the columns which provide the plotting data. The first is a column containing formula 4.2.

[9]In rail reliability assessments, N is equal to the number of recorded failures, for normally the accounting is made only for the recorded failures.

$$\ln\left(\ln\left(\frac{1}{1-Median\ Rank}\right)\right) \tag{4.2}$$

The other column is the transformed data of cumulative tonnage to failure (MGT) (expression 4.3).

$$\ln(MGT) \tag{4.3}$$

An explanation of the need of these two columns is presented in Annex F. The result of the application of expressions 4.1, 4.2 and 4.3 to Table 4.1 is shown in table 4.2:

Table 4.2 Relevant data for Weibull plotting.

MGT to failure	Rank	Median rank	ln(ln(1/(1-Median rank)))	ln(MGT)
454	1	0,067307692	-2,663843085	6,118097198
473	2	0,163461538	-1,723263150	6,159095388
513	3	0,259615385	-1,202023115	6,240275845
532	4	0,355769231	-0,821666515	6,276643489
591	5	0,451923077	-0,508595394	6,381816017
620	6	0,548076923	-0,230365445	6,429719478
654	7	0,644230769	0,032924962	6,483107351
688	8	0,740384615	0,299032932	6,533788838
709	9	0,836538462	0,593977217	6,563855527
812	10	0,932692308	0,992688929	6,699500340

The plot obtained is shown in Figure 4.2, with a linear regression line adjusted to fit the data set:

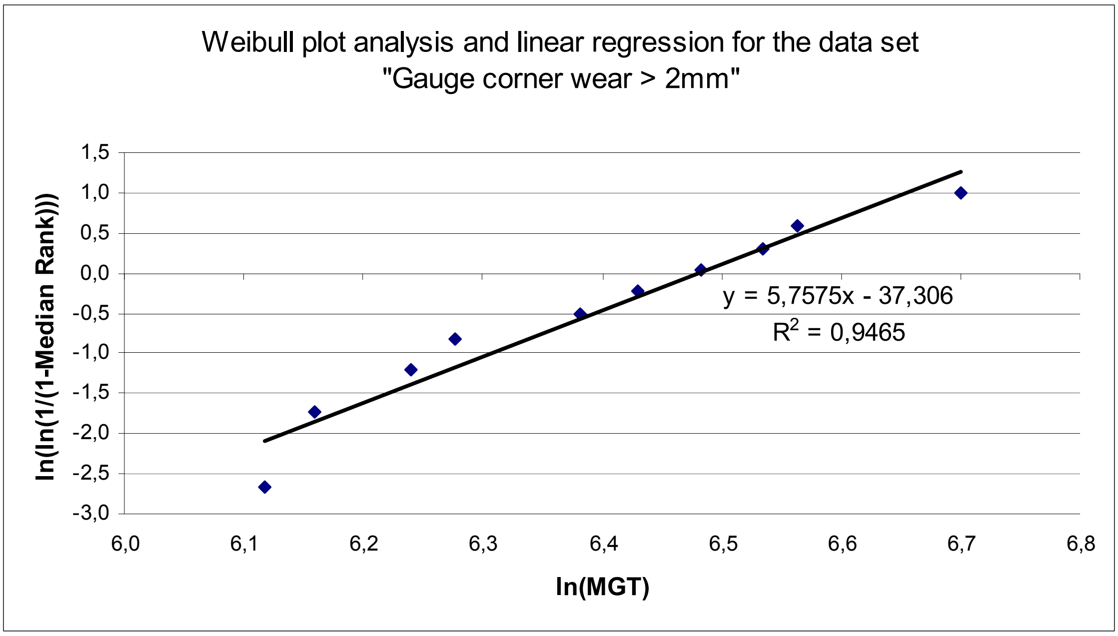


Figure 4.2 Weibull plot analysis and linear regression for the data set of Table 4.2.

Once more the Weibull expressions are presented, this time converting the time unit t to MGT and with the parameter set to null (Annex F):

$$R(MGT) = \exp\left[-\left(\frac{MGT}{\alpha}\right)^{\beta}\right] \quad and \quad \lambda(MGT) = \frac{\beta}{\alpha}\left(\frac{MGT}{\alpha}\right)^{\beta-1} \qquad (4.4; 4.5)$$

As explained in Annex F, the value of β equals the linear regression slope m:

$$\beta = m \qquad (4.6)$$

In the example, this value is of 5,7575. As for the parameter α, it is obtained by (Annex F):

$$\alpha = e^{-\left(\frac{b}{\beta}\right)} \qquad (4.7)$$

Since in the example, *b* equals -37,3056, the value of 651,6106 is returned for α. Taking expressions 4.4 and 4.5, this means that the reliability R(MGT) and failure rate λ(MGT) expressions in this example are[10]:

$$R(MGT) = \exp\left[-\left(\frac{MGT}{651{,}6106}\right)^{5{,}7575}\right] \quad and \quad \lambda(MGT) = \frac{5{,}7575}{651{,}6106}\left(\frac{MGT}{651{,}6106}\right)^{4{,}7575} \qquad (4.8; 4.9)$$

If the railway maintenance manager decides so, reliability growth corrections may be applied to the predictions. This process is explained in Chapter 2.8. In any case since RAMS databases are inexistent among most rail infrastructure maintenance managers, the value of the growth factor α may only be guessed (here, α refers to the parameter presented in equation 2.24). Studies should be carried using larger data sets of reliability parameters in order to obtain the value of the growth factor α. This falls out of the scope of this thesis – further collection and compilation of failure data should be done in order to make estimations on this parameter.

[10] For a detailed analysis of the regression, the *Analysis ToolPak Add-In* may be used. The user barely needs to checkmark it in the *Add-Ins* option of the *Tools* menu, and then it may be used on the *Tools* menu, under *Data Analysis*. Select the option *Regression* to produce the analysis. Under *Input Range Y*, the column ln(ln(1/(1-Median Rank))) should be selected, including its title. Under *Input Range X*, the column ln(MGT) should be selected, also including the title. Then the options *Labels* and *Line Fit Plots* should be checkmarked, and *New Worksheet Ply* should be selected under *Output Options*. This will create a new worksheet with the regression data. In the wider table created, the column *Coefficient* has the values of b, in the row Interception and β on the row ln(MGT). This is also a way to obtain parameters α and β.

4.2.4 Forecasting and prediction range

As a rule of thumb, predictions shouldn't be made too far into the future. This is stated both by RAMS authors (Smith, 2005; Villemeur, 1992-B) and Esveld (2001), the latter one stating that they should be extended by a period no more than 20% larger than the one which refers to the available data.

Forecasts are made by using the expressions of failure rate and reliability obtained (as the ones presented in the example, 4.8 and 4.9), inserting the cumulative MGT value predicted for the desired forecast time. Cumulative MGT is controllable and predictable to a given time if information on train traffic (scheduling and weight) is collected.

As an example, let's assume a train line has yearly train traffic of 30 MGT (Chapter 3.2.1). A conservative forecast of reliability and failure rate for year 11 is made by using 12x30 = 360 MGT. Using the expressions 4.8 and 4.9 obtained from the data set in Chapter 4.2.3, we calculate for that year:

$$R(180) = \exp\left[-\left(\frac{360}{651{,}6106}\right)^{5{,}7575}\right] = 0{,}9677$$

$$\lambda(180) = \frac{5{,}7575}{651{,}6106}\left(\frac{360}{651{,}6106}\right)^{4{,}7575} = 525{,}18\times 10^{-6}\ failures/MGT \qquad (4.10;\ 4.11)$$

4.2.5 Translation into costs

As stated on chapter 2.2, RAMS results may be used as input for LCC assessments. In this chapter, a brief insight into steps 2 and 3 (Chapters 2.2.2 and 2.2.3) is made.

Both reliability and failure rate expressions provide a first look at costs. These costs may refer to accidents or corrective / preventive maintenance actions. This depends on the definition of failures.

To illustrate this, the boundary situations are explained: if the failure of rail is defined as a rail break, when a failure happens the replacement of the rail is necessary. On the other hand, if a failure is defined as the overcoming of a limit of wear or crack opening in mm of a failure mode, then maintenance actions (e.g. grinding, re-profiling) should be scheduled to correct this limit, which is also translated to costs. Estimations on costs related to failures are the subject of life-cycle CBA, being out of the scope of this thesis. Further studies on this issue should be developed.

There was a distinction between steps 1 and 2 of the RAMS cycle (Chapters 2.2.1 and 2.2.2). The process described from Chapter 4.2.1 through Chapter 4.2.4 separates different failure modes, track infrastructure characteristics and train traffic types. This separation allows step 2 (system interaction of components) to be skimped, for each failure rate and reliability expression are defined for different systems. Step 3 is also simplified for it is easy to define costs for the failure modes defined, even if the rail is just a component. This happens due to the reason that there are maintenance actions destined to this component only. So, the reliability engineer may assess the cost of different failure modes from their reliability and failure rate expressions.

Their main uses are:

1. A Pareto analysis of failures can be carried by multiplying unreliability (probability of failure) on reference years by the costs that those failures represent. This returns an expected value of maintenance costs in the year split by failure modes. It allows the rail maintenance manager to identify the failure modes which represent more costs. Plus, a sum of all the expected values of costs on several comparable stretches provides a forecast of the total maintenance costs expected for a train line, network or a certain maintenance area (if maintenance is split by geographical regions). An example is presented in Annex G.

2. λ(MGT) and R(MGT) allow rail maintenance managers to compare different track infrastructure combinations and solutions. The ones that show higher reliability values and smaller failure rates should be better in a first approach.

3. If failures are defined as events where rails turn into a state which is non-repairable, that is, rail breaks, the lower expected value of the R(MGT) expressions obtained for all failure modes shall provide a conservative idea of the mean life of rail.

4. If failures are defined as events where maintenance actions are required or should be planned, a planning of maintenance actions may be done in an early stage. Costs and infrastructure availability may be assessed, for it shall be known when to expect a failure and which maintenance action to take in order to prevent/correct it. Since the forecasting is made for different parts of the network, the predictions made will pinpoint the expected geographical locations of failure, and therefore spares provisioning and maintenance crew assignments may be predicted, leading to lower access times and waiting for provisions/spares times, which will decrease MDT. Costs of maintenance operations may be estimated, as well as duration of maintenance actions (which influences system down time, and therefore availability – see the definition of availability in Annex A). Estimation of maintenance costs is nevertheless complex, since they are divided into fixed and variable costs. For instance, maintenance can be carried by own personnel (normally fixed cost) or outsourced crews (fixed cost if on contract, variable if called in). A good approach is to try to estimate the duration of maintenance activities and assign a cost associated with the use of maintenance equipment and workforce. Further studies should be carried on the field of the assignment of costs to

each failure mode. Finally, the Pareto analysis presented in Annex G also presents a method to analyse the weight of each failure mode in total system MDT in a reference year.

5. The last form presented in Annex D, "*Failure correction report – after failure management*", allows studies of MDT and MTTM to be carried, in order to refine the estimates already known for those operations. Nowadays, estimates for MTTM are fairly well known: estimates can be done if resources like crew size and equipment are used as input. As for MDT, it accounts access time, waiting for spares time, restoration and testing time, among others, which may vary a lot depending on the geographical location of the maintenance actions carried. The data sets obtained on MDT and MTTM may be studied using for instance the LogNormal distribution (Chapter 2.5).

4.3 Availability prediction

Availability is obtained from MDT, as described in Chapter 2.3.4. This is a sensitive subject, for there are several factors which contribute to MDT. In the scope of a rail infrastructure, down time is broken down to:

→ MDT_i: Time when rail traffic is halted due to inspection;
→ MDT_{pm}: Time when rail traffic is halted due to preventive maintenance actions;
→ MDT_{cm}: Time when rail traffic is halted due to corrective maintenance actions;
→ MDT_s: Down time which refers to time loss when rail traffic is rerouted or slowed due to speed restrictions posed by other uses of the line (inspection or maintenance).

As a reminder, the formula for availability (expression 2.9) is once more presented:

$$A = \frac{Up\ Time}{Total\ Time} = \frac{Up\ Time}{Up\ Time + Down\ Time} = \frac{MUT}{MUT + MDT} \qquad (2.9)$$

Applying an estimate of MUT in expression 2.9, as the subtraction of MDT to the total time of scheduled train traffic time (here denoted ST), we obtain expression 4.12:

$$A = \frac{MUT}{MUT + MDT} = \frac{ST - MDT}{ST} = 1 - \frac{MDT}{ST} \qquad (4.12)$$

The definition of scheduled train traffic time is controllable by the rail operators. This information is then easily available to the rail infrastructure managers, so the following chapters will be dedicated to the contributing factors of mean down time. MDT may be seen as the sum of all of them:

$$MDT = MDT_i + MDT_{pm} + MDT_{cm} + MDT_s \tag{4.13}$$

The meanings of each parcel are explained in the following subchapters of 4.3. Following this subject, there are 2 chapters: observations on maintenance and scope for future work.

4.3.1 MDT_i: Down time due to train traffic halting for inspection

Normally, inspection is scheduled in the time slots where no train traffic is scheduled (i.e., usually night slots). This shouldn't be accounted as down time if the slots are large enough to hold complete inspection sessions. If for some reason a part of the inspection time falls out of the free slots, then that share of time should account towards down time. This should be easily controllable by the rail infrastructure maintenance manager and therefore no assessment on this is needed, this factor is chosen while planning.

4.3.2 MDT_{pm}: Down time due to preventive maintenance actions

This contribution of down time is similar. Usually, preventive maintenance actions are carried out in free time slots. So the same considerations should be taken. The reader is also referred to the next chapter for further information and the distinction between preventive and maintenance actions.

4.3.3 MDT_{cm}: Down time due to corrective maintenance actions

Corrective maintenance actions have been defined as "actions destined to restore components into a functioning state". This should be clarified. A functioning state is a definition of the infrastructure maintenance manager. It depends on the definition of what the failed state is. Normally, the rail infrastructure maintenance manager should define degrees of severity of rail defects. All severity levels classify as "failure", but only those with a higher degree of severity should be corrected with corrective maintenance.

So, the distinction between preventive and corrective maintenance is the urgency of the latter, for in a general manner, all maintenance actions must be scheduled. Normally, in rail operation, maintenance is preventive. If inspection leads to the discovery of failures which pose a serious risk or threat to the normal operation of the infrastructure (higher degree of severity of rail defects) then corrective maintenance is needed.

So, to obtain an estimate on MDT for corrective maintenance actions, the reliability engineer needs to learn:

→ When the so-called "high risk" failures happen, i.e., when the corrective actions are needed;

→ How much time is took up by the maintenance procedure, from halting of traffic back to normal operation.

The estimation of the number of times high risk failures happen can be obtained by using the procedure stated in Chapter 4.2. The procedure returns a number of high risk failures *n* for a given MGT value. Assessments should be made for different “high risk failures” to obtain the individual values. So, if there are N “high risk failures”, N predictions need to be done.

Then, the individual times for the corrections of all these failures are to be assessed. Maintenance procedures involve (chapter 2.3.3):

→ Waiting for spares: shall be denoted as T_{Ws}
→ Waiting for crew: shall be denoted as T_{Wc}
→ Waiting for equipment: shall be denoted as T_{We}
→ Maintenance action
→ Testing (when applicable) and restoring of normal operation: Shall be denoted as T_R.

Waiting times are normally grouped for usually maintenance crews bring spares and equipment with them. It is then a reasonable approximation to use the largest of the three, or just one time where all of those are counted, which may be referred to as access time (T_a). This is perhaps the most difficult parameter to assess. Access time depends on the relative geographical location between the failure to repair and the resources needed to repair it. If resources are located near, then the maintenance crews may reach the failure zone more rapidly. A suggestion of how to study these times is presented in Chapter 5.2.

Maintenance action, testing and restoring time may all be included in the definition of MTTM (Annex A). These actions may easily be estimated from equipment and crew outputs if the given task is known in beforehand, which is the case – remember the “high risk failures” may be forecast.

So an individual corrective maintenance action time for a failure *i* is given by expression 4.14:

$$MDTcmi = T_{ai\,i} + MTTM_i \quad (4.14)$$

And the forecasting of all corrective maintenance action times is given by 4.15:

$$MDT_{cm} = \sum_{i=1}^{N} n_i \times MDTcmi_i \quad (4.15)$$

4.3.4 MDT_s: Down time (or time lost) due to traffic slowing or rerouting

Some maintenance actions may be performed with speed restrictions on adjacent lines or by diverting traffic to other lines. This allows operation to run partially on the network instead of a total halt due to the affected stretch. Forecasting of time losses due to partial operation is achieved by subtracting the scheduled travel time in normal operation to the actual travel time in partial operation, as given by 4.16:

MDT_s = [train travel time in affected operation] – [train travel time in normal operation] (4.16)

This value should be used instead of the other values previously discussed whenever traffic is not completely halted due to the operations of inspection, preventive and corrective maintenance. This is very difficult to forecast. The general idea is to use traffic slowing or rerouting when:

→ Planned actions are known to take too long (inspection and preventive maintenance). In this case the prediction should be done in beforehand, while scheduling maintenance actions, trying to reduce their impact on train operation to a minimum.

→ In unplanned actions (corrective maintenance). The idea is to pinpoint the geographical location of forecast failures (Chapter 5.1 – 1) and check the potential of rerouting or need to slow down train traffic. In the case of rerouting, traffic on the line where maintenance is carried may be diverted. In the case of slowing down, traffic on lines adjacent to the line where maintenance is being carried needs to be slowed down for security reasons.

The complexity of this subject is too great for the scope of this thesis. Further research needs to be done in the forecasting of this parameter.

4.3.5 Remarks on maintenance

Ideally, no corrective maintenance should be needed, and this is generally the case in high speed rail tracks. High speed rail infrastructure needs to display very high values of reliability due to safety reasons. This is achieved through extensive inspection and continuous monitoring of rail track components.

Since rail networks are very large structures, most failures remain undetected until inspection. A part of detected failures leads to corrective maintenance (the previously called “high risk failures”), while others are left to repair in maintenance actions carried later for the sake of logistic convenience – in the same grinding operation several rail defects may be corrected – or even left to a rescheduled inspection in order to assess their condition.

It is not always clear to state that preventive maintenance, such as preventive grinding, leads to smaller failure rates. While some failure modes may be improved, in others this may not be the case

(Esveld, 2001). Therefore, the idea that increasing preventive maintenance leads to better results in reliability and consequently availability is not correct.

Chapter 5: Scope for future work

5.1. Failure prediction

1. The first suggestion is to standardize the definitions used to create the "comparable stretches of track" ("homogenous definition of technical locations") in an international level. This means standardizing the intervals (for instance for curve radius, axle load of trains, etc.) used to create the data sets for failure rate and reliability. This would allow the compilation of data sets collected by all rail infrastructure maintenance managers, leading to a larger number of rail defects (failures) per comparable stretch, which would improve the statistical study of failures.
   Another issue which is linked to this is to try to understand how to define these intervals. A more detailed break-up of the rail network into comparable stretches leads to a more refined result, for the "breaking-up" process is the way to deal with different operation and interaction conditions of components in different parts of the infrastructure. It is known that over-breaking leads to insufficient data (Chapter 4.2.2), but it hasn't been assessed which degree of separation (i.e. which set of technical characteristics and intervals defined for them) leads to better results. More studies on this subject should be carried.

2. The second suggestion is to collect information on the whole rail network, track infrastructure and train traffic, and introduce it into a unified GIS (geographic information system), using the standardized definition of comparable stretches of track. The current practice is to introduce this data in company specific GIS, but there lacks a standardization and exchange of information between companies. The use of GIS would allow for an automatic splitting of all studied rail networks into comparable stretches according to desired criteria. If this database is created, the reliability engineer barely needs to identify the correct location, date/time and MGT to failure. The infrastructure characteristics may be obtained by the GIS, or conversely, the failure information may be introduced in the GIS, so that the accounting of failures may be made clearly separating comparable track stretches and the same failure modes. The geographical pinpointing of failures allows a more precise allocation of spares and maintenance crews/equipments. Further studies on the data provided by GIS may lead to the optimization of the use of maintenance connected resources such as depots, warehouses, equipments. This in turn leads to the definition of "maintenance jurisdiction areas", that is, the stretches of the rail network which are maintained by the same resources. See also chapter 5.2. Conversely, the failure information spread over the track networks may help in understanding which are the criteria for the splitting of infrastructure characteristics while building "comparable track stretches". This means the information compiled in the GIS may be of assistance to the previous suggestion.

3. The third suggestion is connected to the way costs are looked at. In Chapter 4.2.5, it was stated that costs are easily obtained for the rail due to the specific maintenance actions destined to correct rail defects. Work has not been made regarding the interaction of components. This was not considered in Chapter 4.2.5, for the procedure stated for the separation of data sets takes in account the fact that different infrastructures are subject to different interactions between components, therefore leading to different values for λ(MGT), R(MGT), etc. of the rail. Nevertheless, the process presented in Chapters 2.2.2 and 2.2.3 should be regarded, with individual failure rates and reliability functions being obtained for each component and methods stated in Annex D applied for system level interaction (where system here is denoted as the rail infrastructure). The work presented on this study should suffice for the estimation of rail-related maintenance action costs, but future modelling work should be done at system level (i.e., regarding the interaction of all components of the rail infrastructure), to assess if the results obtained are better, and to question the possibility to model the costs related to the whole rail track. An idea for a FMEA sheet is presented in Annex H, where an example is shown on how the reliability engineer should deal with the potential causes and consequences of failure. In the example, it is shown how for instance corrugations can affect the ballast bed. This should be used as a starting point for a GFCM or a CQTM (Annex D).

4. The fourth issue regards the assumption that the presence of a failure mode doesn't interfere with the development of other failure modes in its vicinity. Studies should be made regarding the application of GFCM (mentioned in the previous paragraph) to build Internal Gathered Faults (IGFs) for the rail (Annex D).

5. Finally, studies should be carried on how to optimize the treatment of different failure modes with the use of the same maintenance action. For instance, rail grinding (Annex 3.4.2) corrects wheel burns, shelling, head checks, spalling, plastic flow and tongue lipping. A way to save money is to apply rail grinding in the treatment of stretches which display more than one of these rail defects (failure modes). Here, the application of GFCM mentioned in the previous paragraph would be of vital importance to understand the interaction between failure modes, and how much time (or MGTs) would be safe to wait for other rail defects to appear before maintenance actions would be required.

5.2 Availability prediction

1. The suggestion to estimate access time is to firstly break down the network into "maintenance jurisdiction areas". The idea is to assign different parts of the rail network to different maintenance crews who are provided with enough resources to carry any needed corrective maintenance operation. Then, data on access time *vs.* distance to failure should be collected by the reliability engineer each time a procedure like this is needed. The data set should be studied with the use of

a statistical distribution such as the log-Normal distribution to forecast access times individually (Chapter 2.5.4). The input data shall be the distance to the maintenance crew depot. As a reminder, it must be said that it is possible to foresee the geographical location of "high risk failures" if the reliability data collection procedure stated on Chapter 4.2 is used with a high break down level in geographical terms.

2. Studies need to be done in the field of preventive maintenance operations to assess which failure modes' failure rates are improved at the expenses of which (Chapter 4.3.5).

3. Research should be carried in equipment which provides condition monitoring data. The idea behind this is to keep track of the condition of all components of the rail infrastructure in a *quasi* permanent manner, with very regular updates. It is obvious that real time information on the condition of the infrastructure allows to decide when and where to carry maintenance actions, reducing costs and down time (unavailability). On the other hand, keeping track of all of the infrastructure's conditions is very expensive. The ideal level of condition monitoring would be the information update interval on components which leads to the lowest overall cost, which is the sum of the cost of monitoring and the cost of not performing maintenance due to not having more information. Further studies should be made to seek which is the optimal interval of condition monitoring taking in mind all possible failure modes and associated maintenance actions.

Chapter 6: Conclusions

The goal of this study is to assess how RAMS techniques can be applied to the rail in order to forecast failures and their associated costs when data sets are short.

A process to collect failure data was presented, which takes in consideration the recommendations of RAMS authors and the functioning and displayed failure modes (i.e., defects) of the rail. This process was designed to obtain results directly from the rail component, skipping RAMS modelling at system level (here, the word system refers to the whole rail infrastructure). This was achieved through the splitting and separate study of data sets into different train traffic and infrastructure conditions. It was shown that expressions for reliability and failure rate may be obtained by applying the Weibull distribution to the data sets obtained by the previously mentioned process. The Weibull distribution was shown to be suitable to short data sets, since it works with a least amount of 6 points. Also, the Weibull distribution was shown to be suitable for the study of the rail due to the fact that it allows to model the behaviour of components with non-constant failure rates. An Excel-based process to obtain the expressions for failure rate and reliability was displayed. It was also discussed how to use these expressions on Pareto analyses, allowing to compare the relative importance of failures in maintenance costs and down time and to obtain an order of magnitude of their values. The problem of the accuracy of failure rate and reliability predictions was approached, being stated that RAMS authors clearly indicate that the absolute values of predictions should be relied upon just to display the expected order of magnitude, and that RAMS predictions should be used mainly in comparative analyses. This was another reason which led to the use of Pareto analyses in the study of the costs connected to failure, since they are mainly comparison techniques.

Prediction of availability was discussed. It was shown that availability may be calculated based on estimations of mean up time and mean down time. MUT is set by the train operator. A break down of MDT into a sum of several different components was displayed. This was concluded to be of hard assessment. Firstly, these components are dependant on each other. Secondly the different logistic factors and maintenance strategies available lead to a fluctuation of the observed MDT values. Furthermore, other studies have already shown that preventive maintenance may lead to lower failure rates in some failure modes at the expenses of higher failure rates in other failure modes. It was recommended to collect more data on MDT, which is achieved by using a dedicated section of the data collection forms. As stated by other authors, prediction of MDT is achieved through the log-Normal distribution, which may be done in the future if the stated procedure is applied. With the prediction of MDT it is then possible to estimate availability.

Bibliography

AWS (2004). *AWS Welding Handbook Volume 2 – Part 1: Welding Processes (WHB V2)*. American Welding Society.

CANNON, D. F. (2003). *An international cross reference of rail defects, 2nd edition*. The Steering Group of UIC/WEC - Union Internationale des Chemins de Fer, Paris, France.

DORNER, William (1999). *Using Excel for Weibull Analysis*. Quality Digest.

DOWSON (1999). *Lubrication at the Frontier: The Role of the Interface and Surface Layers in the Thin Film and Boundary Regime*. Elsevier.

ESVELD, Coenraad (2001). *Modern Railway Track.* 2nd edition. MRT Productions, Delft University of Technology.

H.J. Skelton Canada, Ld. *http://skelton-metals.com/railsections.htm*

IEC (2002). *International Standard IEC 62278: Railway applications – Specification and demonstration of reliability, availability, maintainability and safety (RAMS).* First Edition 2002-09. International Electrotechnical Commission.

KHUMAR, S. (2007). *Study of Rail Breaks: Associated Risks and Maintenance Strategies.* Luleå Railway Research Center (JVTC)

KRULL, R. Hintze H., LUKE, THOMAS M., POHL R., RÜHE S. (2002). *Eddycurrent Detection of Head Checks on the Gauge Corners of Rails: Recent Results.* DB AG, the Research and Technology Centre; BAM Federal Institute for Materials Research; Prüftechnik Linke & Rühe GmbH, Germany

LUNDTEIGEN, Mary Ann & RAUSAND, Marvin (2006). *Assessment of Hardware Safety Integrity Requirements.* Proceedings of the 30th ESReDA Seminar, June 07-08, 2006, Trondheim, Norway.

MANLY P., Harold (2006). *Oxy-acetylene welding and cutting.* Indypublish.com.

MATHER, Daryl (2005). *The Maintenance Scorecard – Creating Strategic Advance.* Industrial Press, Inc.

MILLS (1956). *Materials of Construction, Their Manufacture and Properties*. Wiley.

NAHMAN, Jovan M. (2002). *Dependability of Engineering Systems – Modeling and Evaluation.* Springer.

NOWLAN, F. Stanley & HEAP, Howard F. (1978). *Reliability-Centered Maintenance.* United Airlines, USA.

NUNNS, S. R. (2002). *Guiding principles to the implementation of IEC-61508.* Proceedings of the SIPI Workshop 3, the 14th and 15th of November 2002, London, UK.

PROFILLIDIS, Vassilios A (2000). *Railway Engineering.* 2nd edition. Ashgate Publishing Limited.

SMITH, David J. (2005). *Reliability, Maintainability and Risk – Practical Methods for Engineers. 7th edition*. Butterworth-Heinemann.

SPG Media PLC. Web source: http://www.railway-technology.com/projects/germany/

Transportation Safety Bureau Canada.
Web source: http://www.tsb.gc.ca/en/publications/reflexions/rail/2003/issue_19/rail_issue19_sec2.asp

U.S. Dept. of the Army (1991). *Railroad Track Standards: Appendix B - Field identification of rail defects.* Departments of the Army and Air Force.

VILLEMEUR, Alain (1992-A, 1992-B). *Reliability, Availability, Maintainability and Safety Assessment,* A: *Vol. 1: Methods and Technique.* B *Vol.2: Assessment, Hardware, Software and Human Factors*: John Wiley & Sons.

VITEZ, I., ORUČ, M., KRUMES, D., KLADARIČ, E. (2007). *Damage to railway rails caused by exploitation.* METALURGIJA 46 (2007) 2, 123-128

VITEZ, I., KRUMES, D., VITEZ, B. (2005). *UIC recommendations for the use of rail steel grades.* METALURGIJA 44 (2005) 2, 137-140

WEMAN, K. (2005). *Welding Processes Handbook.* ESAB Research, Sweden. Woodhead Publishing Limited.

Annex A: Technical terms and definitions in RAMS

Assessment

Investigation process with the goal of judging the suitability of a system

Availability - A(t)

"Ability of a certain entity to be in the state of providing a certain function under certain conditions, at a given time instant (Villemeur, 1992-B)." It can be measured by the probability of an entity E not being failed at a generic time instant t, that is:

$$A(t)=P[\text{entity E not failed at time t}] \tag{A.1}$$

It can alternatively be expressed as the ratio of up time over total working time (Smith, 2005):

$$A=\frac{Up\ Time}{Total\ Time}=\frac{Up\ Time}{Up\ Time+Down\ Time} \tag{2.9}$$

Confidence interval

Range of values for a random variable limited by two bounds such that the studied variable will have a probability of (1-α) of falling into this range, where 1-α is the confidence level (Villemeur, 1992-B)

Common cause failure

Failure which results of an event that creates a coincidental failure state in more than one component at the same time, resulting in the defeat of redundancy of the failed components

Corrective maintenance

Operation carried after a component failure is identified intended to restore the component back into a functioning state

Down time

Time interval during which a component is in a failed state, being unable to perform its functions

Failure

Departure of a component's functionality targets from specification (Smith, 2005). Alternatively, the definition "termination of the ability of an entity to perform a required function under specified conditions" (Villemeur, 1992-B) can be used[11].

[11] For RAMS studies purposes, in order to avoid misinterpretations, every possible failure for each component must be clearly defined *a priori* and contractually. This is done by defining the abovementioned "specification" and "ability to perform a required function" by means of absolute values and predicted variation ranges in diverse parameters (Smith, 2005).

Failure cause

Events or circumstances that lead to a failure, due to design, manufacturing, or operation (Villemeur, 1992-B)

Failure mode

Predicted or observed consequences of a failure in a component or system in relation to the operating conditions at the time of failure (IEC, 2002). Effect by which a failure is observed (Villemeur, 1992-B)

Failure rate

Represented by $\lambda(t)$, it is by definition the transition function between a working state and a failed state of a component, sub-system or system. It can be written analytically as the probability of a failure to occur in a time interval given that the component was working up until then. Being a transition function, it deals with short time intervals (Nahman, 2002), and as so it becomes[12]:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P\{t < T \leq t + \Delta t | T > t\}}{\Delta t} \qquad \text{(A.2)}$$

Field data

Data obtained during field observations (Villemeur, 1992-B), i.e. while the systems are in operation and subject to specific environmental conditions and maintenance policies

Hazard

Situation which has the potential to cause: damage to the system, damage to its surrounding environment, injuries or loss of human lives. The term “accident” can also be used

Life cycle

Set of activities that occur involving a system, from its conception phase up until the disposal and decommissioning phase

Maintainability

Ability of an entity to be restored into or be kept in a condition or state that enables it to perform a required function, when maintenance operations are performed under given conditions and are carried using stated procedures and resources (Smith, 2005). In short, it is the ability of an entity to be repaired in a given time (deadline). It is normally measured by the probability that the maintenance procedure of a certain entity E performed under certain conditions is finished at time t given that the entity failed at time t = 0 (Villemeur, 1992-A), that is:

M(t)=P[maintenance of E is completed by time t], when E fails at time t = 0 (A.3)

---

[12] The time interval definition provides a failure rate in the units of failure per time. Some applications might define failure rate in terms of failure per cycles of usage or distance travelled, if it is found to be more convenient.

Maintenance

Actions destined to keep or restore an item into a functioning state, that is, a state in which it can perform a required function (Villemeur, 1992-B)

MDT

Acronym for Mean Down Time. It is the expected value of all experienced component down times. Please refer to the definition of “down time”, and of “MTTR”.

Mission

Objective description of the fundamental task performed by a system (IEC, 2002)

Mission profile

Definition of the expected variation in parameters that influence a system’s mission such as travel time, speed, loading, etc.

MTBF (Mean Time Between Failures)

Also represented by $\hat{\theta} = T/k$ where T is the time elapsed (operation, standby, or combined), and k is the number of observed failures in a studied component. It is the expected value for operating time between the occurrences of two failures (Smith, 2005). For constant failure rates, the function $\theta(t)$ that returns time elapsed between two failures becomes a constant function as well and MTBF equals $\hat{\theta} = 1/\hat{\lambda}$ (Smith, 2005).

MTTF (Mean Time to Failure)

MTBF of the first failure. Due to their definition, in non-repairable items MTTF≡MTBF

MTTR (Mean Time to Repair)

Expected value of all the repair times experienced by a component

MUT (Mean Up Time)

Expected value of up (operating) time of a component

Preventive maintenance

Maintenance operations carried out at specified periods (time, cycles, distance, etc.) aimed at reducing the probability of failure of a component (Villemeur, 1992-B)

Probability of failure-on-demand

Sometimes written PFD. Also known as unavailability. Defined as (Smith, 2005):

$$\overline{A} = 1 - A = \frac{\lambda MDT}{1 + \lambda MDT} \cong \lambda MDT \qquad (2.7)$$

Random failures

Failures which happen when stress peaks overlaps the strength peaks in the components' population. Mostly associated with hardware, it's the failure type normally displayed in available failure data collections. Usually a constant rate of occurrence of such failures is assumed (Lundteigen & Rausand, 2006).

RAMS

Acronym for Reliability, Availability, Maintainability and Safety

Reliability - R(t)

Ability of an entity to perform a required function under given conditions for a given time interval (Villemeur, 1992-A). In other words, an entity is reliable if it hasn't failed – i.e., stayed within the specifications – over a time interval. Measured by:

R(t)=P[entity E not failed over time [0,t]]; the entity is assumed to be operating at time t=0

(A.4)

Reliability growth

Process of improvement of reliability parameters for a given component with time

Repair

Mechanical or manual operations performed on the failed component during a corrective maintenance procedure

Repair rate

Limit, if any, of the ratio of the conditional probability that the corrective maintenance action ends in a time interval, [t, t+Δt] over the length of this time interval, Δt, when Δt tends to zero, given that the entity is faulty at time t=0 (Villemeur, 1992-B). Also represented by μ(t)

Repair time

Time during which repair actions occur

Risk

Result of the crossing of two criteria: probable frequency of occurrence and degree of severity of the impact of a hazard

Safety

Freedom from unacceptable risk of harm (IEC, 2002)

System component
Smallest unit in a system which can be accurately tested in terms of RAMS parameters

Systematic failures
Failures which are caused by errors in any safety life cycle activity, under some particular combination of inputs or a particular environmental condition (IEC, 2002). They are mostly associated with software or systems with a high degree of complexity and interaction between components, and need to be accounted separately from the values on failure rate provided by failure rate data collections (Smith, 2005).

Up time
Time during which a component is in a functioning (available) state, i.e. the component has the ability to perform a required function under stated conditions

Annex B: Interrelationships of reliability terms

B.1 Boolean algebra and probabilistic definitions

When a trial is done, a set of possible occurrences is yielded. This is known as set of observables. Let's take for instance the throwing of two coins. The set of observables, Ω, will be:

Ω = {heads-heads, heads-tails, tails-heads, tails-tails} (B.1)

An event can be defined as a subset of Ω. An event can be for instance "the result heads-heads", or as well "the result heads in one of the coins". Each of these events, let's call them A and B, have a probability of occurrence. These are written:

$$P[A] = x_1 \quad and \quad P[B] = x_2 \qquad \text{(B.2; B.3)}$$

It is said that A and B are identical if they have the same probability of occurrence. It is said B is the inverse of A if B occurs when A does not occur and vice versa. In that case, B can be written as Ā, that is:

$$P[\overline{A}] = 1 - P[A] \quad \Leftrightarrow \quad P[A] + P[\overline{A}] = 1 \qquad \text{(B.4)}$$

Another definition is the impossible event, which never occurs and has a null probability of occurrence, and is represented by Ø. Identically, the certain event is the event which always occurs, and it can be represented by the entire set Ω.

$$P[\varnothing] = 0 \quad and \quad P[\Omega] = 1 \qquad \text{(B.5; B.6)}$$

For a study of the occurrence of events simultaneously or alternatively, a diagram of Venn is represented in Figure B.1.

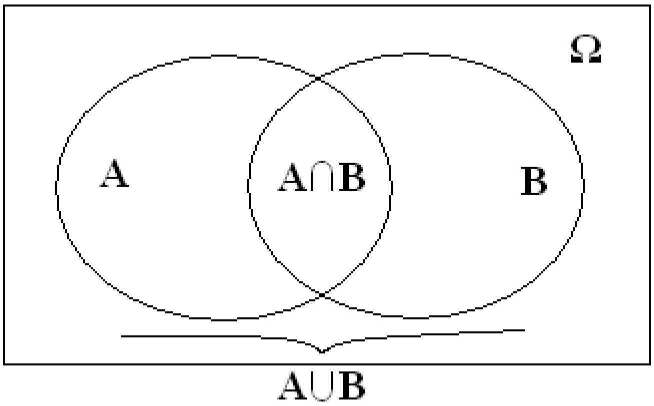


Figure B.1 Venn diagram.

Event A AND B is defined as the occurrence of A and B in simultaneous. It can be written as yet another subset of Ω, C. It is said that events A and B are mutually exclusive when A∩B = Ø.
Event A OR B is defined as the occurrence of either (at least) event A or B. In the same manner, it can be written as another subset of Ω, for instance C.
There are other symbols which can represent AND, OR, and the certain and impossible events. Those are presented in Table B.1.

Table B.1 Probabilistic conventions.

Operator/subset	Probabilistic symbols	Boolean algebra symbols
AND	∩	.
OR	∪	+
Certain event	Ω	1
Impossible event	Ø	0

The last column represents the symbols most commonly used in Boolean algebra equations. These equations are the ones commonly used to track down minimal cut paths, i.e., the smallest combination of component failures that leads to system failure. Several handy Boolean algebra rules can then be written:

Commutative laws:

$$A \bigcap B = B \bigcap A \quad A.B = B.A$$
$$A \bigcup B = B \bigcup A \quad A + B = B + A$$

(B.7; B.8; B.9; B.10)

Associative laws:

$$A \bigcap (B \bigcap C) = (A \bigcap B) \bigcap C \quad A.(B.C) = (A.B).C$$
$$A \bigcup (B \bigcup C) = (A \bigcup B) \bigcup C \quad A + (B + C) = (A + B) + C$$

(B.11; B.12; B.13; B.14)

Distributive laws:

$$A \bigcap (B \bigcup C) = (A \bigcap B) \bigcup (A \bigcap C) \quad A.(B + C) = A.B + A.C$$
$$A \bigcup (B \bigcap C) = (A \bigcup B) \bigcap (B \bigcup C) \quad A + B.C = (A + B).(A + C)$$

(B.15; B.16; B.17; B.18)

Idempotent laws:

$$A \bigcap A = A \quad A.A = A$$
$$A \bigcup A = A \quad A + A = A$$

(B.19; B.20; B.21; B.22)

Laws of absorption

$$A \bigcap (A \bigcup B) = A \quad A.(A + B) = A$$
$$A \bigcup (A \bigcap B) = A \quad A + A.B = A$$

(B.23; B.24; B.25; B.26)

Complementation laws

$$A \bigcap \overline{A} = \varnothing \quad A.\overline{A} = 0$$
$$A \bigcup \overline{A} = \Omega \quad A + \overline{A} = 1$$
$$\overline{\overline{A}} = A$$

(B.27; B.28; B.29; B.30; B.31)

Ø and Ω related (0 and 1)

$$\varnothing \bigcap A = \varnothing \quad 0.A = 0$$
$$\varnothing \bigcup A = A \quad 0 + A = A$$
$$\Omega \bigcap A = A \quad 1.A = A$$
$$\Omega \bigcup A = \Omega \quad 1 + A = 1$$
$$\overline{\varnothing} = \Omega \qquad \overline{0} = 1$$
$$\overline{\Omega} = \varnothing \qquad \overline{1} = 0$$

(B.32; B.33; B.34; B.35; B.36; B.37;

B.38; B.39; B.40; B.41; B.42; B.43)

De Morgan's theorems

$$\overline{(A \cap B)} = \overline{A} \cup \overline{B} \quad \overline{A.B} = \overline{A} + \overline{B}$$
$$\overline{(A \cup B)} = \overline{A} \cap \overline{B} \quad \overline{(A + B)} = \overline{A}.\overline{B}$$

(B.44; B.45; B.46; B.47)

Bayes' theorem (Smith, 2005)

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]}$$

(B.48)

Other useful theorems (Villemeur, 1992-A)

$$A \cup (\overline{A} \cap B) = A \cup B \qquad A + (\overline{A}.B) = A + B$$
$$\overline{A} \cap (A \cup \overline{B}) = \overline{A} \cap \overline{B} = \overline{(A \cup B)} \quad \overline{A}.(A + \overline{B}) = \overline{A}.\overline{B} = \overline{(A + B)}$$

(B.49; B.50; B.51; B.52)

B.2 Mathematical deductions regarding failure rate, reliability, unreliability and the probability density function

Reliability, R(t), is defined as the probability that at time t, the entity is in a sound state. If the duration of the sound state is T, then reliability can be written as:

$$R(t) = P\{T > t\}$$ (B.53)

This meaning that the time instant t is smaller than the sound state. Unreliability, Q(t), can then be expressed as the probability that the sound state T (state under which the entity presents no failure) is smaller or equal to the considered time instant "t":

$$Q(t) = P\{T \leq t\}$$ (B.54)

Applying basic probabilistic laws (expression B.4), we find that:

$$R(t) + Q(t) = 1$$ (B.55)

The probability density function for the variable T is the differentiation of unreliability:

$$f(t) = \frac{dQ(t)}{dt} \tag{B.56}$$

Also applying arguments of probability theory, it can be stated that:

$$dQ(t) = Q(t + dt) - Q(t) = P\{t < T \leq t + dt\} \tag{B.57}$$

Differentiating the sum of reliability and unreliability (B.55 and B.56), it can be concluded that:

$$\frac{d[R(t) + Q(t)]}{dt} = \frac{d1}{dt} \Leftrightarrow \frac{dR(t)}{dt} + \frac{dQ(t)}{dt} = 0 \Leftrightarrow \frac{dR(t)}{dt} = -\frac{dQ(t)}{dt} \Leftrightarrow \frac{dR(t)}{dt} = -f(t) \tag{B.58}$$

Taking expression B.56 and solving the probability density function in order to unreliability returns:

$$f(t) = \frac{dQ(t)}{dt} \Leftrightarrow Q(t) = \int_{-\infty}^{t} f(t)dt = \int_{0}^{t} f(t)dt \tag{B.59}$$

The lower bound limit in expression B.59 can be cut to 0 since f(t) is a null function for negative time values – it only has (positive) non-null values from the time of construction or installation. Reliability can also be expressed through the probability density function, through expressions B.55 and B.59:

$$R(t) = 1 - Q(t) = 1 - \int_{0}^{t} f(t)dt = \int_{0}^{\infty} f(t)dt - \int_{0}^{t} f(t)dt = \int_{t}^{\infty} f(t)dt \tag{B.60}$$

Failure rate, $\lambda(t)$, is by definition the transition function between a sound state and a failed state. It can be written analytically as the probability of a failure to occur in a time interval given that it was working up until then, and since it is a transition function, it deals with short time intervals (Nahman, 2002), and as so it becomes:

$$\lambda(t) = \lim_{\Delta t \to \infty} \frac{P\{t < T \leq t + \Delta t | T > t\}}{\Delta t} \tag{A.3}$$

Using Bayes theorem (expression B.48) and expression B.53 it is possible to write:

$$P\{t < T \leq t + \Delta t | T > t\} = \frac{P\{t < T \leq t + \Delta t, T > t\}}{P\{T > t\}} = \frac{P\{t < T \leq t + \Delta t\}}{R(t)} \tag{B.61}$$

Here the event “B” could be eliminated from the first element of the fraction in expression B.61 since it was included in event “A”.

Using A.3, B.56, B.57 and this last simplification (B.61), the following conclusion can be taken (Nahman, 2002):

$$\lambda(t) = \frac{f(t)}{R(t)} \tag{B.62}$$

Now, using B.62 and taking the probability density function written expressed through reliability (B.58), the differential equation can be solved, leading to:

$$\frac{dR(t)}{dt} = -f(t) \Leftrightarrow \lambda(t) = \frac{-dR(t)}{dt}.\frac{1}{R(t)} \Leftrightarrow R(t) = \exp\left[-\int_0^t \lambda(t)dt\right] \tag{B.63}$$

Mean Time To Failure is the mathematical expectation of an entity’s up-time, in the case of a non-repairable entity. MTTF’s expression can also be used for a repairable entity’s MTBF if it is considered that the repair operation returns the entity to a state equivalent to the state it had when it was new (Nahman, 2002). The expression is then:

$$\theta = E\{T\} = \int_0^\infty t.f(t)dt = -\int_0^\infty t.R'(t)dt = -t.R(t)\Big|_0^\infty + \int_0^\infty R(t)dt = \int_0^\infty R(t)dt \tag{B.64}$$

The original integral in B.64 was solved by partial integration, and the first term of the resulting equation could be eliminated due to the fact that reliability decreases faster than time (Nahman, 2002).

If the failure rate $\lambda(t)$ is assumed to be constant (equal to λ), then it is possible to simplify B.63 to:

$$R(t) = \exp\left[-\int_0^t \lambda(t)dt\right] = \exp\left[-\lambda t\Big|_0^t\right] = e^{-\lambda t} \tag{B.65}$$

And MTBF (B.64) can be simplified to:

$$\theta = \int_0^\infty R(t)dt = \int_0^\infty e^{-\lambda t}dt = \frac{1}{\lambda} \tag{B.66}$$

As a consequence of the way the expressions are defined, it is possible to obtain either reliability, unreliability, the probability distribution function of T and failure rate from any of the other three (Nahman, 2002).

Annex C: List of significant databanks/sources

Military sector:

- Military Handbook 217E (MIL HDBK 217E), compiled in the US, dedicated to the electronic parts of military equipment.
- HRD5 Handbook of Reliability Data for Electronic Components used in Telecommunications Systems, a database compiled by British Telecom. It usually presents optimistic results, frequently in one order of magnitude, since failures are excluded from the data as soon as corrective measures for them are applied in manufacturing processes. Maintenance-induced errors are also excluded from the data.
- RADC NPRD 2 (Nonelectronic Parts Reliability Data), a document made in the US which handles mechanical and electromechanical parts used on military equipment.

Nuclear sector:

- System Reliability Service Data Bank (SYREL), which was started by the United Kingdom Atomic Energy Authority as an inventory of incidents in its nuclear power plants since 1961, that subsequently led to the creation of SYREL in 1967 by the “Systems Reliability Service” (SRS), being extended to other industrial plants.
- Nuclear Plant Reliability Data System (NPRDS), developed by the Institute of Nuclear Power Operations (INPO), regarding nuclear power plants in the US.
- Other US sources for Nuclear Reliability: Appendix III of WASH 1400 study, NUCLARR (Nuclear Computerized Library for Assessing Reactor Reliability), NUREG.

Electrotechnical, telecommunications, and other industrial sectors:

- Reliability database of the Centre National d’Études des Télécommunications, a database developed by CNET dedicated to the failures of telecommunication equipment.
- Bellcore (Reliability Prediction Procedure for Electronic Equipment) TR-NWT-000332 Issue 5 1995, data compiled by Bell telephone companies for telecommunication devices. Other phone companies have failure rate databanks but these are not available for purchase. Companies such as Nippon Telephone Corporation, Ericsson and Thomson CSF use these internally. It’s also worth to mention Siemens and RM Consulting, which are in the same situation and not exclusively dedicated to telecommunications.
- Offshore reliability data (OREDA), a database concerning oil-rig equipment studied in the North Sea and Adriatic.
- SINTEF, which is a part of the Norwegian Institute of Technology at Trondheim, collects data related to Fire and Gas Detection equipments.
- TECHNIS, a databank compiled by Dr. David Jay Smith, which comprises industry and generic failure rates and some repair times.
- European Reliability Data System (ERDS), created by the EC Research Center in Ispra (Italy), which comprises four data banks regarding components of PWR and BWR power plants.

- Scandinavian Nuclear Power Reliability Data System (ATV-System), which monitors nuclear power plants in Scandinavia.

Other power industries sectors:

- EPRI (Electric Power Research Institute), of General Electrics Co., New York, deals with gas turbine failure data in the USA.
- GADS (General Availability Data System), maintained by NARC, North American Electric Reliability Corporation, which bases its data on statistics collected in power plants in the USA and Canada

Data source documents:

- Probabilistic risk assessment in nuclear power plants, a report on PWR and BWR power plants in the US.
- Analysis of significant events in nuclear power plants; it has been systematically recorded in the Licensee Event Reports for several years.
- In-plant reliability data system, a program developed by the Institute of Electrical and Electronics Engineers (IEEE) started in 1977 which collects data from several nuclear reactors' components and maintenance programs.
- IEEE Std 500-1984 Document, published by IEEE firstly in 1977 based in the efforts made on the In-plant reliability data system, aimed at improving data quality and its scope.

Older sources. Despite their age, these collections are still used nowadays as references (Smith, 2005):

- Reliability Prediction Manual for Guided Weapon Systems (UK MOD) - DX99/013–100
- Reliability Prediction Manual for Military Avionics (UK MOD) - RSRE250
- UK Military Standard 00–41
- Electronic Reliability Data - Inspec/NCSR (1981)
- Green and Bourne (book), 1972. *Reliability Technology.* Wiley.
- Frank Lees (book). *Loss Prevention in the Process Industries*. Butterworth-Heinemann.

Annex D: Quantitative approach methods of RAMS simulation

D.1 Success diagram method (SDM)

This is the first method applied in the assessment of systems reliability. It is also known as "Reliability Block Diagrams (RBD)". It is a relatively effective tool of predicting reliability quantitatively – especially in non-repairable systems – but proves insufficient in the evaluation of safety aspects.

It consists mainly in the representation of system components (or the functions they perform) as blocks. A diagram of the system is built using links between the blocks, and system failure can be defined through two basic linkage methods:

- Series diagram: system failure is caused if one of the several series blocks experiences a failure. For example, in Figure D.1, there is only going to be an output signal "O" if c_1 and c_2 are working simultaneously.

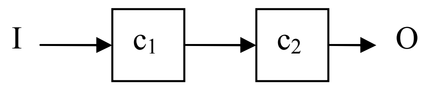


Figure D.1 Series block diagram. (Adapted from Villemeur, 1992-A)

- Parallel diagram: system failure is caused if all the blocks in parallel fail. In the figure below, there will be output if either c_1 or c_2 is working. Figure D.2 shows the schematic way to represent redundancy.

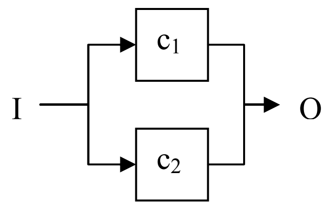


Figure D.2 Parallel block diagram. (Adapted from Villemeur, 1992-A)

Other special blocks can be used to build systems, according to the needs raised by each system's complexity level, like m/n parallel systems where m out of n components must be working (in parallel) in order to create an output signal, or systems where passive redundancy is used (redundancy systems which are in standby and only work upon request, i.e., when the previous component fails).
Needless to say, a study of how component failures affect the system functions and cause the system to fail needs to be carried in order to correctly depict and represent the system diagram. FMEA can prove itself useful in this assessment.

Knowing the failure rates of the components c_i, the analyst can then use probabilistic calculations to determine the system's failure rates or MTTFs. These won't be looked at in detail for they are not relevant in this study's scope, but it must be stated that this method will only work if the probabilistic events (in this case, the failures) are independent. When this is not verified, other methods such as Cause Tree Method (CTM) should be used.

As a complement of this method, the Counts Part Method or CPM can be used. This method assumes that all the components in the system need to be working in order not to cause a system failure. Being so, the equivalent success diagram is a series diagram and the failure rate is simply the sum of every component's failure rate. If the assumption is wrong, the value obtained can be used as a conservative estimate of the system failure rate. However, it is hard to quantify how conservative it is, but it is still useful to provide a general picture of the order of magnitude of the system failure rate.

D.2 Cause tree method (CTM)

This method was firstly developed in Bell Telephone Laboratories by Watson, to study the reliability of the Minuteman Missile Launch Control System. It was successful at improving its reliability and identifying and eliminating weak links in the project. Since then its use has been growing and was extended to other industrial sectors and nowadays it stands as the most used reliability assessment method.
Its principle is similar to the SDM, since the final result is a schematic representation of the system's failures. What basically is done is an identification of the combinations of single component failures that lead to a system failure (undesirable event), which are then drawn in a tree-like structure, using logic gate operators.

It is important to firstly define the undesirable event that is going to be studied, so the other events (single failures or combination of failures) that lead to it can be identified, and so on. The process keeps going until the events are broken down into a level that is basic enough to allow them to be considered independent and their probability known (single component failures, normally). The "undesirable event" defined will then sit on top of the tree-diagram and the events that lead to it will be on the lower levels and so on. It is important not to choose an "undesirable event" too comprehensive as it might be difficult to identify the causes that lead to it, but it mustn't also be too specific for the analysis may then by-pass the important elements of the system. In this stage, a FMEA is helpful to identify the "undesirable events".

The connections between events are, by obligation, the above mentioned logic gates. These will dictate which are the interactions between the events of the same level that lead to the next event further up in the tree. Logic gates can be designed according to system specifications, but the mainly used ones are the AND and OR logic gates. Other examples include IF, AND with condition, OR with condition, m/n, EXCLUSIVE OR, PRIORITY AND, DELAY, MATRIX, QUANTIFICATION, SUMMATION, COMPARISON, among others (logic gates are treated mathematically in a way comparable to the block diagrams used in SDM).

Events can be represented in three ways: as a circle, representing a primary failure; as a diamond, representing a secondary failure, or as a rectangle, representing a command failure.
Primary failures are the ones intrinsic to the components, for instance in a motor, a primary failure would simply be motor breakdown. Secondary failures include environmental conditions and human operator errors, and command failures are the ones associated with the components which give out the instructions to parts of the system (like switches, levers or buttons), and may cause problems in their operation.

The tree-like structure that represents the system is not always easy to build and some guidelines must be followed:

- The undesirable events must have their immediate, necessary and sufficient (INS) causes identified, in order to develop the lower levels of events
- The intermediate events must be correctly classified and must be connected with component defects
- The previous steps must be repeated until the "basic level" of events is obtained
- Successive iterations can be used to improve the model where it lacks supporting data on failures
- No-miracles rule: if the system is in a failed state, no component failure can make it go back into the non-failed state
- Complete the gate rule: events must first be linked to a logic gate before they are studied in detail
- No gate-to-gate: between logic gates there must be events and vice-versa.
- Causes precede consequences rule: looking further down in the tree is like looking back in time. If a consequence is found to be similar to the cause that creates it, that "cause branch" must be eliminated

Finally, reduction of this method through Boolean algebra needs to be done, that is, verifying of the possible cause combinations which can be eliminated from the tree due to the fact that they are already included as subsets of other combinations. This will simplify the study and lead to the so called minimal cut set or path to failure, which identifies which is the simplest combination of events (here, causes) leading to a failure.

When this method is being used, causes for failures are being analysed. In the majority of cases, the analyst normally chooses the most likely causes that lead to the events, especially if he is taking in consideration the first rule stated above. This may be seen as a problem in a pure modelling point-of-view since it may omit other possible (though unlikely) causes, but it proves to produce solid results, and leads to more practical model solutions – it can be a burdensome task to account every cause, and in practice eliminating the low probability causes won't cause a great impact in the final outcome. Again it's important to stress that each model has its own flaws and this is no exception to other branches of engineering; models are just attempts to describe reality; they don't represent reality itself. As stated above, this method is in widespread use in many industrial fields and there is a large number of software applications developed for its quantitative analyses, but as a downside it is extremely hard to implement in complex systems with many interacting elements, since the causes leading to failures might be hard to track down. This method should be complemented with the FMEA in order to provide assistance on this last subject. Also, systems with a low number of decision processes in safety procedures, i.e. sequential event oriented systems, are of more suitable analysis with the consequence tree, cause-consequence diagram, and space state method.

A good current software application for calculations involving CTM is TTREE (Smith, 2005).

D.3 Truth table method (TTM)

The truth table method is a relatively simple method in terms of theory, and therefore it is applicable to almost any field, but is especially used in the fields of electrical and electronic engineering.

The method requires a FMEA to assess which are all the possible operating and failed states of components and the truth table per se will list all the possible combinations of these states. The table presents the components in columns and the combinations of operating or failed states of the components are presented in rows. The identification of operating or failed state is made, correspondingly, by 0 or 1, so for instance, for a series or parallel system of two components c1 and c_2, the truth tables will be the ones presented in Figure D.3:

C_1	c_2	System
0	0	0
1	0	1
0	1	1
1	1	1

c_1	c_2	System
0	0	0
1	0	0
0	1	0
1	1	1

Figure D.3 Truth tables for series (left) and parallel (right) component arrangements. (Adapted from Villemeur, 1992-A)

The result of the state combinations in the system is read in the "System" column. What needs to be done is to identify all the combinations leading to a non-functioning system, and furthermore, combinations included in other combinations need to be cut out in order to obtain a reduced truth table, which will allow a simplified study and the definition of the minimal path to failure. This reduction can be done employing the use of Boolean algebra, being similar to what happens in the cause tree method. Also, incompatible combinations must be ruled out.

As it is simple to observe, the truth table construction uses a very basic method. This is why it is applicable to so many fields, but as a downside it will be very difficult to apply in the study of large systems with many components or systems where more states need to be defined or where there is a large number of interdependencies between component states. The latter is easier to study through the use of the GFCM. Still, it is the most accurate method in existence, due to the fact that in theory a listing of all the possible state combinations has to be made in beforehand.

D.4 Gathered fault combination method (GFCM)

This method first appears as a complement to the previous methods, namely the FMEA. Whereas FMEA lists only the effects of single failures, GFCM's aim is to study the interactions and failure combinations which result in undesired events. This is reached in an inductive way through study of firstly the failures inherent to elementary systems, and then the interaction between those systems and listing of failures which affect other elementary systems.

There are 4 basic steps (Villemeur, 1992-A) when applying the GFCM. First, a FMEA is to be applied to each elementary system – the elementary systems are broken down into components and a listing of all the single failures effects is obtained. Here the term "elementary system" is used to describe the smallest possible part of the system that can be tested by itself in a practical way.
The second step is to determine and list the internal gathered faults (IGF). IGFs are the grouping of failure modes in each elementary system which have the same effect on itself or other elementary systems. That is if for instance failure A and failure B have the same effect, they can be grouped into an IGF or if failure A.B (combination of failures A and B through an AND gate) and failure C have the same effect the same can be done. IGFs result barely from internal causes. Obviously, if the FMEA is applied, it will identify failure modes that are unique and different in a strict sense, so the question "how is it possible to group failure modes into IGFs?" will arise. This is done by applying engineering judgement, by defining the main and most important parameters in the elementary system and creating variation ranges for those. If different failure modes end up causing variations in parameters within the same range, then they can be considered similar and thus grouped into IGFs. This is an iterative process and must be applied carefully and while looking at the way systems work, and not only at the parameter variation data.
The third step is to analyse which of these IGFs are also EGFs or external gathered faults, i.e., the IGFs of an elementary system that affect other elementary systems operations. These can be of two types. Either:

- IGFs which by themselves cause effects on other elementary systems or
- IGFs from an elementary system which combined with one or more IGFs from other elementary system affect the operation of the latter.

Provisions must also be made to account combinations of EGFs as EGFs on their own if they produce different effects when combined. EGFs result from external causes.

The fourth and last step is to determine the GGFs, or global gathered faults. These will group both IGFs and EGFs or combinations of them which have the same effects on all the elementary systems studied.

The idea behind the whole method is to make a direct link of the failure modes of the system to the failure modes of components. This allows for, in future analyses which include the same elementary systems, the skipping of the failure modes analysis and a direct study of the system using IGFs, that is, a simplification of the process is achieved.

A difficulty that may arise is that the listing of failure combinations can be too burdensome or even impossible to accomplish. It is also predictable that a big number of combinations will have an extremely low probability of occurrence, and thus can be excluded from the study without a significant impact on results. For these two reasons, one of two limitations can be set: either combinations only up until a certain order are considered or only failure modes over a certain probability of occurrence are studied. The latter limitation will obviously force a simultaneous and iterative process in the qualitative and quantitative analyses of IGFs, EGFs and GGFs.
In the creation and listing of any of the gathered faults, the truth table method can provide assistance in the studying of the outcome of fault or failure mode combinations. Accordingly, the GFCM can solve a truth table method of the system where combinations of GGFs are made. This is a way of reducing the truth table method, for as stated above, the use of TTM in large and complex systems is too onerous.

There is no specific quantitative analysis method since this will depend on the way that faults or failure modes are grouped. Quantitative analyses can be carried out using auxiliary methods such as the CTM or TTM, for instance. Though this might sometimes prove to have a high workload, the GFCM is well adapted to the study of complex and large systems.

D.5 Consequence tree method (CQTM)

This method is mainly applied – and was firstly applied too – in the nuclear industry. The idea behind the method is to, from an "initiating event", foresee the consequent events which build upon it, and try to understand if there is any sequence of events which causes an "undesired event", i.e. an accident. In fact, this method is recommended in the American Guide to Performance of Probabilistic Assessments for Nuclear Power Plants and was even firstly called "event tree method", due to the fact that the safety systems in the nuclear industry require a very low number of decisions (in a strict sense) when accidents happen (Villemeur, 1992-A) – meaning the majority of systems are automatic and need no human operator decisions.

CQTM is best applied in systems with a high number of safety mechanisms, where there are many redundancies and interactions, and whenever there is a need to understand the consequences of certain events in the system – hence its big use in the nuclear industry.

The method always starts by determining an initiating event and then builds up possible consequences of that event in a tree-like structure, so that the next degree on the tree is an "initiating event" for the following, and so on. This will come to an end when all the safety mechanisms and redundancies have been exhausted, and in that time the events which are a consequence of the initiating event can be analysed qualitatively and quantitatively. The final events can either have unacceptable consequences (UC) or acceptable consequences (AC). Normally, the tree-like structure is built from left to right, being the UC – and thus unacceptable sequences – drawn towards the bottom, and the AC drawn towards the top.

The biggest issues with this method are: how to define the initiating event and all other potential initiating events, how to define the consequence events that happen throughout the tree, and how to remove inconsistencies and reduce the consequence tree.

In a deductive approach, what firstly needs to be done is a listing of all the safety mechanisms and functions of the system, understanding which purposes they serve, and what kind of events trigger a response from their part.
The initiating events can then be determined using engineering judgement, which takes in account the safety mechanisms and all past history of the system in study and its accidents, analogous systems, physical understanding of the phenomena involved, or even sometimes the conception of scenarios where external elements from the system are taken account (i.e., bombing of a nuclear power plant, etc.). Alternatively, the CTM can be used to determine causes of failures previously known, and so by using CTM and CQTM at the same time, theoretically the specialists can list exhaustively all types of events.
After writing down the initiating event, before the other consequent events are determined, the levels of safety must be understood and written down to its right. Using the knowledge about the safety functions, the specialist will have to understand which safety mechanisms and redundancies act triggered by the first event, and which will act as a second safety level and so on. This is where the first reductions are made, that is to say only safety functions and mechanisms which are relevant to the initiating event and the conditions set by it should be accounted in the analysis. Also, interactions between these functions must be taken in account. The failure of a safety function may trigger directly the failure of another thus making the latter irrelevant for the analysis.
As most of the previous methods, this one is also iterative, since the generic consequent events interact and depend on the safety functions regarded as relevant to the analysis. In the other way around, generic events may turn safety functions irrelevant for the analysis.

Aspects like time of events, functional interaction and interactions between elementary safety systems will help to sort out the tree and understand which generic events are to be accounted for. GFCM can help out in this stage, while looking for the interactions between elementary systems.

In an inductive approach, GFCM is applied to list out all the possible GGF. These will provide an overview of the failure paths (i.e. the chain of events needed in order to create a failure or more pointedly an accident), and so will give out which are the possible initiating events and elementary systems which are relevant for the failures. In fact, GFCM may disregard CQTM in the understanding of failure paths. In its essence, what CQTM does is a reduction of GFCM since if a set of two-possibility-state events in n safety systems is analysed, a number of 2^n possible paths is obtained – and this not accounting with the event order, which would make the number of possible paths $n!2^n$.
The selection of the initiating event is based on GGFs of the first order, that is, of the simplest type of combination between EGFs or IGFs. The idea is to then use GGFs as generic events as well, combining GGFs until all the levels of redundancy of safety systems are defeated.

One special case to be aware of is the one of feedback loops, in which an outcome of an event revisits a decision box of a previous event. This may be the representation for a standby process in which the system is “waiting” for a set of circumstances to arise, or it can represent continuous processes. This is harder to model by hand, and is preferably treated with computer software (Smith, 2005).

In any case the final result is a number of sequences, or paths, which combine all the possible states of safety functions with regard to the events that trigger them. This is a kind of reduction of the truth table method. Obviously reductions can be made by eliminating generic events or elementary systems which do not bring any consequences in future events or influence or change the states of other elementary systems.

D.6 Cause-consequence diagram method (CCDM)

This method is basically a conjunction of CTM with CQTM, that is, it links causes with consequences. It was first developed in Laboratory Riso in Denmark in the early 1970s for the application in Scandinavian power plants (Villemeur, 1992-A).

The idea is to start from an initiating event and draw out a logical diagram of the possibilities this event has. Typically, the initiating event leads to a “decision box”, where the analyst can ask a specific “yes” or “no” question, normally being “no” associated with failure. If this is the case then the YES branch proceeds to other levels of non-failed state functioning and the NO branch proceeds to the redundancy mechanisms which will allow for operation in the case of a failed state in the previous level. At the same time, causes determining the NO will be pointed out.

To an extent this is very similar to the cause tree method, especially in terms of appearance and symbols used. The big change is the introduction of an initiating event and decision boxes, which work as input and decision over that input respectively, and prompt the analyst to split his analysis into failed or non-failed states of components or elementary systems.
Construction is made from bottom to top, where the final events can be seen. Then minimal cut sets or failure paths can be deduced qualitatively, being the paths leading up to undesirable events, and thus these can be studied quantitatively.
A sensible aspect one must not overlook is the selection of the initiating event. The FMEA can be of assistance in this stage. Initiating events are typically failures of components or elementary systems which are vital to the operation of the system, events leading to complete loss of redundancy, abnormal events in or outside the system, events known to disturb a big number of control parameters in the system, failures in auxiliary systems, and the first event in important procedures of a system, such as start-up or shut-down (Villemeur, 1992-A).

The great advantage of the CCDM is the almost simultaneous analysis of event consequences and their causes. It provides a clear vision of how systems work, but is of very difficult application in complex or larger systems.

D.7 State-space method (SSM)

The state space method is a technique that aims to make an exhaustive study of the transition of states of components of a system. This allows for a study of repairable systems, and is often irreplaceable in doing so. It is particularly useful in the study of Markovian processes, and as with any other method, easier to implement with small systems. The cause tree method can be used together with SSM.
Non Markovian processes can also be modelled, but calculation becomes harder unless a semi-Markovian process is being studied. Under special conditions, consequence trees can be used as an approximation to semi-Markovian processes.

The modelling follows three major steps. The first step is to make an exhaustive list of system states by verifying the combinations of every elementary system or component state (operating, failed, other…). If a system has n components with x possible states for each component then there will be $p=x^n$ possible system states. With two possible states and ten components, a value of $2^{10}=1024$ possible system states is obtained – this number climbs particularly fast with the size of the system. FMEA can be useful during this stage as an auxiliary method.
The second step consists in carrying out a study of the possible state transitions of components or elementary systems accounting the interactions between them. Here, all failures, repair stages and other maintenance procedures must be accounted. Interactions are important for sometimes failures in one component can trigger failure in others. There can also be procedures which return a failed

system to the operating state once they are completed. CTM and GFCM can be of assistance during this stage.

Finally, the last step is the calculation of the probabilities of being in certain states or of different reliability parameters (MTBF, MTTF, MTTR, failure rate, etc). This can be achieved by calculations applied from the qualitative study done in the last stage, through the transition rate matrix. The qualitative analysis can be represented through a diagram, where every node is a given system state, and every arc linking them represents the transition between states. The transition rates are represented through λ or μ, that is the failure and repair rates of components or elementary systems.

The basic representation – a two state component – is shown in Figure D.4.

If, for instance, a parallel system (active redundancy) is represented, there will be four different states, which are the four possible combinations of component 1 and 2 being in an operating state (represented by 0) or in a failed state (represented by 1). The diagram is represented in Figure D.5.

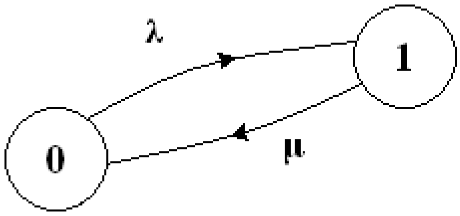


Figure D.4 Basic representation of a two state component as a SSM diagram.
(Adapted from Villemeur, 1992-A)

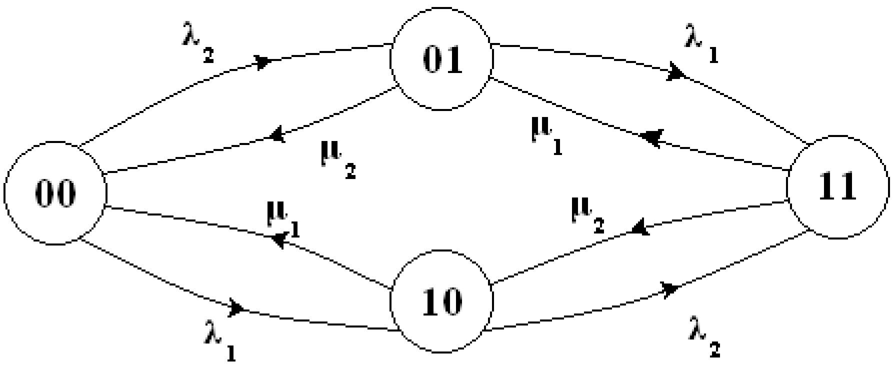


Figure D.5 Representation of a repairable two component system.
(Adapted from Villemeur, 1992-A)

In Figure D.5:

State 1 (00) represents both components working;

State 2 (01) represents component 1 operating and component 2 failed;

State 3 (10) represents component 2 operating and component 1 failed;

State 4 (11) represents both components failed.

For reliability purposes, what basically is done is to turn the system failed states, i.e. the states that cause a system failure, into "absorbing states". This means once that state is reached there is no way to make a transition from the failed state into any operating state. This changes the diagram from Figure D.5 into the one shown in Figure D.6:

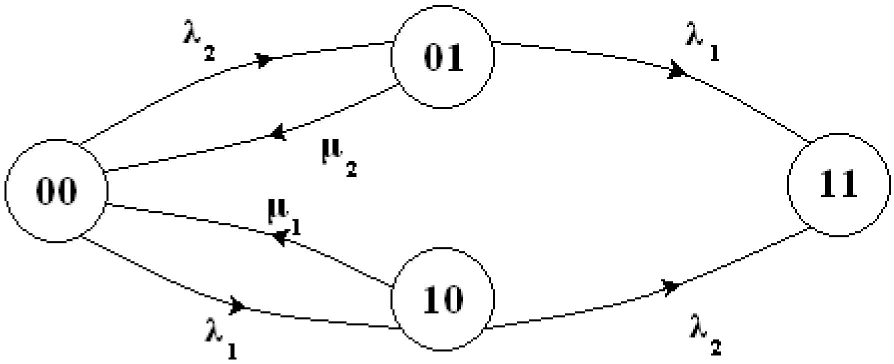


Figure D.6 Representation of a non-repairable two component system[13].
(Adapted from Villemeur, 1992-A)

What is done in Markov-type repairable systems is to write down the state equations of the system and solve them through the transition rate matrix.
The system state equations are represented by the probability $Q_i(t+dt)$ that the system is in state i at time t+dt. Knowing that the transition rate from state i to a state j is represented by a_{ij}, one can calculate $Q_i(t+dt)$ knowing that:

Firstly, if the system is in a state i at time t, then it remained in that state to comply with the event described above. This "non-transition" has a probability of:

$$1-\sum_{j\neq i} a_{ij}\,dt \tag{D.1}$$

If the system was in a state j at time t then it necessarily changed from j to i to comply with the event. This has a probability of:

$$a_{ji}\,dt \tag{D.2}$$

Consequently, the probability $Q_i(t+dt)$ will be:

[13] Here the "non-reparability" is represented by the elimination of the repair rates from the 11 state, and is used only for reliability purposes, i.e., simulation is over whenever the system reaches the 11 state.

$$Q_i(t+dt) = P[system\ is\ kept\ in\ state\ i\ at\ time\ t] + \\ + \sum_{j \neq i} P[system\ changes\ from\ state\ j\ to\ state\ i\ over\ [t, t+dt]] \quad \text{(D.3)}$$

Therefore, using D.1, D.2 and D.3:

$$Q_i(t+dt) = Q_i(t)\left[1 - \sum_{j \neq i} a_{ij}dt\right] + \sum_{j \neq i} Q_j(t)a_{ji}dt \quad \Leftrightarrow$$

$$\frac{Q_i(t+dt) - Q_i(t)}{dt} = -Q_i(t)\sum_{j \neq i} a_{ij} + \sum_{j \neq i} Q_j(t)a_{ji} \quad \Leftrightarrow$$

$$\frac{dQ_i(t)}{dt} = -Q_i(t)\sum_{j \neq i} a_{ij} + \sum_{j \neq i} Q_j(t)a_{ji} \quad \text{(D.4)}$$

Taking in consideration that:

$$a_{ii} = -\sum_{j \neq i} a_{ij} \quad \text{(D.5)}$$

It is then possible to yield, from D.4 and D.5:

$$\frac{dQ_i}{dt}(t) = \sum_{j=1}^{p} Q_j(t)a_{ij} \quad \forall i = 1,2,..., p \quad \text{(D.6)}$$

Here, p represents the total number of possible states of the system. In the case of a homogeneous process the terms a_{ij} are constant, it is possible to differentiate Q_i, and this set of differential equations for each state makes up the system state equations, which in matrix form can be expressed as:

$$\left[\frac{dQ_1}{dt}(t), \frac{dQ_2}{dt}(t), ..., \frac{dQ_p}{dt}(t)\right] = [Q_1(t), Q_2(t), ..., Q_p(t)]A \quad \text{(D.7)}$$

In expression D.7, matrix A is square, of dimensions $p\ x\ p$. The terms in each row of the matrix sum up to 0, which means matrix A is a singular matrix, i.e. its determinant is equal to zero. This will allow for the solution to be found if a distribution – normally the initial distribution $Q_i(0)$ – is known.

The calculations can be made by solving simple first-order linear differential equations systems with Laplace transforms, discretization, or the calculation of eigenvalues of matrix A – with matrix exponentials:

$$Q(t) = Q(0)e^{At}$$

(D.8)

The last method can prove itself problematic for very reliable systems, due to the fact that the largest eigenvalue $-\beta_1$ is a lot smaller than the others when in absolute value, and precision issues can make it quite inaccurate when there is a large number of p. Furthermore, this eigenvalue will forecast how the system works once there are large enough values of t. These disadvantages make this last method to be used very seldom.

Availability is defined as the probability of the system being in one of its operating states. If it is considered that, out of the p states, there are l operating states, availability will be written as:

$$A(t) = \sum_{i=1}^{l} Q_i(t) \tag{D.9}$$

To reach this value, first the asymptotic probability of the system being in a state Q_i as time t tends to infinity ($Q_i(\infty)$) is defined by means of:

$$\sum_{j=1}^{p} Q_j(\infty) a_{ji} = 0 \quad \forall i = 1, 2, ..., p \tag{D.10}$$

$$Q_i(\infty) = -\frac{\sum_{j \neq 1} Q_j(\infty) a_{ji}}{a_{ii}} \tag{D.11}$$

Solving the system state equations through the Laplace transform or the equation above the previous returns:

$$Q_i(\infty) = \frac{1}{\Delta} \begin{vmatrix} a_{11} & \cdots & a_{1,p-1} & 0 \\ \vdots & & \vdots & 0 \\ a_{i1} & \cdots & a_{1,p-1} & 1 \\ \vdots & & \vdots & 0 \\ a_{p,1} & \cdots & a_{1,p-1} & 0 \end{vmatrix} \qquad with \quad \frac{1}{\Delta} = \begin{vmatrix} a_{11} & \cdots & a_{1,p-1} & 1 \\ a_{i1} & \cdots & a_{1,p-1} & 1 \\ \vdots & & \vdots & \vdots \\ a_{p,1} & \cdots & a_{p,p-1} & 0 \end{vmatrix} \tag{D.12}$$

This means the asymptotic availability can be written as:

$$A(\infty) = \frac{1}{\Delta}\begin{vmatrix} a_{11} & \cdots & a_{1,p-1} & 1 \\ \vdots & & \vdots & \vdots \\ a_{l1} & \cdots & a_{l,p-1} & 1 \\ a_{l+1,1} & \cdots & a_{l+1,p-1} & 0 \\ \vdots & & \vdots & \vdots \\ a_{p,1} & \cdots & a_{p,p-1} & 0 \end{vmatrix} \tag{D.13}$$

Unavailability will then be, as defined:

$$\overline{A}(\infty) = 1 - A(\infty) = \frac{1}{\Delta}\begin{vmatrix} a_{11} & \cdots & a_{1,p-1} & 0 \\ \vdots & & \vdots & \vdots \\ a_{l1} & \cdots & a_{l,p-1} & 0 \\ a_{l+1,1} & \cdots & a_{l+1,p-1} & 1 \\ \vdots & & \vdots & \vdots \\ a_{p,1} & \cdots & a_{p,p-1} & 1 \end{vmatrix} \tag{D.14}$$

Due to the extent of this method's calculation procedures, if further interest persists, the reader should be referred to Villemeur, 1992-A.

Annex E: Forms destined to field data collection

These forms are meant to be kept together, so that they become part of a process which starts upon inspection and detection of each failure, until its correction and restoring of operation. The presented values are merely a suggestion, for their definition falls out of the thesis' scope.

Infrastructure and train traffic form – GIS compilation

Location (coordinates, distance from reference point):

Special stretch? □Yes □ No

If Yes, description (bridge joint, tunnel joint, switch, other):________________________________

____

Rail characteristics

Profile type and weight: □ UIC________ □ S___________ □ BS_________

□ other_______

Steel grade: □ R200 □ R220 □ R260 □ R260 Mn
□ R320 Cr □ R350 HT □ R350 LHT

Track gauge: □ Standard (1435 mm) □ Wide (1666/7 mm) □ Other______

Installation date (dd/mm/yy):

MGT withstood to date:

Weld zone? □ Yes □ No

If Yes, weld type: □ Electric flash-butt □ Thermit □ Electric arc □ Other_______

Welding date (dd/mm/yyyy):

MGT withstood to date:

Fastening characteristics

Model: □ DE clip □ Pandrol standard □ Pandrol Fastclip
□ Vossloh □ Nabla □ Other_______

Installation date (dd/mm/yy):

MGT withstood to date:

Sleeper characteristics

□ Monoblock □ Twinblock:

Weight (kg): □ 0-75 □ 75-150 □ 150-250 □ 250-350 □ 351-450
□ 451-over

Spacing (cm): □ 40 □ 60 □ 80 □ Other_______

Installation date (dd/mm/yy):

MGT withstood to date:

Support system

□ Ballast □ Slab

Thickness (cm): □ 0-20 □ 20-40 □ 40-60 □ 60-80 □ 80-more

Subgrade type and quality

UIC classification: □ S_3 □ S_2 □ S_1 □ S_0

Geometrical characteristics

Curve radius (m): □ 0-700 □ 700-2500 □ 2500-5000 □ 5000-7500
□ >7500 (considered as straight stretch)

Cant / superelevation (cm): □ 0-5 □ 5-10 □ 10-15 □ 15-20
□ 20-25 □ 25-more

Gradient transition zone? □ Yes □No

If No, which gradient? □ 0-1% □1-2%

Traffic characteristics

Predominant direction? □ Yes □ No

If Yes, which?_______________

Traffic type: □ Passenger □ Freight □ Mix

Axle load: □ A (max 16t) □ B (max 18t) □ C (max 20t) □ D (max 22.5t)

□ More (how much?)__________

Traffic load (estimate of MGT per year):_______

Maintenance

Inspection frequency: □ once a year or more □ more than once a year

Rail grinding frequency: □ every 2 years or more □ once a year □ more than once a year

Lubrication on rail? □ Yes □ No

Tamping frequency: □ every 2 years or more □ once a year □ more than once a year

Ballast cleaning frequency: □ every 2 years or more □ once a year □ more than once a year

Failure registry form

Failure registry number:

Location of failure (coordinates, distance from reference point):

Radius < 7500 m? (i.e., Curve?) □ Yes □ No

If Yes: □ Lower / inner rail □ Higher / outer rail

Rail failure mode detected (UIC coding system):

Zone near rail ends (discontinuities):

□ UIC 111 – Transverse cracking

□ UIC 113 – Longitudinal vertical cracking of the rail head

□ UIC 133 – Longitudinal vertical cracking of the rail web

Zone away from rail ends:

□ UIC 211 – Transverse cracking / *Tache ovale*

□ UIC 212 – Horizontal cracking

□ UIC 213 – Longitudinal vertical cracking of the rail head

□ UIC 2201 – Short-pitch corrugations

□ UIC 2202 – Long-pitch corrugations

□ UIC 2203 – Lateral (gauge corner) wear

□ UIC 221 – Rolling surface disintegration

□ UIC 2221 – Shelling of the running surface

□ UIC 2222 – Gauge-corner shelling

□ UIC 2223 – Head checking / spalling

□ UIC 2251 – Isolated / single wheel burn

□ UIC 2252 – Repeated wheel burns

□ UIC 227 – Squats

□ UIC 233 – Longitudinal vertical cracking of the rail web

<u>Weld zones:</u>

□ UIC 411 – Transverse cracking

□ UIC 412 – Horizontal cracking of weld

□ UIC 421 – Transverse cracking of profile in Thermit weld sections

□ Buckling

□ Plastic flow and tongue lipping

□ Other

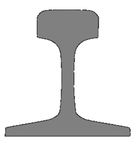
Installation date of the rail (dd/mm/yy):

MGT withstood to date by rail (i.e. MTBF in [MGT]):

□ First time failure mode is detected in the zone □ Reincidence:________ time

<u>Failure mode / rail defect description</u>

Location: Description:____



____

Wear / Crack opening / other measurement: ____________________

____

Inspection method:

□ Visual □ Noise / excessive vibration □ Track inspection car □ Ultrasound handheld equip.

Inspection date (dd/mm/yyyy):

Inspection time (24h):

Picture (s):

Insert picture(s) here

Management of failure

Action(s) to take:

□ Rail grinding □ Rail reprofiling □ Weld correcting □ Rail replacement □ New inspection

Scheduling of action (dd/mm/yyyy):

Hour (24h):

If welding is made, weld method used:____________________

Failure correction report – after failure management

Failure registry number:

Management of failure

Action taken:

□ Rail grinding □ Rail reprofiling □ Weld correcting □ Rail replacement

Date of action (dd/mm/yyyy):

Time (24h):

Total duration of maintenance time [accounts for MTTM]:

Total duration of down time [accounts for MDT]:

Number of interventions and partial intervals:

1st	2nd	3rd	4th	…
___hours	___hours	___hours	___hours	___hours

Resources used:

Spares used:____

Maintenance crew:____________________________________

Equipment used:_____________________________________

Costs:

Spares costs:_______________________________________

Maintenance costs:___________________________________

Equipment costs:____________________________________

Total costs:____

Other observations:

____

____

____

Annex F: Explanation on Weibull method calculations

The Weibull cumulative distribution function in order to the variable *x* is written as:

$$F(x) = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}} \tag{F.1}$$

This contains the assumption that the γ parameter is null, which might be faced as true for at 0 MGT no failures are to be expected. The conversion to a type y=mx+b linear equation is achieved through:

$$1 - F(x) = e^{-\left(\frac{x}{\alpha}\right)^{\beta}} \quad \Leftrightarrow$$

$$\Leftrightarrow \quad \ln(1 - F(x)) = -\left(\frac{x}{\alpha}\right)^{\beta} \quad \Leftrightarrow$$

$$\Leftrightarrow \quad \ln\left(\frac{1}{1 - F(x)}\right) = \left(\frac{x}{\alpha}\right)^{\beta} \quad \Leftrightarrow \tag{F.2}$$

$$\Leftrightarrow \quad \ln\left[\ln\left(\frac{1}{1 - F(x)}\right)\right] = \beta \ln\left(\frac{x}{\alpha}\right) \quad \Leftrightarrow$$

$$\Leftrightarrow \quad \ln\left[\ln\left(\frac{1}{1 - F(x)}\right)\right] = \beta \ln x - \beta \ln \alpha$$

Therefore, parameter β is simply the slope *m* of the regression line fitted to the Weibull plot, and α is calculated through the expression:

$$\alpha = e^{-\left(\frac{b}{\beta}\right)} \tag{4.7}$$

Annex G: Simple example of a Pareto analysis

A simple demonstration of the carrying of a Pareto analysis is made here. Assuming for instance the following set of fictitious defined failures, which upon discovery imply the performance of a maintenance action with an associated defined cost and MDT as shown in Table G.1:

Table G.1 Example of defined failures and associated costs and maintenance times.

Failure	Average Maintenance Action Cost	MDT associated to Maintenance Action
Gauge corner wear > 2 mm	500€	3h
Corrugations with amplitude > 1 mm	300€	2h
Squat > 1 mm depth	600€	3,5h
Plastic flow on the field side > 2 mm	600€	4h
Rail break due to *tache-ovale*	1600€	5h

Reliability expressions may be obtained using the procedure stated in 4.2.3, and reliability values may be forecast using the procedure stated in 4.2.4. Assuming the values forecast for a given year for **Unreliability** for a whole network (mean value) are the ones show in Table G.2:

Table G.2 Unreliability for year 10 for each defect.

Failure	Q (year 10)
Gauge corner wear > 2 mm	0,2
Corrugations with amplitude > 1 mm	0,3
Squat > 1 mm depth	0,25
Plastic flow on the field side > 2 mm	0,15
Rail break due to *tache-ovale*	0,05

The most useful way to carry the Pareto analysis is to multiply costs and MDT by the unreliability for that year. The data may be ordered by any criteria, in this case in ascending cost order, like shown in Table G.3.

Table G.3 Pareto analysis of the considered rail defects.

Failure	Cost x Q (year 10)	MDT x Q (year 10)
Rail break due to *tache-ovale*	80€	0,25h
Corrugations with amplitude > 1 mm	90€	0,6h
Gauge corner wear > 2 mm	100€	0,6h
Squat > 1 mm depth	150€	0,875h
Plastic flow on the field side > 2 mm	240€	0,6h

The conclusion in the example is that the forecast cost is larger for the maintenance action associated to the correction of the failure “squat > 1 mm”. The example shown here is made for just one “comparable track stretch”. It should be repeated should for all track stretches of each line (or maintenance intervention area) in order to obtain the total forecast costs.

Annex H: Example of a (fictitious) FMEA for the rail infrastructure

Component identification	Functions	Failure modes	Possible failure causes (Internal, external causes)	Failure effects (consequences)	Detection method	Action required	Time needed for maintenance action	Spares, equipment, crew size needed
Rail	Load bearing, Train guiding	UIC2203 gauge corner wear	Higher traffic speed than design speed, excessive load	Wear on rail head	Visual, ultrasonic	Reprofiling	3h	Reprofiling car, etc.
Rail	Load bearing, Train guiding	UIC2201 corrugations	Excessive braking force, cold climate	Excessive noise / vibration, Effect on ballast bed (deformation if vibration is excessive)	Visual, ultrasonic, noise	Grinding	2h	Grinding car, etc.
…	…	…	…	…	…	…	…	…
Ballast	Load bearing, Rain water drainage, noise dampening	Deformation	Geological settlings, excessive loading, insufficient sleeper spacing, Vibration of ballast due to corrugations	Deformed rail track geometry, risk of cant or other	Visual, other	Tamping	1h	Tamping machine, stone blowers, hand tools
…	…	…	…	…	…	…	…	…